Depth-aware salient object segmentation

Nguyen Hong Thinh[1], Tran Hoang Tung[2], Ha Vu Le[1]

[1] University of Engineering andTechnology, Vietnam National University Hanoi
[2] University of Science and Technology Hanoi (USTH)

Correspondence: Nguyen Hong Thinh, hongthinh.nguyen@vnu.edu.vn

Communication: received ..., revised ..., accepted ...
Online publication: ..., Digital Object Identifier: ...
The associate editor coordinating the review of this article and recommending it for publication was ....

Abstract– **Object segmentation is an important task which is widely employed in many computer vision applications such as object detection, tracking, recognition, and retrieval. It can be seen as a two-phase process: object detection and segmentation. Object segmentation becomes more challenging in case there is no prior knowledge about the object in the scene. In such conditions, visual attention analysis via saliency mapping may offer a mean to predict the object location by using visual contrast, local or global, to identify regions that draw strong attention in the image.However, in such situations as clutter background, highly varied object surface, or shadow, regular and salient object segmentation approaches based on a single image feature such as color or brightness have shown to be insufficient for the task. This work proposes a new salient object segmentation method which uses a depth map obtained from the input image for enhancing the accuracy of saliency mapping. A deep learning-based method is employed for depth map estimation. Our experiments showed that the proposed method outperforms other state-of-the-art object segmentation algorithms in terms of recall and precision.**

Keywords– **saliency map, depth map, deep learning, object segmentation.**

1 Introduction

Object segmentation has been studied for decades. Many researchers have pointed out that it is hard to separate unknown objects from images of complex scenes because it can not rely on pre-existing object models to detect and to split the object out of the background. A recent approach is to utilize visual attention information.

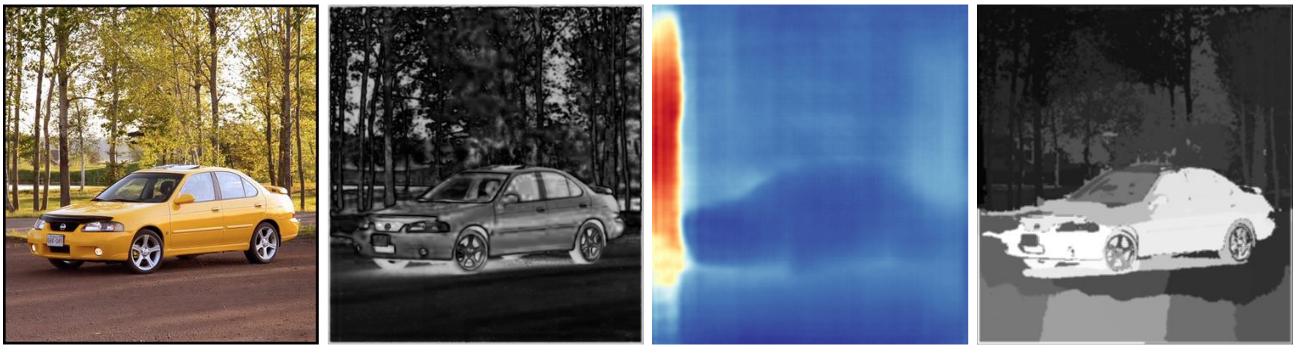
Visual attention is an inherent and powerful ability of the visual system which helps human quickly capture the most conspicuous regions of a scene [1]. It reduces the complexity of visual analysis and makes the human visual system considerably efficient in complex environments. Based on this hypothesis, the object of interest can be detected by finding regions with stronger attention in the image. The level of visual attention at every pixel in the image is given by a weight matrix called saliency map. Saliency map has shown to be beneficial to the detection and segmentation of unknown objects in images [2–8].

In general, saliency-based object segmentation approaches consist of two phases:

- **Object detection via saliency mapping**: Firstly, from the input image, the saliency map is computed in order to locate the objects in a scene, usually at positions with hight saliency weights. Saliency computation algorithms can be roughly divided into*the bottom-up* and *the top-down* approaches. The *bottom-up methods* [2, 3, 6–8] focus on low-level cues like color contrast and luminance contrast. The *top-down methods* [4, 5], on the other hand, are often task-driven. They utilize supervised learning with high-level cues, such as shape learning, categories learning, and etc. Recently, deep learning is also applied for saliency map computation [9–11]. Despite their effectiveness, the *learning-based top-down methods* have limited use due to their need for large sets of training data, especially labeled ground truth images. In this paper we focus only on unsupervised learning-based methods.
- **Segmentation**: Secondly, from the obtained saliency map, a binary mask is calculated to mark whether each pixel belongs to the object or the background. A simple way to do it is to set a threshold on the saliency map. However, using thresholds often fails to identify the exact boundaries of the objects, thus it requires another extra segmentation step using such algorithms as Mean-shift [12], Grab-cut [13], and Saliency-cut [7] in order to improve the accuracy of object segmentation. That makes the process become more complicated.

Saliency-based unknown-object segmentation is difficult because the saliency computation relies mostly on unsupervised perceptual cues which give low accuracy in complex scenes. A number of computation models have been proposed to improve the accuracy of salient object segmentation, by using additional information such as location[7], shape prior [14, 15], or contextual information[3, 15].

More recently, there are several works on 3D saliency and RGB-D saliency [16–18] that proposed to use depth information to to detect and to extract objects out of images. These studies show that using depth information may improve saliency object detection and



(a) Input image (b) RGB-saliency map (c) Depth map (d) Depth-based saliency

Figure 1: Illustration of our proposed saliency-from-depth approach: an accurate saliency map for object segmentation obtained by using information from the RGB-saliency map and depth map. Both the RGB-saliency map and depth map are computed from the original input image.

segmentation even in cases when the object appears very similar to the background. In addition, object boundaries can be recovered from the depth channel.

Motivated by such results, we propose a simple and efficient method that estimates depth-like information and then uses estimated depth information to improve the accuracy of salient object segmentation. Our proposed idea is shown in Fig.1. The difference between ours and other methods [16, 17] is that we compute a depth map directly from the 2D input image without the need for precise depth information from special hardware such as 3D cameras or depth sensors. Our proposed method, thus, can be applied for normal RGB images.

The paper is organized as follows. Related works on visual saliency, depth map computation, and depth-based saliency computation are reviewed in Section 2. Section 3 introduces our proposed method. Experiments with the proposed method and results are shown in Section 4. Finally, in Section 5 we conclude this work with a brief discussion about the future directions of the current work.

2 Literature review

2.1 Saliency map computation

Based on the understanding of human visual attention [1], Itti et al [2] was first to present a theoretical framework for saliency map computation. In that study, the saliency map is calculated by combining *local contrast* feature maps of color, intensity, and orientation. Then, the final saliency value at each pixel position is determined by merging all the feature maps using the “winner take all" method. Motivated by the work of Itti et al, many saliency map models which are based on different computational paradigms were introduced. A recent survey of popular saliency map computation approaches can be found in [8].

Saliency computation methods can be divided into four main categories:

- Contrast based methods: exploit visual contrast cues, i.e., salient objects are expected to exhibit high contrast to the background within certain context [2, 5, 7, 20–22]. The contrast cues could be local or global. Local methods compute the contrast within a small neighborhood of pixels by using color difference [20] or shape/edge difference [5]. Different from the local methods, global methods produce the saliency map by estimating the contrast over the entire image. They consider statistics of the whole image and rely on image features such as intensity contrast [21], global color histogram contrast [7], or fusion of color, luminance, texture, and depth contrast features [22].
- Spectral methods: estimate the saliency map based on spectral analysis using amplitude spectrum [23], or both phase and amplitude spectra [24], or HSV image and amplitude spectrum [25].
- Spatial context-based methods: integrate location information in computing the saliency map [3, 6, 7, 26], based on the assumption that spatial information has an important role in locating the object in the scene [26, 27].
- Depth-based methods: use depth feature in 3D images as a cue to improve the accuracy of the saliency map. Some remarkable works are [16, 21, 22, 28–31]. In [16], Ciptadi et al. proposed an RGB-D saliency computation algorithm which constructs a 3D layout and the shape features from depth measurements. 3D salient object detection algorithms in [21] calculate the contrast regions of the depth map and background and orientation priors, then reconstruct the saliency map globally. In [30], color and depth contrast features are used to generate saliency maps, then multi-scale enhancement is performed on the saliency map to further improve the detection precision. Xue et al. [31] proposed using manifold ranking to fuse RGB and depth saliency maps.

2.2 Depth map estimation from a single image

Depth information, in general, may be obtained by using special hardware such as depth sensors, stereo cameras, structured light cameras [32], or by applying depth reconstruction techniques such as depth multiple

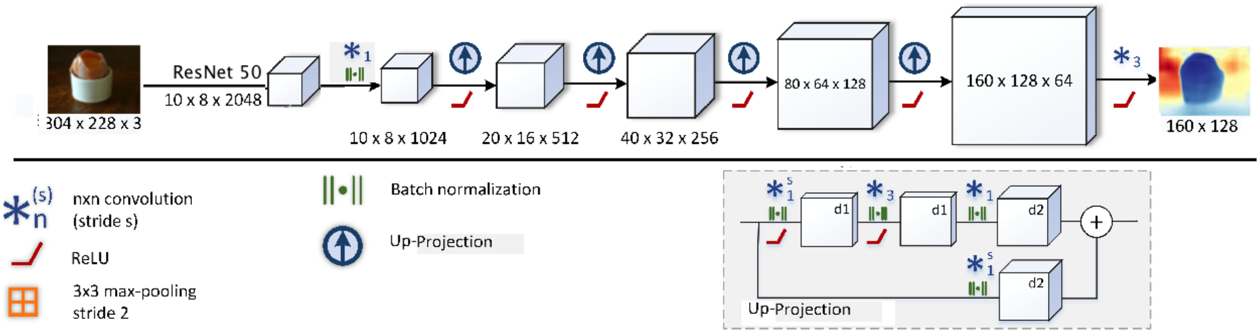


Figure 2: Architecture of the neural network proposed by [19], used in this paper in order to estimate depth from the input RGB image.

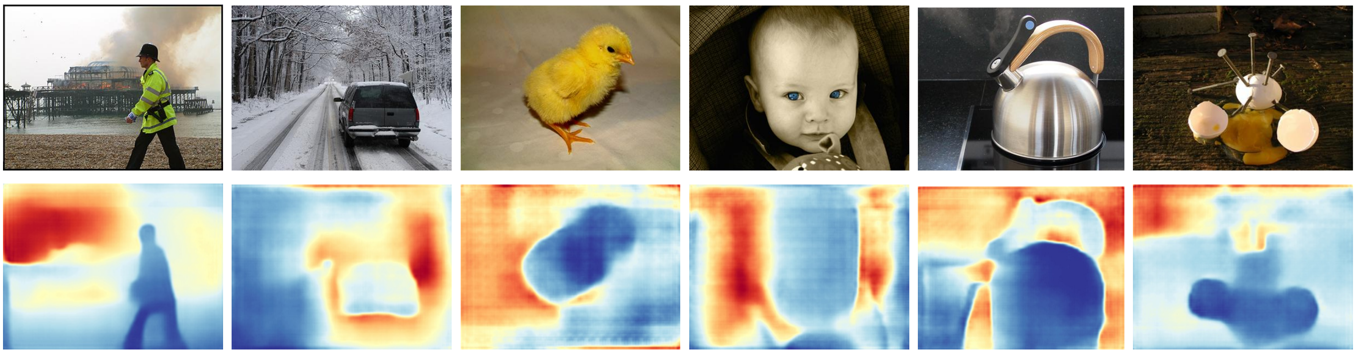


Figure 3: Depth map prediction on the MRSA-B dataset using our retrained CNN model. The first row shows the original images, the second row shows the corresponding depth maps.

views of a scene [33], depth from motion on video sequences [34], and depth from imaging conditions (i.e, shading, defocus aberration,...) [35].

Several methods for depth map prediction from a single RGB image has been proposed [36–40]. For indoor images, [36, 38] used geometric cues for reconstructing the spatial layout of cluttered rooms such as walls, ceilings, and floors. However, these models make strong assumptions about the structure of indoor environments, hence they can not adapt when the assumed structure scene is unfit for the scene.

In case of outdoor images, [41] proposed a method to categorize image regions into geometric structures (i.e., ground, tree, sky, and etc.), which they use to compose a simple 3D model of the scene. The model was later improved by [39, 40] by incorporating a broader range of geometric subclasses, or information of semantic classes.

Saxena et al. [37] are among the first authors to propose a method to estimate depth applicable for images of both indoor and outdoor scenes. They applied supervised learning with linear regression on a training set of RGB-D images and the Markov Random Field to predict the value of the depth map as a function of the image.

Several other machine learning based methods for depth computation have been proposed recently. [42] introduced a depth transfer model which relies on feature-based matching between input RGB images and RGB-D training images. [43] presented a “learning from examples" method to estimate depth from correspondences between RGB and RGB-D images. The main drawback of these methods [42, 43] is that they always need the RGB-D training set for matching when estimating the depth map for an input RGB image. Leterly, deep learning techniques have shown remarkable advances in computer vision, and several works have also proposed to apply deep networks to predict the depth information [19, 44, 45]. Deep learning methods are complicated in the training phase, but after the weighted graph has been obtained, it is efficient in predicting depth-like image of the input image. In this work, we use a pre-trained deep learning model for estimating a type of depth-like information.

3 Proposed method

The idea of our method is shown in Fig. 4. It includes four main steps: compute saliency map from input RGB image (RGB-saliency), estimate depth information from the input using deep learning technique, verify the reliance of obtained depth image, and fuse the depth map (if be reliable) with saliency map to obtain high accuracy saliency map. In this section, we first detail the method to estimate depth using a pre-trained CNN model and then present the method verify the confident of the predicted depth based on quantized depth contrast. Finally, we introduce the method to combine RGB saliency map and a depth map since the

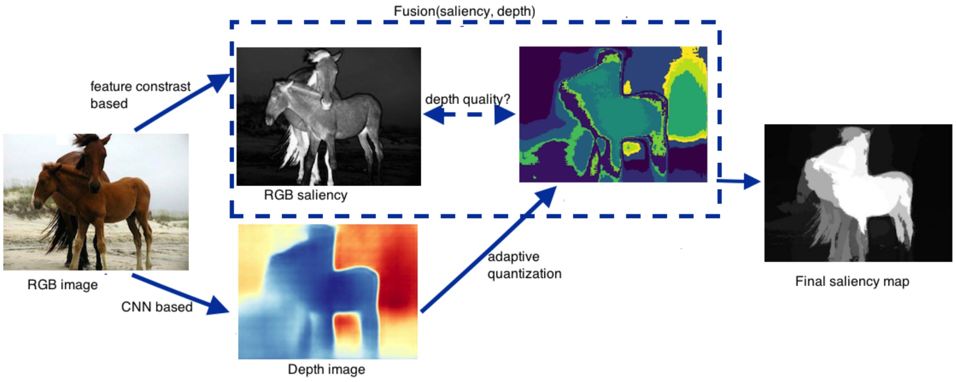


Figure 4: Frame work of our proposed method

depth image is confirmed correctly estimated.

3.1 Estimating depth information

As mention before, in this research we intend to use the deep learning technique to estimate depth-like information of an input image. We use the CNN model proposed by [19] to compute the depth map from the RGB input image. As shown in Fig. 2, the architecture of the network model builds upon the pre-trained ResNet-50 without the last fully-connected layer and the polling layer. Since the ResNet model introduced skip layers that by-pass two or more convolutions and are summed to their outputs, including batch normalization after every convolution, using ResNet-50 structure makes it possible to create much deeper networks without facing degradation or vanishing gradients [19]. The extracted output features have the dimensions of 10x8x2048. Moreover, the network also consists of five up-projection blocks in order to obtain the depth map with a higher resolution. The obtained final depth map has the sizes of 128x160. To evaluate the performance of the CNN model, they trained and tested the neural network on the NYU Depth Dataset V2 [41]. The NYU Depth Dataset V2 is a 4K indoor scene dataset, captured with Microsoft Kinect. In this research, we mean to obtain the network model may adapt to various types of input images; for that reason, we fine-tuned the model by re-training it with a combination of two other datasets: KITTI dataset which incorporates street likes scene depth map information [46], and an RGB-D object detection dataset [29]. Since we obtained the weighted graph of the CNN model, it can directly apply to predict depth image of an RGB image. The implementation is quite fast. Normally, it takes a second to compute a depth matrix for 300x400 input image. Fig. 3 shows several results of depth maps successfully computed by using this CNN model.

3.2 Saliency map computation with depth cue

In our proposed method, we intend to incorporate information from the depth map to improve the accuracy of the saliency map. As we observed before, the depth map, in some cases, is inaccurately estimated. The quality of predicted depth map must be verified first. In order to do this, we proposed to used the global depth contrast. Assume that objects in images are captured with cameras focusing on them, meaning that an object usually appears in front at the center of the image. Its depth, if any, is usually smaller than that of the surrounding area. For easy of computation, we first normalized depth values from 0 to 255 levels; then do adaptive quantization on the depth map. As consequent, the areas with similar depths will be represented by the same values. Using the calculated information from the RGB saliency map, we can have the prediction of the object. Based on that, compare the depth value in the area to the depth values in the vicinity. We can check that the results are accurate or not. If not; resulting saliency map is RGB saliency. In case we have gained reliable depth information, we start apply features contrast on color cue, intensity cue and depth cue. As suggest by [7], using all colors are too expensive in computation and may reduce the performance of calculating color contrast in the image. So, we apply an adaptive quantization on both color map and depth map. We chose 80 most frequent colors and and 16 levels for depth values. The quantization color image and depth map then are used to compute saliency map, and depth-based contrast map based on global contrast method. We apply the average pooling to fuse the values on saliency map and depth-based contrast map. For depth cue, further, we apply object clustering (two clusters for object and non-object) on depth point cloud. In final, we got the binary mask correspond for object boundary based depth cluster.

4 Experiment, Result and Analysis

In this section, we assess our method for saliency object segmentation. The performance is evaluated on **MSRA10k** dataset. This dataset is introduced by [7], contains 10.000 images with large variation themes. The images including in-door, out-door, animal, naturally scenes.The saliency objects are manually segmented for all images in the dataset.

4.1 Evaluation

There are several measures for evaluating a saliency object detection model, usually based on counting

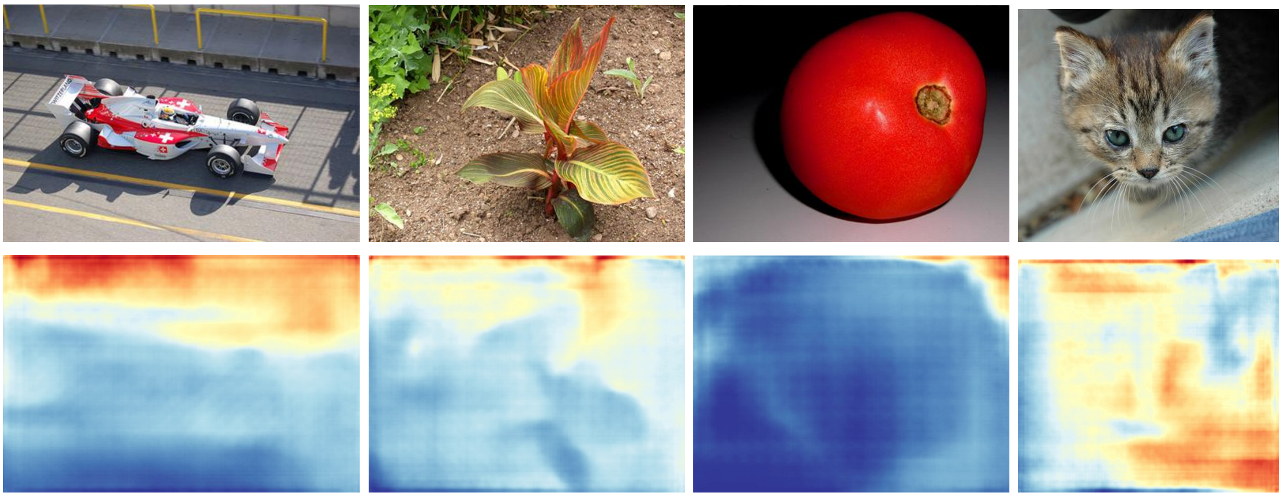


Figure 5: Depth map prediction was performed on MRSA10k dataset using our retrained CNN model, however, for these cases the network failed to predict the depth maps.

the overlap between a tagged regions (i.e ground trush) and the model predictions. Following [6–8, 12], to evaluate the performance of our method, we use standard *Precision-Recall* curves (PR curves), F-Measure and Mean Absolute Error (MAE). At first, it needs to converts the obtained saliency map S into a binary mask M by using a threshold. When the binary mask M is compared against the ground truth G, we can calculate the precision and recall values following:

$$Precision = \frac{|M \cap G|}{|M|}\ ,\ Recall = \frac{|M \cap G|}{|G|},$$

resulting in a pair of *Precision* and *Recall* values. A *Precision-Recall* curve is then obtained by varying the threshold. Furthermore, it is also can chose a adaptive threshold. [12] proposed the image-dependent adaptive threshold for binarizing saliency map S, which is computed as twice as the mean saliency of S:

$$Threshold\ T = \frac{2}{WxH} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x,y),$$

where W and H are the width and the height of the saliency map S, respectively.
Second, since high *Precision* and high *Recall* are both desired in many applications, the F-Measure is proposed as a weighted harmonic mean of both *Precision* and *Recall*:

$$F_\beta = \frac{(1+\beta^2).Precision.\ Recall}{\beta^2 Precision + Recall},$$

where β^2 is set to 0.3 as suggested in [12] to weight *Precision* more than *Recall*.

4.2 Comparison with the State of the Art

We compare our proposed saliency model with a number of existing state-of-the-art methods, including the Spectral Residual approach**(SR)** [23], Spatial Weighted Dissimilarity approach **(SWD)**, Histogram Contrast approach **(HC)** [7], Context-Aware saliency **(CA)** [3], Maximum Symmetric Surround approach **(MSS)** [47], Context-Based and shape prior approach **(CB)** [15], Segmenting saliency approach **(SEG)** [48]. A visual comparison is given in Fig. 7. The figure show the sample results of proposed method and nine other comparison methods. As can be seen, our method performs well compared with the other, in a variety of challenging cases, e.g., indoor scene, natural scene, object scene, thank to depth-like information obtained from the image. As part of the quantitative evaluation, we then evaluate our method using precision-recall curves. As shown in the Fig 6, in general, the proposed method achieves the highest precision compare with other method. Since our method is based on HC method if depth is inaccuracy estimated and is combination of depth map and HC-saliency map; so the positive gap between our method curve and HC curve confirm the advantage of using depth-information in order to compute the saliency values.

5 Conclusion

In this paper, we have proposed a simple and effective method to produce accurate saliency map for the purpose of salient object segmentation. The method uses deep learning to predict the depth map from the input image and incorporate such information to compute the saliency values. The experimental results demonstrated that the additional depth information is useful in improving the performance of object segmentation.

References

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

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

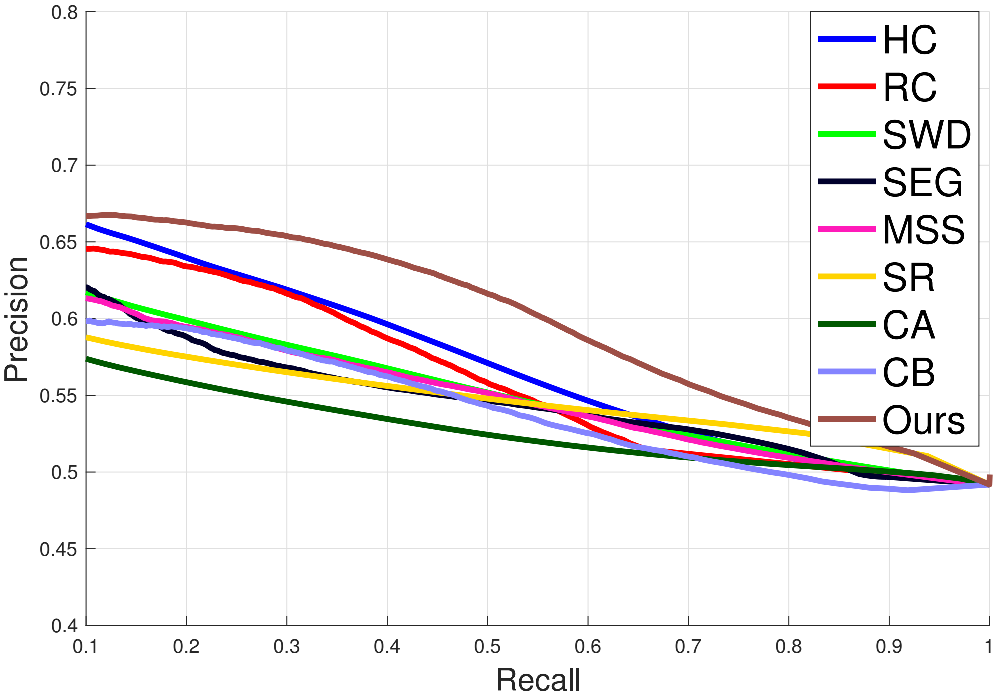


Figure 6: Precision-Recall curves of different saliency detection methods on MRSA10k dataset

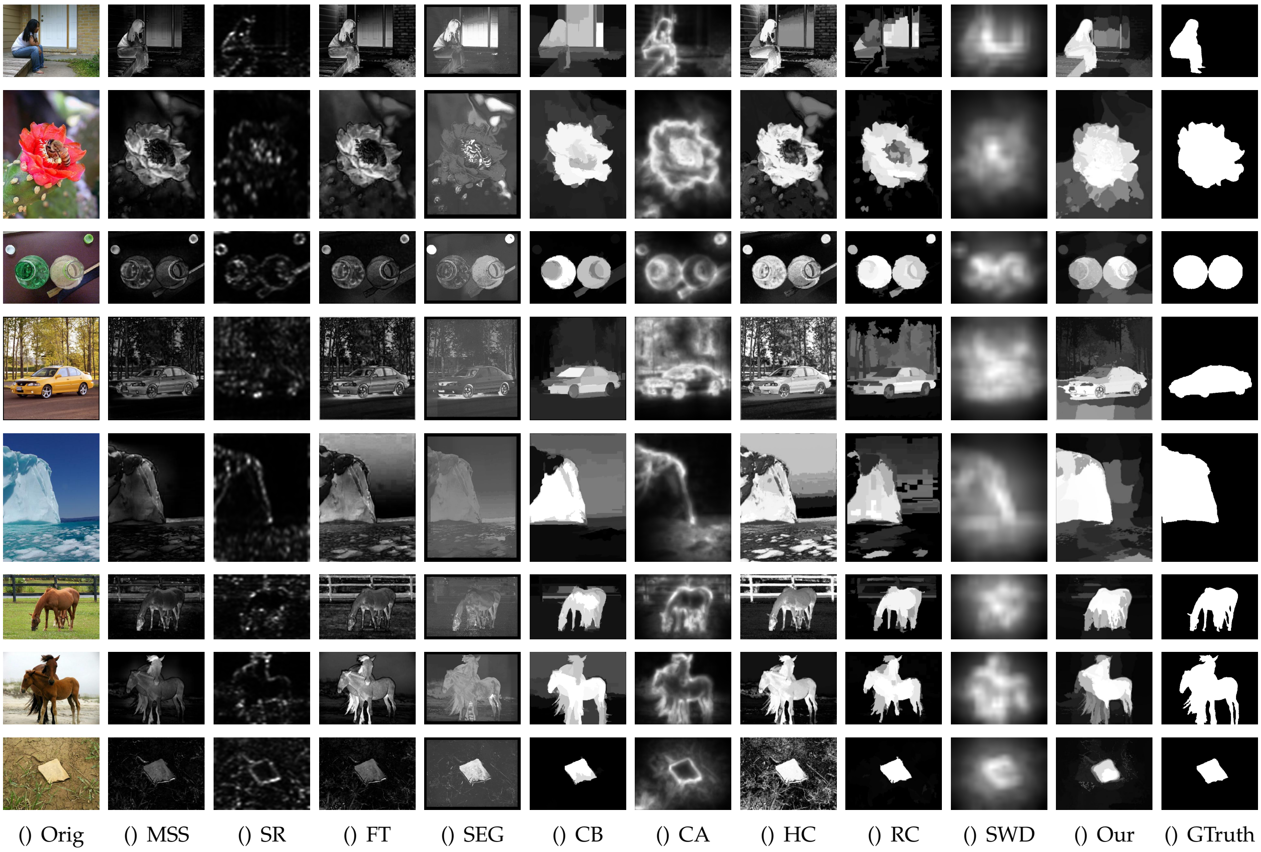


Figure 7: Result of saliency map obtained by several state of the art method and our method.

[3] 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.

[4] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, “Sun: Top-down saliency using natural statistics,” *Visual cognition*, vol. 17, no. 6-7, pp. 979–1003, 2009.

[5] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang,

and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[6] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 733–740.

[7] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. 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.

[8] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[9] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[10] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.

[11] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 678–686.

[12] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned saliency detection model," *CVPR: Proc IEEE*, pp. 1597–604, 2009.

[13] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[14] E. Borenstein and J. Malik, "Shape guided object segmentation," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 969–976.

[15] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior." in *BMVC*, vol. 6, no. 7, 2011, p. 9.

[16] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency." Georgia Institute of Technology, 2013.

[17] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth." in *BMVC*, 2013.

[18] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813.

[19] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 239–248.

[20] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2006, pp. 155–162.

[21] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang, "Exploiting global priors for rgb-d saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 25–32.

[22] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images." *IEEE Trans. Image Processing*, vol. 23, no. 6, pp. 2625–2636, 2014.

[23] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[24] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Computer vision and pattern recognition, 2008. cvpr 2008. ieee conference on*. IEEE, 2008, pp. 1–8.

[25] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, and C.-W. Lin, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 187–198, 2012.

[26] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Computer vision–ECCV 2012*. Springer, 2012, pp. 101–115.

[27] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3d video," in *International Conference on Multimedia Modeling*. Springer, 2010, pp. 314–324.

[28] J. Wang, M. P. Da Silva, P. Le Callet, and V. Ricordel, "Computational model of stereoscopic 3d visual saliency," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2151–2165, 2013.

[29] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 92–109.

[30] P. Wu, L. Duan, and L. Kong, "Rgb-d salient object detection via feature fusion and multi-scale enhancement," in *CCF Chinese Conference on Computer Vision*. Springer, 2015, pp. 359–368.

[31] H. Xue, Y. Gu, Y. Li, and J. Yang, "Rgb-d saliency detection via mutual guided manifold ranking," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 666–670.

[32] S. Katz and A. Adler, "Depth camera based on structured light and stereo vision," Mar. 8 2012, uS Patent App. 12/877,595.

[33] P. Chatterjee, G. Molina, and D. Lelescu, "Systems and methods for determining depth from multiple views of a scene that include aliasing using hypothesized fusion," Mar. 21 2013, uS Patent App. 13/623,091.

[34] L. Matthies, T. Kanade, and R. Szeliski, "Kalman filter-based algorithms for estimating depth from image sequences," *International Journal of Computer*

Vision, vol. 3, no. 3, pp. 209–238, 1989.

[35] Y. Y. Schechner and N. Kiryati, "Depth from defocus vs. stereo: How different really are they?" *International Journal of Computer Vision*, vol. 39, no. 2, pp. 141–162, 2000.

[36] E. Delage, H. Lee, and A. Y. Ng, "A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2418–2428.

[37] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2009.

[38] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering the spatial layout of cluttered rooms," in *Computer vision, 2009 IEEE 12th international conference on*. IEEE, 2009, pp. 1849–1856.

[39] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1253–1260.

[40] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 89–96.

[41] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[42] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 978–994, 2011.

[43] J. Konrad, M. Wang, and P. Ishwar, "2d-to-3d image conversion by learning depth from examples," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 16–22.

[44] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5162–5170.

[45] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2800–2809.

[46] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[47] R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," in *Image processing (ICIP), 2010 17th IEEE international conference on*. IEEE, 2010, pp. 2653–2656.

[48] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 366–379.