Hindawi
Journal of Sensors
Volume 2018, Article ID 8249180, 11 pages
https://doi.org/10.1155/2018/8249180

Research Article

Saliency Detection by Multilevel Deep Pyramid Model

Hai Wang,[1,2] **Lei Dai**,[1] **Yingfeng Cai**,[2,3] **Long Chen**,[3] **and Yong Zhang**[4]

[1]*School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China*
[2]*Robotic and Automation Lab, The University of Hong Kong, Hong Kong*
[3]*Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China*
[4]*School of Automotive and Traffic Engineering, Nanjing Forestry University, Nanjing 210037, China*

Correspondence should be addressed to Yong Zhang; zy.js@163.com

Received 22 March 2018; Accepted 24 July 2018; Published 14 August 2018

Academic Editor: Calogero M. Oddo

Traditional salient object detection models are divided into several classes based on low-level features and contrast between pixels. In this paper, we propose a model based on a multilevel deep pyramid (MLDP), which involves fusing multiple features on different levels. Firstly, the MLDP uses the original image as the input for a VGG16 model to extract high-level features and form an initial saliency map. Next, the MLDP further extracts high-level features to form a saliency map based on a deep pyramid. Then, the MLDP obtains the salient map fused with superpixels by extracting low-level features. After that, the MLDP applies background noise filtering to the saliency map fused with superpixels in order to filter out the interference of background noise and form a saliency map based on the foreground. Lastly, the MLDP combines the saliency map fused with the superpixels with the saliency map based on the foreground, which results in the final saliency map. The MLDP is not limited to low-level features while it fuses multiple features and achieves good results when extracting salient targets. As can be seen in our experiment section, the MLDP is better than the other 7 state-of-the-art models across three different public saliency datasets. Therefore, the MLDP has superiority and wide applicability in extraction of salient targets.

1. Introduction

Visual saliency aims to extract the most significant regions and targets in a scene by simulating the human visual attention system. In recent years, visual salient object detection models has been applied to many applications such as video summary [1], specific object retrieval [2], and object detection [3].

In order to calculate visual saliency, traditional models usually based on the image contrast. For example, the global contrast-based salient model [4] divides the image into several small image regions, and the contrast between the small image regions is used to highlight salient targets. This kind of model has a good overall contrast and location of salient targets is accurate, but its contours of salient targets are relatively fuzzy. Salient models [5] obtain the saliency map by comparing the neighboring pixels in the image. This method can extract the contours of salient targets but is susceptible to complex backgrounds, which reduce the accuracy of target detection.

At the same time there are many salient models based on low-level features, the most famous of which is the traditional pyramid model proposed by Itti and Koch [6]. In this model low-level features such as color, direction, and brightness are extracted at different channels. This model can simulate the biological central-surrounding suppression mechanism in the human visual system, and the saliency map is obtained by multiscale feature fusion, but only low-level features are extracted. So the contours of salient targets are fuzzy due to background noise. Therefore, in order to avoid background noise interference, a model based on fusion of background and foreground [7] was proposed. This model is useful in avoiding background noise interference, but are still not accurate enough. Meanwhile, with the continuous improvement of the deep learning network and deep convolutional network [8] in the area

of computer vision and image processing, a salient model using deep convolution network based on high-level features was proposed [9] which divides the original image into small image blocks. The small image blocks are convoluted and pooled at different levels and iterated to obtain the feature dictionary of the original image. Then, the saliency score of each pixel is calculated by a feature dictionary and support vector machine (SVM). This kind of model combines the advantages of models which is based on local contrast and extracts high-level features (such as details of human faces). However, due to the lack of color, space, and other low-level features, the salient map is easily affected by background noise.

In order to solve these problems, we propose a deep pyramid model called the MLDP (multilevel deep pyramid) model, which integrates low-level features, high-level features, and local contrast. The MLDP model is based on the structure of the pyramid [6], the VGG16 deep model, the superpixel segmentation mapping structure, and the background noise filtering structure [7].

First, we place the entire image into the VGG16 model to get an initial salient map. Then, we apply a pyramid structure simulating the central-surrounding suppression mechanism, because this pyramid structure can extract features based on local contrast. We then divide the initial saliency map into six different scales to form the initial saliency map pyramid. In this way, we can compare the multiscale images and extract local contrast features. The VGG16 model has a relatively small convolution kernel, small steps, and has better accuracy; thus, we selected the VGG16 model for our pyramid to extract high-level features. The VGG16 model extracts high-level features for each scale image of the pyramid to form a deep pyramid with high-level features, and a multiscale feature map formed by center-surround difference. For the obtained feature map, the "winner-takes-all" policy and inhibition of return are used to create the saliency map based on the deep pyramid. It uses the VGG16 deep model to extract high-level features, so it has better feature extraction accuracy compared with shallow model. The superpixel segmentation mapping structure is designed to extract color, brightness, texture, and other low-level features, because superpixels are the small areas whose pixels are similar to each other in position, color, brightness, texture, and other low-level features. By mapping between superpixels and a saliency map based on a deep pyramid, low-level features can be added to form a saliency map fused with superpixels. Because the saliency map fused with superpixels is sensitive to background noise, the background noise filtering structure (based on low-level factors such as color and spatial distance) is used to eliminate the effect of background noise. The background noise filtering structure can also enhance the extraction of low-level features in order to obtain a saliency map without background noise (a saliency map based on foreground). The final saliency map is a fusion of the saliency map based on the foreground and the saliency map fused with superpixels, so the feature extraction of the final saliency map is more comprehensive and accurate.

Our MLDP model makes the following four contributions:

(1) We do not use the traditional pyramid structure to extract the low-level features of three channels such as color, direction, and brightness. Instead, we creatively use the VGG16 model to extract high-level features to form an initial saliency map as the deep pyramid structure input.

(2) We add an extra spatial pyramid pool layer to the VGG16 model in order to adapt the deep pyramid structure for different scales.

(3) We use the VGG16 model with a spatial pyramid pooling layer in each scale of the pyramid to construct a deep pyramid structure. With the local contrast feature extraction, the high-level features can be extracted more completely and accurately.

(4) We create a superpixel contrast mapping structure. The superpixel segmentation is based on the characteristics of low-level features, and these low-level features are added to saliency map based on the deep pyramid to simultaneously extract low-level features.

2. Our Approach

In recent years, researchers have been inspired by the human visual attention system and have proposed many visual salient object detection models [10–15]. In this section, we will introduce our MLDP model, as shown in Figure 1. It consists of five parts: (1) the VGG16 model is used to extract the high-level features of the original image, forming an initial saliency map; (2) we obtain the image pyramid by multiscale segmentation and apply the VGG16 model with a spatial pyramid pooling layer to each scale of the image pyramid to form the deep pyramid; (3) the saliency map based on the deep pyramid is mapped with superpixel segmentation to extract low-level features; (4) background noise filtering structure; and (5) weighted fusion structure of multilevel saliency maps. Next, we will introduce these five parts in turn.

2.1. Forming the Initial Saliency Map. We take the original image as the input for a VGG16 model to extract high-level features. The VGG16 model in our MLDP model is similar to the traditional VGG16 model [8] with regard to feature extraction, which contain convolution and pooling iterations. The difference is that we add a spatial pyramid layer [16] in front of the full connecting layer of the VGG16 model in order to adapt to the different scales of images in the pyramid. The structure of our VGG16 model is shown in Figure 2. We use five convolution layers and obtain the initial global saliency map through the activation of the full connecting layer. The five convolution layers can extract the global high-level features, and the spatial pyramid pooling layer can avoid the change of parameters in the full connecting layer due to the changed size of the initial saliency map, which can make the training easier. However, because the initial saliency map is based on the global high-level feature

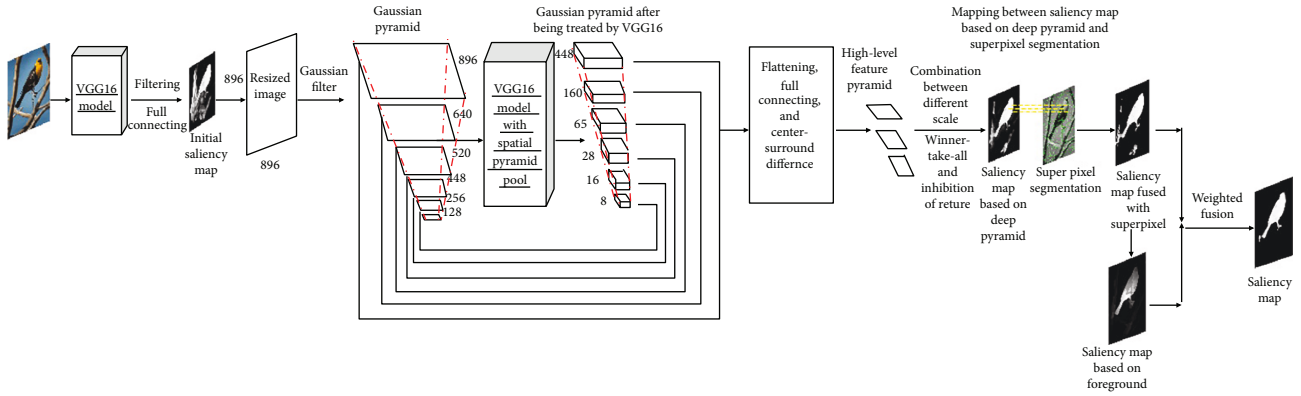


Figure 1: The overall architecture of the proposed MLDP model.

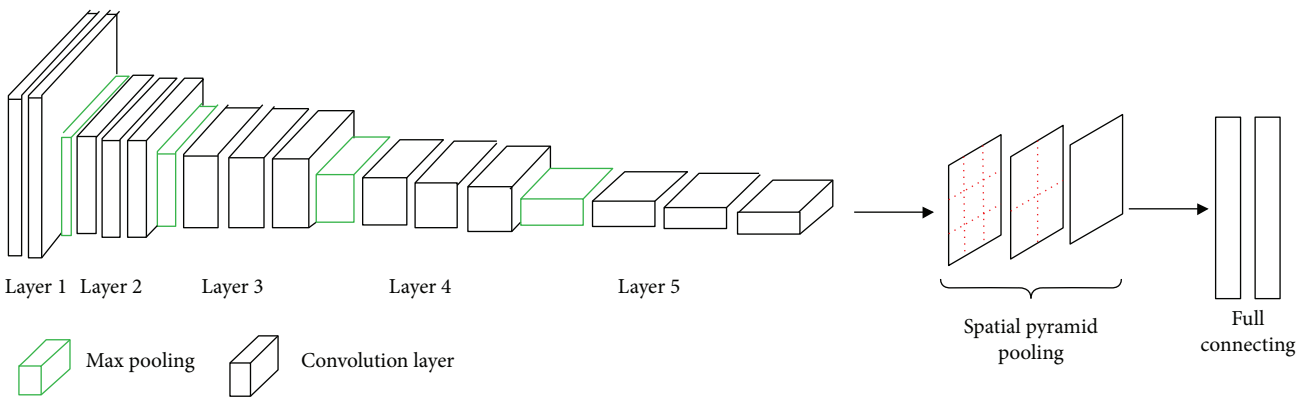
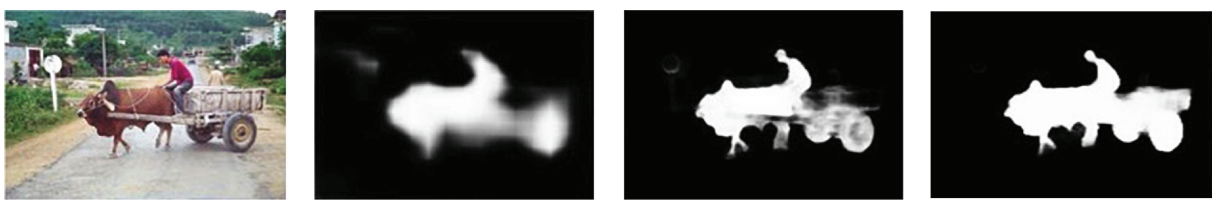


Figure 2: The architecture of the VGG16 model with spatial pyramid pooling.



(a) Image

(b) Saliency map based on deep pyramid without VGG16 model

(c) Saliency map based on deep pyramid without Gaussian pyramid

(d) Saliency map based on deep pyramid with Gaussian pyramid and VGG16 model

Figure 3: The importance of VGG16 and Gaussian pyramid.

information, ignoring the local contrast feature information and low-level feature information, the initial global saliency map cannot extract the details of salient targets. Therefore, we use the deep pyramid structure and superpixel segmentation to extract the local and low-level features.

2.2. Multiscale Deep Pyramid. The multiscale image pyramid is a feature extraction method based on local contrast. Unlike a traditional pyramid, we ignore low-level features like color, brightness, and direction, because the pyramid in our model is used to extract the high-level features with local contrast. We use the initial global saliency map as the input for the pyramid structure and apply the VGG16 model with a spatial pyramid pooling layer to each scale in the pyramid. The main contribution of applying the VGG16 model to the Gaussian pyramid is that the VGG16 model has excellent ability to extract the features, but VGG16 model lacks the features confrontation mechanism which exists in the pyramid, the confrontation mechanism has been proved significant in the salient object detection of the human visual system [17]. Without the confrontation mechanism, the performance of VGG16 model will fall in the salient object detection. On the other hand, if the pyramid loses the VGG16 model, the pool performance in extracting the features like color, brightness, and direction of the traditional pyramid will restrict the performance in the salient object detection. Therefore, it is important to apply the VGG16 model to the Gaussian pyramid as shown in Figure 3.

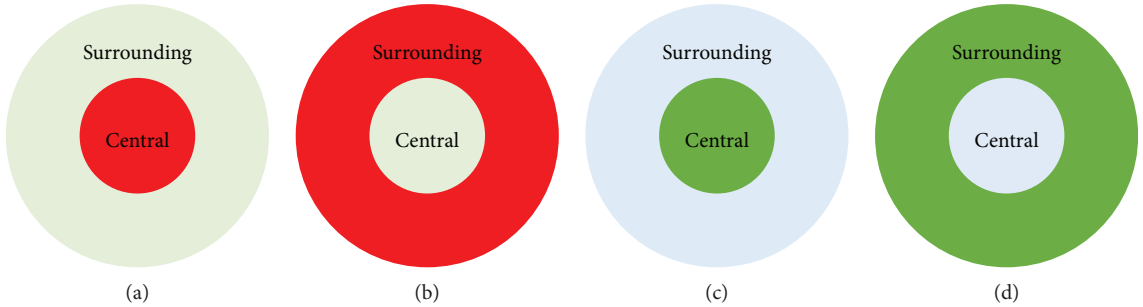


Figure 4: The sketch map of confrontation mechanism in human visual system.

Table 1: The number of layers in different scale of pyramid.

The scale of pyramid	896×896	640×640	520×520	448×448	256×256	128×128
The number of layers	2	3	4	5	5	5

The VGG16 model has a fixed scale requirement to the input; our Gaussian pyramid has different scales; if we want to use the VGG16 model in the Gaussian pyramid, the multiscale requirements of the pyramid must be solved. In view of the above problem, we add a spatial pyramid pooling layer [16] to deal with the problem which is another main contribution in applying the VGG16 model to pyramid. As shown in Figure 2, the spatial pyramid pooling divides the input into the fixed grid 4×4, 2×2, and 1×1. Through the fixed grid, the final output in the full connecting layers will normalize to 4096×1 of different scale input. In our model, the VGG16 model containing a spatial pyramid pooling layer can adapt to the multiscale requirements of the pyramid and form a deep pyramid.

The idea in our model inspired by the human visual attention is multiscale deep pyramid. As shown in the research of human visual attention system [17], the human visual system can detect the visual salient object due to the confrontation between the central area and the surrounding area of feeling field in the visual cell, if the central area is more salient than the surrounding area, the salient object in the human visual system is the central area as shown in Figures 4(a) and 4(c), otherwise salient object is the surrounding area as shown in Figures 4(b) and 4(d). Therefore, the central and surrounding are confronting each other to produce the final result. In our model, the scale we choose as the surrounding area is the scale 896×896, 640×640, and 520×520 in the Gaussian pyramid, and the scale we choose as the central area is the scale 448×448, 256×256, and 128×128 in the Gaussian pyramid. As similar to the Gaussian pyramid, we choose the scale 448×448, 160×160, and 65×65 as the surrounding area and the scale 28×28, 16×16, and 8×8 as the central area. Through the following central-surrounding difference mechanism, our model can simulate the confrontation between the central area and the surrounding area of feeling field in visual cell of human visual system.

The number of layers in VGG16 model varies with the scale of pyramid to avoid excessive image size which may cause distortion, as shown in Table 1:

The flattening and full connecting layer are then used to obtain the high-level feature map of the deep pyramid at multiscale. Its formula is as follows:

$$v(x) = fc(f(V(x))), \tag{1}$$

where $v(\cdot)$ is the extraction of high-level features with the VGG16 model, x is the high-level feature saliency map at multiscale, $f(\cdot)$ is the flattening operation, $fc(\cdot)$ represents the full connecting layer, and $v(x)$ is the multiscale high-level feature map of the deep pyramid.

The deep pyramid simulates the central-surrounding difference of the human visual system, which can extract local contrast features. We subtract between the different levels of the high-level feature map of the deep pyramid. The formula is as follows:

$$q(c, s) = |v(c) \otimes v(s)| (c \in \{2, 3\}, w \in \{3, 4\}, s = c + w), \tag{2}$$

where $v(c)$ and $v(s)$ represents the multiscale high-level features map of the deep pyramid, respectively, $\otimes$ indicates the point-to-point subtraction between the multiscale high-level features map of the deep pyramid, and $q(c, s)$ is the obtained multiscale local contrast feature map of the deep pyramid.

In order to fuse the multiscale local contrast feature map of the deep pyramid, the deep pyramid defines a normalized function, which has the following formula:

$$N(x) = i^* (M - m)^2, \quad i \in x, M = \max(x), m = \bar{x}, \tag{3}$$

where x is the inputted multiscale local contrast feature map of the deep pyramid, i represents the feature scores of the local contrast feature map, M is the maximum value of the local contrast feature map, and m is the feature average of the local contrast feature map.

We normalize the local contrast features at different scales using the normalization function, and perform multiscale fusion. The formula is as follows:

$$J = \bigoplus_{c=2}^{3} \bigoplus_{s=c+2}^{c+4} N(q(c,s)), \tag{4}$$

where J is the saliency map based on the deep pyramid.

On the basis of the high-level features of the initial saliency map, the deep pyramid is used to further extract high-level features, and the local contrast features are also fused to enhance the extraction of salient targets.

2.3. Using Super-Pixel Segmentation to Mitigate the Missing Salient Scores of Low-Level Features. Superpixel segmentation is based on similarities between pixels in low-level features such as color and spatial distance. These pixels with similar low-level features are classified as a region, in order to segment regions whose pixels are similar in low-level features. These regions are mapped onto a saliency map based on a deep pyramid, so as to carry out an image region segmentation operation on a deep pyramid saliency map based on the principles of low-level feature similarity. The average of pixel points' salient scores in the region is calculated for each region. Because pixels in the same region are similar in color, spatial distance, and other low-level features, if the salient scores of the pixel are lower than the average score in the same region, the pixel's salient score will be replaced by the average score. This method is actually based on the high-level salient scores, according to the similarity of pixel's low-level features in the same region:

$$\begin{aligned} a &= \frac{\sum_{i\in n}\sum_{j\in n} d(i,j)}{n}, \\ D(x,y) &= \begin{cases} d(x,y), & d(x,y) \geq a \\ a, & d(x,y) < a, \end{cases} \end{aligned} \tag{5}$$

where n indicates the total number of pixels in a small region, $d(x,y)$ is the salient scores of each pixel occupying a small region of the saliency map based on a deep pyramid, and x, y is the coordinates of the pixel. Therefore, this method can compensate for the reduction of salient scores caused by a lack of low-level features.

2.4. Background Noise Filtering. We use the saliency map extraction method based on foreground clues [7] to filter the interference of background noise. This is primarily divided into two parts as detailed below.

2.4.1. The Choice of Foreground Clue. We use the method of adaptive thresholds [18] to segment the saliency map fused with superpixels, and select those pixels whose salient scores are greater than the threshold as the foreground clue. We use adaptive thresholds rather than a fixed threshold because the adaptive thresholds can be adapted to the different origins of the input and have good accuracy.

2.4.2. Saliency Map Filtering Background Noise. We measure the salient scores of a region by calculating the color and spatial distance between the regions obtained by superpixel segmentation which match the foreground clue and the regions obtained by superpixel segmentation that do not match the foreground clue. The formula is as follows:

$$S_i = \sum_{j\neq i, j\in FS} \frac{1}{d(a_i,a_j) + d(k_i,k_j)}, \tag{6}$$

where FS is the set of foreground clues, $d(a_i,a_j)$ represents the color distance between the regions obtained by superpixel segmentation which match the foreground clue and the regions obtained by superpixel segmentation that do not match the foreground clue, and $d(k_i,k_j)$ represents the spatial distance between the regions obtained by superpixel segmentation which match the foreground clue and the regions obtained by superpixel segmentation that do not match the foreground clue. In order to avoid the self-similarity of zero in the foreground clue, we calculate the salient score using the following formula:

$$\begin{aligned} S_i &= \frac{(S_i + (1/|FS| - 1)\sum_{Z\in FS} p(z,i) S_Z)}{|FS|}, \\ p(z,i) &= \begin{cases} 1, & z = i \\ 0, & z \neq i, \end{cases} \end{aligned} \tag{7}$$

where $|FS|$ is the cardinality of the foreground clue set FS, S_i is the salient scores of each region segmented by superpixel. Because our foreground clue consists of pixels which have high salient scores selected by the adaptive threshold method from our saliency map fused with superpixels, extracting foreground clues can filter out those pixels with low salient scores caused by background interference. Thus, a saliency map composed of S_i can filter out the interference of background noise.

2.5. Weighted Fusion of Different Saliency Maps. In order to avoid the weakening effect that extracting salient targets caused by background noise filtering may have, we use weighted fusion [19] between the saliency map (fused with superpixels) and saliency map (based on foreground). The formula is:

$$Q = w_1 S + w_2 D, \tag{8}$$

where w_1 and w_2 are the weights of the saliency map S and D, respectively, obtained by the least squares estimation, and Q is the final saliency map.

3. Experiments

3.1. Datasets. In this section, in order to test and reflect the effect of our model, we select the MSRA dataset, ECSSD dataset, and PASCAL dataset as the processing target dataset. The MSRA dataset contains 5000 images with different complex backgrounds, and each image of the ECSSD dataset has a

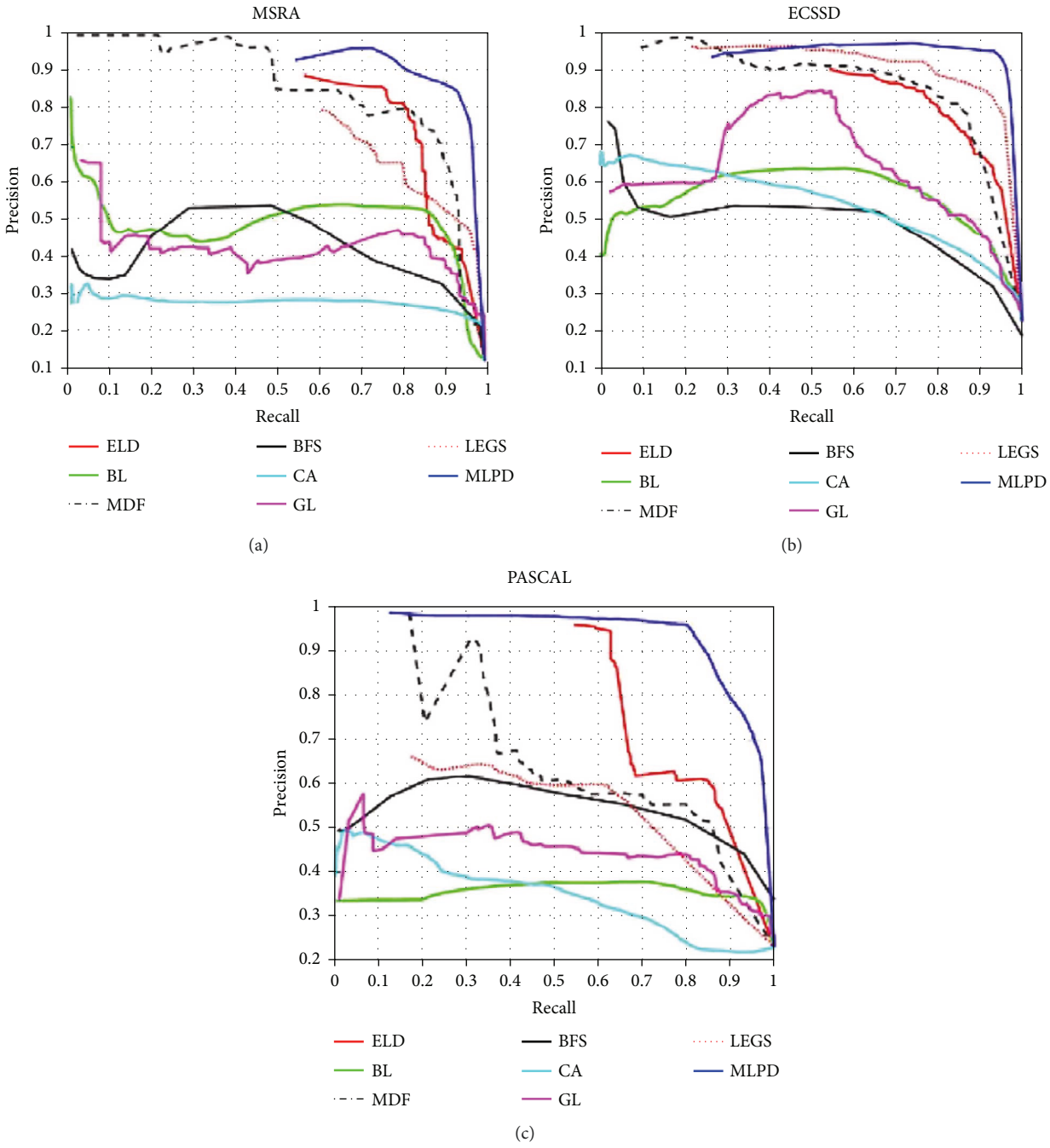


Figure 5: Quantitative model comparisons. This shows PR curves on MSRA, ECSSD, and PASCAL datasets.

well-defined saliency target, PASCAL contains 1000 real-world images which has more than one salient objects.

3.2. Evaluation Metrics. In addition to the PR curve, we use the F-measure score [20] to evaluate the extraction of the saliency target. The F-measure score is calculated as follows:

$$F = \frac{\left(1 + x^2\right) \cdot Precision \cdot Recall}{x^2 \cdot Precision + Recall}, \tag{9}$$

where x^2 is set to 0.3, and Precision and Recall are obtained by the adaptive threshold segmentation method.

3.3. Results. To demonstrate our findings, we compare our MLDP model with 7 other state-of-the-art models: CA [21], BFS [7], BL [22], GL [23], MDF [24], LEGS [25], and ELD [26].

We evaluate the extraction results of the salient targets in these models using different datasets, as shown in Figures 5 and 6.

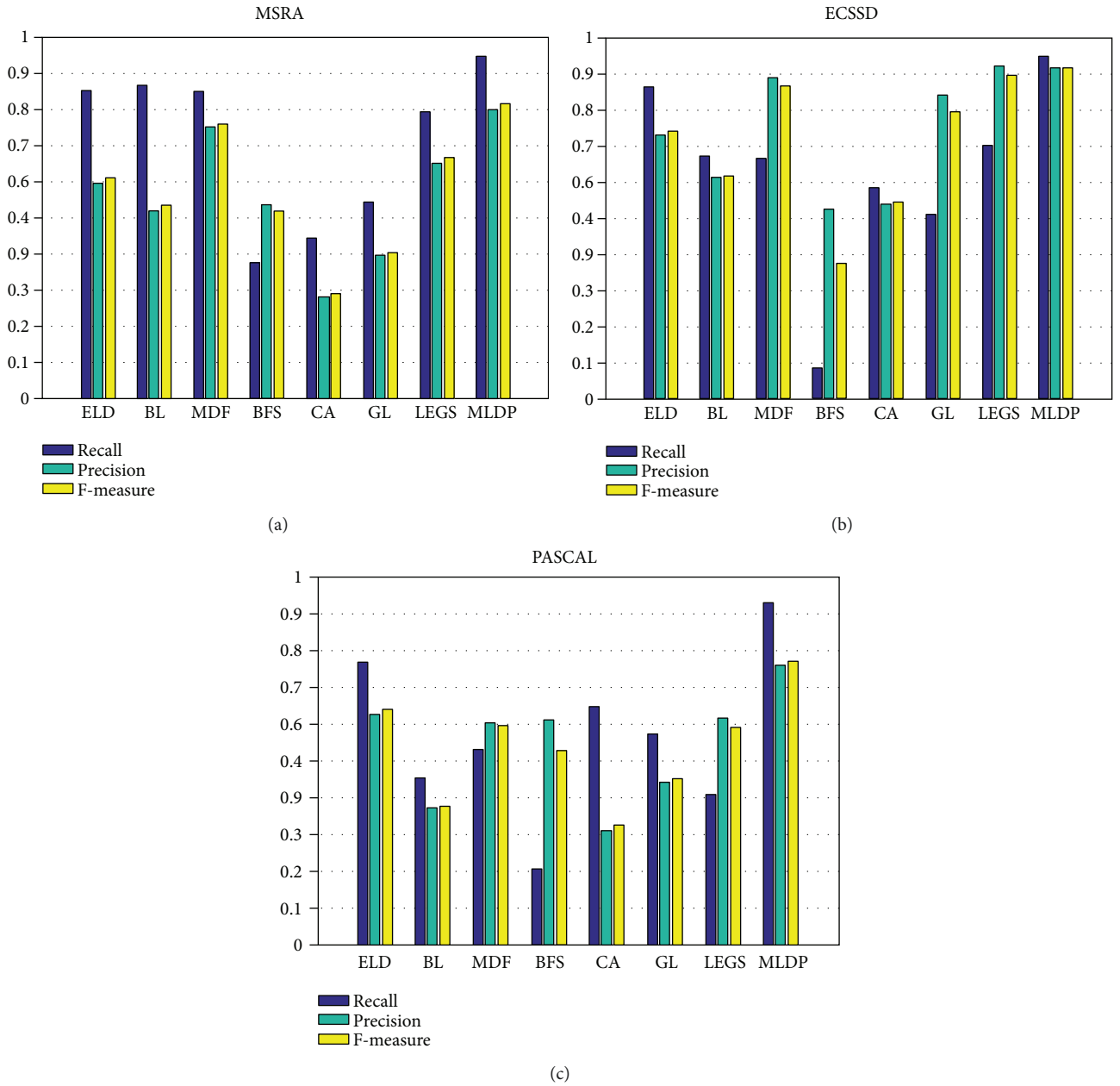


Figure 6: Quantitative model comparisons. This shows F-measure scores on MSRA, ECSSD, and PASCAL datasets.

As shown in Figure 5, our MLDP model achieves good precision and recall across both the MSRA, ECSSD, and PASCAL datasets. Although the other models' recall rate is slightly higher than our MLDP model when the recall is between 0.28 and 0.3 in ECSSD datasets, the MLDP model is better across most other ranges in terms of precision and recall. Thus, our MLDP model has wide-ranging applicability.

We can see from Figure 4 that our MLDP model is better than other models evidenced by the higher F-measure scores. The comprehensive results show our MLDP model is quite effective in recall, precision, and F-measure.

We present a visual comparison of results for each model in Figure 7. As shown, our MLDP model not only accurately locates the salient target but also extracts significant details of salient targets with clear contours, particularly in the case of complex backgrounds such as lines 2 and 8. Most other models confuse the salient targets and the background. Because the MLDP model can eliminate the interference of complex background factors, the extraction of salient targets is better than most other models. Whether the salient targets are small (lines 11) or large (lines 10), our MLDP model has better salient target extraction, especially when the salient targets are large and close to the edge of the image (lines 10 and lines 5). Most other models will be affected by the edges of the image, and this impacts the clarity of salient targets. The MLDP model also show good results for low contrast color (lines 1, lines 3, and lines 6). MLDP not only has a good results on a single salient target but also works well on multiple targets (lines 4, lines 7, and lines 9).

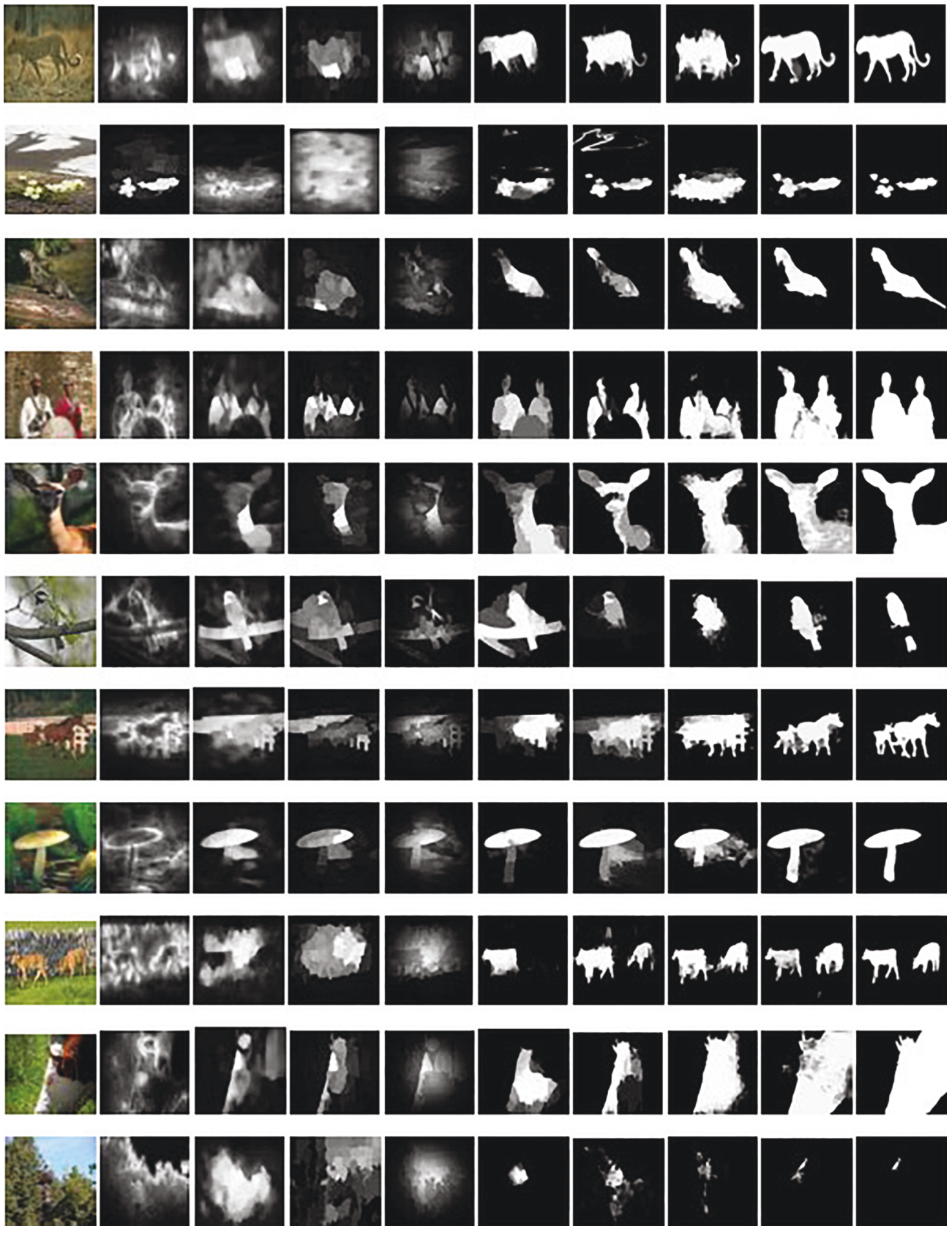
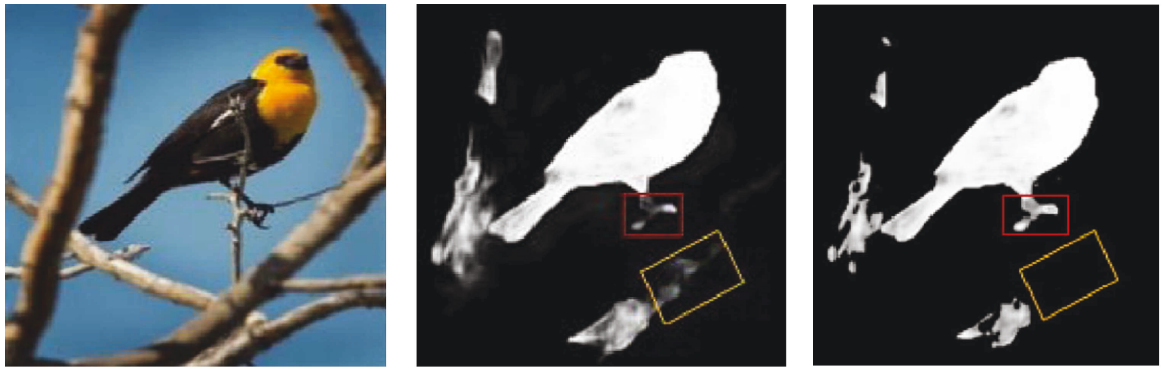


Figure 7: Quantitative model comparisons. The last column is GroundTruth.

(a) Image (b) BFS (c) MLDP

Figure 8: Comparisons between BFS without a deep pyramid and MLDP.



(a) Image (b) MLDP without mapping on superpixels (c) MLDP with mapping on superpixels

Figure 9: Comparisons between MLDP without mapping on superpixels and MLDP.

3.4. Model Component Analysis. One of the advantages of our MLDP model is the way it combines the traditional pyramid model with the deep learning VGG16 model to form a deep pyramid. The VGG16 model is used to extract the high-level features at every level of the pyramid, so the results of the MLDP model are greatly improved. The background noise filtering in our model is also significant, particularly with regard to complex background environments where it can reduce the interference of background factors in the extraction of salient targets.

3.4.1. The Importance of the Deep Pyramid. The deep pyramid is a structure based on the traditional pyramid in extraction of local features and extracts high-level features by the introduction of a deep learning VGG16 model. In order to demonstrate the significance of the deep pyramid, we compare the results of our MLDP model with the BFS model [7], which does not use a deep learning framework to extract high-level features, as shown in Figure 8.

Figure 8 shows that the MLDP model (with a deep pyramid) is better than the BFS model (without a deep pyramid). Although both MLDP and BFS models have a background filtering structure, the BFS model is still based on low-level features extracted by traditional methods. The MLDP model, however, uses a deep pyramid to extract high-level features resulting in greatly improved outcomes.

3.4.2. The Importance of Mapping on Superpixels. We do not extract low-level features directly according to the traditional ideas, but indirectly extract low-level features with mapping on superpixels based on low-level features, we can measure its benefits from Figure 9. From the red block diagram, we can see that the mapping on superpixels can make up for the lack of shape in salient targets. Mapping on superpixels can also reduce the background interference as shown in the yellow block diagram. The mapping on superpixels has two big benefits in our MLDP model.

3.4.3. The Importance of Background Noise Filtering. Because the extraction of salient targets is easily affected by background noise factors, particularly in cases with complex backgrounds, we adapt background noise filtering to eliminate this effect. To illustrate the importance of background noise filtering, we compare the results of a model without background noise filtering and our MLDP model in Figure 10.

Compared with the model without the background noise filtering, the MLDP model eliminates the interference of background noise factors. Figure 10 demonstrates that the

(a) Image (b) MLDP without background filtering (c) MLDP

Figure 10: Comparison between MLDP without background filtering and MLDP.

model without the background noise filtering will reduce the accuracy of salient targets due to the interference from a complex background, and the background interference will appear around the salient targets. Meanwhile, the MLDP with background noise filtering can eliminate these errors caused by background noise factors.

4. Conclusions

In this paper we propose the MLDP model, which is based on a pyramid to extract low-level features with a deep learning model added to extract high-level features. The results of the MLDP model are better than most state-of-the-art methods, and it is able to address the issues of identifying salient targets against complex backgrounds by eliminating the interference of background noise factors.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request. The data sources in this work are public.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (U1762264, 61601203, U1664258, U1764257, and 61773184), National Key Research and Development Program of China (2018YFB0105003), Key Research and Development Program of Jiangsu Province (BE2016149), Key Project for the Development of Strategic Emerging Industries of Jiangsu Province (2016-1094, 2015-1084), Key Research and Development Program of Zhenjiang City (GY2017006).

References

[1] Y. Jae Lee, A. A. Efros, and M. Hebert, "Style-aware mid-level representation for discovering visual connections in space and time," in *2013 IEEE International Conference on Computer Vision*, pp. 1857–1864, Sydney, NSW, Australia, December 2013.

[2] Y. Cai, Z. Liu, H. Wang, and X. Sun, "Saliency-based pedestrian detection in far infrared images," *IEEE Access*, vol. 5, pp. 5013–5019, 2017.

[3] H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao, and D. Li, "Object classification using CNN-based fusion of vision and LIDAR in autonomous vehicle environment," *IEEE Transactions on Industrial Informatics*, p. 1, 2018.

[4] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[5] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *2011 International Conference on Computer Vision*, pp. 2214–2219, Barcelona, Spain, November 2011.

[6] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10–12, pp. 1489–1506, 2000.

[7] J. Wang, H. Lu, X. Li, N. Tong, and W. Liu, "Saliency detection via background and foreground seed selection," *Neurocomputing*, vol. 152, pp. 359–368, 2015.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[9] C. Shen and Q. Zhao, "Learning to predict eye fixations for semantic contents using multi-layer sparse network," *Neurocomputing*, vol. 138, pp. 61–68, 2014.

[10] H. Wang, L. Dai, Y. Cai, X. Sun, and L. Chen, "Salient object detection based on multi-scale contrast," *Neural Networks*, vol. 101, pp. 47–56, 2018.

[11] Y. Y. Zhang, C. Yang, and P. Zhang, "Two-stage sparse coding of region covariance via log-Euclidean kernels to detect saliency," *Neural Networks*, vol. 89, pp. 84–96, 2017.

[12] H. Du, Z. Liu, H. Song, L. Mei, and Z. Xu, "Improving RGBD saliency detection using progressive region classification and saliency fusion," *IEEE Access*, vol. 4, pp. 8987–8994, 2016.

[13] X. Li, D. Li, Z. Yang, and W. Chen, "A patch-based saliency detection method for assessing the visual privacy levels of objects in photos," *IEEE Access*, vol. 5, pp. 24332–24343, 2017.

[14] Y. Cai, X. Sun, H. Wang, L. Chen, and H. Jiang, "Night-time vehicle detection algorithm based on visual saliency and deep

learning," *Journal of Sensors*, vol. 2016, Article ID 8046529, 7 pages, 2016.

[15] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3668–3677, Las Vegas, NV, USA, June 2016.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[17] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence*, pp. 115–141, Springer, Dordrecht, 1987.

[18] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[19] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: a discriminative regional feature integration approach," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2083–2090, Portland, OR, USA, June 2013.

[20] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, Miami, FL, USA, June 2009.

[21] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[22] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1884–1892, Boston, MA, USA, June 2015.

[23] N. Tong, H. Lu, Y. Zhang, and X. Ruan, "Salient object detection via global and local cues," *Pattern Recognition*, vol. 48, no. 10, pp. 3258–3267, 2015.

[24] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5455–5463, Boston, MA, USA, June 2015.

[25] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3183–3192, Boston, MA, USA, June 2015.

[26] G. Lee, Y. W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 660–668, Las Vegas, NV, USA, June 2016.