2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Lightweight Contrast Modeling for Attention-Aware Visual Localization

Lili Huang[1], Guanbin Li[1], Ya Li[2] and Liang Lin [1]

Abstract— **Salient object detection, which aims at localizing the attention-aware visual objects, is the indispensable technology for intelligent robots to understand and interact with the complicated environments. Existing salient object detection approaches mainly focus on the optimization of detection performance, while ignoring the considerations for computational resource consumption and algorithm efficiency. Contrarily, we build a superior lightweight network architecture to simultaneously improve performance on both accuracy and efficiency for salient object detection. Specifically, our proposed approach adopts the lightweight bottleneck as its primary building block to significantly reduce the number of parameters and to speed up the process of training and inference. In practice, the visual contrast is insufficiently discovered with the limitation of the small empirical receptive field of CNN. To alleviate this issue, we design a multi-scale convolution module to rapidly discover high-level visual contrast. Moreover, a lightweight refinement module is utilized to restore object saliency details with negligible extra cost. Extensive experiments on efficiency and accuracy trade-offs show that our model is more competitive than the state-of-the-art works on salient object detection task and has prominent potentials for robots applications in real time.**

I. Introduction

Attention-aware salient object detection is the indispensable preprocessing for intelligent robots to understand and interact with the configurations of unknown and complicated environments. It aims at locating the most visually distinctive object regions in images and facilitates a convenient way to direct the agent's attention, which benefits many robotic tasks, such as object grasping (e.g. Fig. 1), manipulation and scene exploration. In case of disaster and emergency, attention-aware salient object detection, as the preprocessing step, is expected to be handled with high accuracy and high efficiency.

With the prevalence of deep convolutional neural networks (CNNs) in computer vision and robot-oriented applications, the performance of salient object detection has dramatically improved. Nevertheless, the performance gain comes at the cost of consuming huge computing resources, which greatly limits its application in the field of robotics. In this paper, we aim at seeking a lightweight network for attention-aware salient object detection by improving the efficiency while maintaining its accuracy as much as possible.

(a) scene image (b) salient object detection

Fig. 1. The application of salient object detection in the field of robot control. By analyzing the most eye-attracting visual objects in the scene, the robot can quickly perform precise grabs and subsequent operational tasks.

Various researches have focused on lightweight model designs, such as exploring new algorithmic architectures through network pruning, connectivity learning and hyper-parameter optimization [1], [2], [3]. However, it would not be appropriate to directly apply them in attention-aware salient object detection because visual contrast is the most significant factor [4], [5], [6] for accuracy improvement, while these models are not tailor-designed for capturing the subtle visual contrast in an image.

Various visual saliency detection approaches are based on local or global contrast cues. In early works, the visual contrast is illustrated by sophisticated hand-crafted low-level features, such as color, intensity and texture. Recently, CNN based models have been employed to obtain high-level semantic features, which is more robust than hand-crafted features, achieving better results than early attempts. Most of these methods infer visual saliency by learning contrast from a single input, and their output is derived from receptive fields with a uniform size. Accordingly they may not perform well enough when handling images with salient objects at different scales. Resorting to a multi-scale fully convolutional network is the most intuitive solution. Furthermore in order to obtain the global context information, the general approach is to expand the receptive field. However, consecutive down-sampling of CNNs makes the resolution of final detected salient object only a very small fraction of the original input image, which is infeasible to accurately locate the salient object.

In this paper, inspired by a recent work [3] which incorporates depth-wise separable convolutions to build lightweight deep neural networks, we propose a lightweight multi-scale network (LMNet) to simultaneously capture contrast information at different scales for saliency detection, and to

*This work was supported in part by National Natural Science Foundation of China under Grant No.61702565, in part by Guangzhou Municipal Universities No. 1201620302 and in part by the Science and Technology Planning Project of Guangdong Province No. 2015B010128009. This work was also sponsored by SenseTime Research Fund. (Corresponding author: Ya Li)

[1]Lili Huang, Guanbin Li, and Liang Lin are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China huanglli3@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn, linliang@ieee.org

[2]Ya Li is with Guangzhou University, Guangzhou, China liya@gzhu.edu.cn

reduce model parameters for efficiency improvement. Our LMNet consists of the basic feature representation, multi-scale visual contrast learning and the lightweight refinement module for fine-tuning. We utilize lightweight bottleneck blocks to learn features, which has been proved to be very efficient in [3]. Taking the output features as input, a multi-scale fully convolutional network with pyramid average pooling is designed to encode rich contextual information for visual saliency reasoning. Finally, we incorporate a lightweight refinement module to capture sharper salient object inference, which gradually recovers the spatial information by resolution expansion. It is worth mentioning that we apply the depth-wise separable convolution to both the multi-scale contrast module and the refinement module, resulting in an efficient end-to-end solution.

In summary, this paper has the following contributions:

- We propose a novel lightweight multi-scale convolutional network for salient object detection which consists of the basic feature representation, multi-scale visual contrast learning and the lightweight refinement module for fine-tuning. The linear bottleneck blocks used in feature extracting and the depth-wise separable convolution applied in multi-scale visual contrast learning and the refinement module tremendously decrease the number of parameters and model size, ensuring the efficiency while maintaining accuracy.
- We introduce a multi-scale contrast module for capturing visual contrast, which works by first encoding rich contextual information with pyramid pooling, followed by feeding feature maps to the depth-wise separable convolution for subsequent model acceleration.
- This work presents intensive experiments on the trade-off of efficiency and accuracy. Experimental results demonstrate its superiority over state-of-the-art works on salient object detection.

II. Related Work

In this section, we discuss the most relevant work on salient object detection and lightweight deep models.

Salient Object Detection. Traditional approaches on visual salient object detection can be roughly categorized into bottom-up and top-down methods. The bottom-up approaches [7], [8], [9], [10] identify contrast of image regions according their low-level visual attributes such as color, intensity, texture and orientation. Contrarily, the top-down approaches [11], [12], [13], [14] normally incorporate high-level knowledge learning to obtain a saliency map. Recently, salient object detection research has been pushed into a new phase by the advancement of deep CNNs. The rapidly sprung up deep models can be further separated into two categories, i.e., patch based multi-stage deep feature leaning approaches and end-to-end FCN-based approaches. The former patch based approaches [15], [16], [17] first partition an image into patches and treat each patches as independent samples for training and testing, resulting in inefficient learning and redundancy among overlapping patches. By contrast, the FCN-based approaches [18], [19], [20], [21] with encoder-decoder structure, have been developed to directly map the whole input image to corresponding saliency map in an end-to-end trainable way. Particularly, some approaches [19], [21] consider multi-scale features extracted from extra stacked convolution layers to capture high-level contrast. Although these end-to-end networks improve accuracy and become the fundamental component, high computational resources requiring is beyond the capabilities of robots.

Lightweight Deep Model Design. Lightweight deep model designs for resource-constrained applications have been a new and fascinating research field for the last several years. Early research utilizes manual tuning parameters and training techniques to optimize networks [22], [23], [24], [25]. Subsequently, many works [26], [27], [28], [29], [30] explore new architecture through network pruning, connectivity learning and hyper-parameter optimization. Recently, a growing number of works [1], [2], [3], [31], [32] are devoted to restructuring the connectivity of the internal convolutional blocks. However, none of them have explored the attention-aware visual localization model, which is essential in the field of robotic cognition. This paper proposes a lightweight but very effective neural network for attention-aware visual localization.

III. Lightweight Multi-Scale Network

As shown in Fig. 2, our designed LMNet architecture is composed of lightweight bottleneck blocks, a multi-scale contrast module, and a lightweight refinement module. Given an input image, the feature representation, i.e. pre-trained tailored VGG, is utilized to extract the feature map. Afterward, the multi-scale contrast module is utilized to capture three-level contrast context priors and further fuse them as the global prior. Then the prior is concatenated with the original feature map to produce the final encoder feature map. It is followed by the lightweight refinement module to generate the final saliency prediction. Therefore, our LMNet directly maps a raw input image to its corresponding saliency map in an end-to-end trainable fashion, through which the multi-scale contrast module, the lightweight refinement module and the feature representation can be optimized simultaneously. The following subsections are dedicated to a detailed description of the proposed approach.

A. Lightweight Bottleneck Module

Since our proposed LMNet is basically constructed from the lightweight bottleneck module to learning features while requiring the less computational cost, we first detail the module in this section. Specifically, the lightweight bottleneck module is characterized by depth-wise separable convolutions [33], linear bottlenecks, and inverted residuals. It first expands the input to high dimension through a point-wise convolution, i.e. 1×1 convolution, and then filters the output with a lightweight depth-wise convolution. Finally, another point-wise convolution is used to reduce

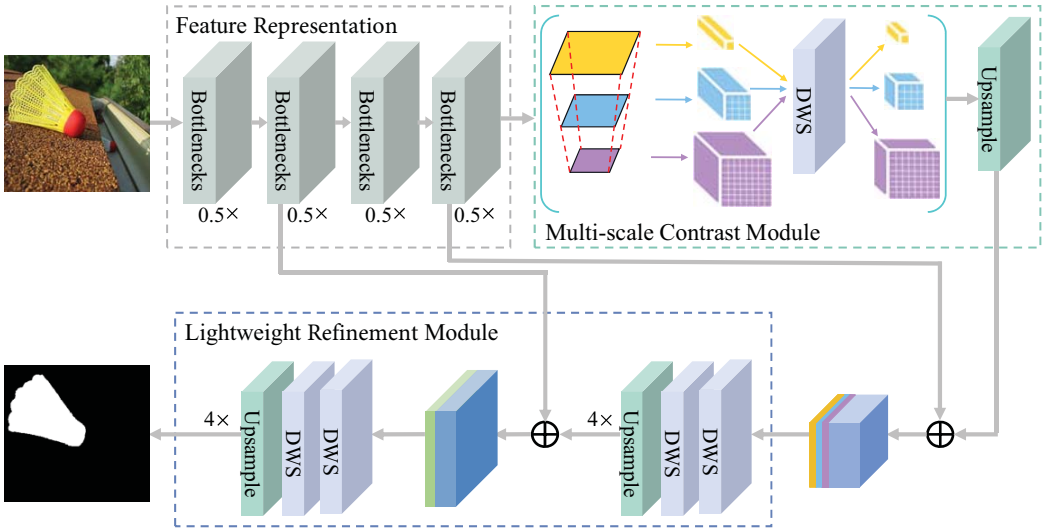

Fig. 2. Overview of our proposed lightweight multi-scale network. Bottlenecks are built up of as in [3]. ***DWS*** and ⊕ denote depth-wise separable convolutions and concatenation operations, respectively

the features back to low dimension again. More specifically, a depth-wise separable convolution is decomposed into one depth-wise convolution and a point-wise convolution. The depth-wise convolutions deploy single convolutional filter into each channel features, while the point-wise convolutions further linearly combine the each channel features to yield new feature representation. Thus, given an input tensor T_i with dimension $C_i \times H_i \times W_i$ and convolved it with kernel size $K \times K$, assuming that the resulted feature map is of size $C_j \times H_i \times W_i$, the corresponding computational cost M of a depth-wise separable convolution is:

$$\begin{aligned} M =& (K^2 \times C_i \times H_i \times W_i) + (C_i \times C_j \times H_i \times W_i) \\ =& (K^2 + C_j) \times C_i \times H_i \times W_i, \end{aligned} \tag{1}$$

on the contrary, the cost R of a regular convolution operation is:

$$R = K^2 \times C_j \times C_i \times H_i \times W_i. \tag{2}$$

Therefore, the computation cost M of a depth-wise separable convolution is much smaller than that of a regular convolution.

The inverted residuals refer to building shortcut connection between the input layer before expanding operation and the layer after dimensionality reduction of filtered features, which are devoted to speed up the procedures of training and inference. Meanwhile, to prevent useful information from being damaged, the bottleneck module eliminates the non-linear activation in the last layers of bottlenecks.

However, both the vanilla CNNs and the lightweight CNNs realize different levels of feature modeling by stacking convolution operations and limited receptive field information to learn the context information of the region and hence can not accurately describe the contrast information of different levels and are arduous to accurately detect salient object of various scales. To address this issue, we propose a multi-scale contrast module to reasonably combine both global and local contextual prior for accurate salient object detection.

B. Multi-Scale Contrast Module

Our proposed LMNet replaces the regular convolutions with lightweight bottleneck blocks to extract feature maps. However, as discussed in the aforementioned section, the network, constitutive of bottleneck blocks, insufficiently incorporates the crucial global contrast prior for the loss of the empirical receptive field. To address this issue, we propose a multi-scale contrast module to reasonably combine both global and local contextual prior for accurate salient object detection.

The spatial pyramid average pooling has been successfully applied to semantic segmentation [34] and image classification [35] tasks where spatial statistics provide a good descriptor for overall image interpretation. We further extend pyramid average pooling to capturing multi-scale contrast context for salient object detection. As shown in Fig. 2, the multi-scale contrast module consists of pyramid average pooling and a depth-wise separable convolution (i.e., a depth-wise convolution coupled with a point-wise convolution).

The pyramid average pooling is built up of three different pyramid sizes, i.e. small, middle and large size. The large-size pooling highlighted in yellow is the coarsest global pooling to generate a single bin output. The following other size pooling partitions the feature map into different sub-regions and forms pooled representation for each corresponding sub-region with bin sizes of 4×4, 16×16 respectively. Furthermore, to lighten the model, we reduce the dimension of different-size output features to 1/3 of the original input one after through one depth-wise separable convolution layer. Compared with that only using one 1×1 convolution to reduce the feature dimension, our strategy using one depth-wise separable convolution can extract the more expressive features and further improve the accuracy performance by

more than 0.5% with the negligible computation cost. Consequently, the outputs are of disparate sizes. Then we directly upsample the shrunken outputs via bilinear interpolation to achieve the final pyramid contrast features with the same resolutions as the one of original feature map. Finally, the original features and the final pyramid contrast features are concatenated to generate the final encoder feature maps.

C. Lightweight Refinement Module

The features from the feature representation are zoomed out with output stride 16. The common decoding technologies are one-step bilinearly upsampling by a factor of equivalent stride, and multi-step bilinearly upsampling via skip connections as done in [36]. Nonetheless, the former one-step bilinearly upsampling technology is too naive to restore object saliency details, while the latter multi-stage skip-connection decoder marginally improves the performance at the cost of overmuch additional computing resources. We thus propose a plain and effective strategy, i.e. lightweight refinement module, as shown in Fig. 2.

We first bilinearly upsample the encoder features by a factor of 4 and then concatenate them with the corresponding low-level features from the feature representation that have the same spatial resolution. The channel number of the encoder features is much larger than that of the low-level features, which overshadows the importance of the low-level features and make the network harder to train. Therefore, we apply a few 3×3 depth-wise convolutions and one 1×1 convolution on the encoder features to reduce their channels while retaining crucial contrast features. After the concatenation, we apply another 3×3 depth-wise convolutions to refine the features and obtain sharper contrast results, and one 1×1 convolution to further lighten the module. In what follows, another simple bilinear upsampling by a factor of 4 is used.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

Datasets. We evaluate the performance of our method on five public visual saliency datasets, including MSRA-B [37], DUTOMRON [38], HKU-IS [15], ECSSD [39], and PASCAL-S [13], all of which are available online and have been widely used recently. MSRA-B [37] contains 5000 images with diverse image contents. Most images in this dataset have only one coarsely annotated salient object. DUTOMRON [38] is another large challenging dataset containing 5168 images, most of which have multiple salient objects in relatively complex and cluttered backgrounds. HKU-IS [15] contains 4447 challenging images, each of which has either low contrast or multiple salient objects. ECSSD [39] contains 1000 semantically meaningful but structurally complex natural images acquired from the Internet. PASCAL-S [13] was built upon the validation set of the PASCAL VOC 2010 segmentation challenge. It contains 850 images with the ground-truth masks labeled by 12 subjects. In our experiments, the threshold is set as 0.5 to obtain binary masks as suggested in [13]. Many images in this dataset have multiple salient objects either with low contrast or overlapping with the image boundary. To obtain a fair comparison with other methods, as done in [20], [40], [15], we combine the training sets of both the MSRA-B dataset [37] and the HKU-IS dataset [15] as our training set for salient region detection. The validation sets in the aforementioned two datasets are also combined as our validation set. Then we directly applied the trained model to test over all of the datasets.

Evaluation Criteria. We evaluate the performance on both accuracy and efficiency. The efficiency is measured by multiply-adds (MADD), actual latency (i.e., running time on GPU or CPU), and the number of parameters as in [3]. The accuracy is evaluated using precision-recall (PR) curves, F-measure and mean absolute error (MAE). Note that the predicted saliency map is converted to a binary mask using a threshold. The precision and recall is calculated by comparing the binary mask against the ground truth. Averaging precision and recall over saliency maps of a given dataset yields the PR curve. The F-measure is defined as

$$F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \tag{3}$$

where β^2 is set as 0.3 to highlight the importance of the precision as suggested in [41], [18]. The maximum F-measure (maxF) calculated from the PR curve is reported. MAE [9] pixel-wisely measures the numerical distance between an estimated saliency map M and the ground truth G,

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |M(i,j)G(i,j)|, \tag{4}$$

where W and H denote the width and height of the saliency map, $M(i,j)$ stands for the saliency value of the pixel at (i,j) and the same for $G(i,j)$.

Implementation Details. Our LMNet is implemented on the tensorflow [42], a flexible open source architecture with strong support for deep learning. In this paper, we use the tailored VGG as feature representation (downsampling the inputs by a factor of 16) and pre-trained over the ImageNet dataset [43]. The expansion rates in the bottleneck blocks are all set to 6 as in [3]. We take the tailored lightweight VGG followed by one-step bilinearly upsampling operations (i.e., stride = 16) as a baseline model (BS). Then, the baseline model integrating with the multi-scale contrast module and lightweight refinement module is serviced as our final model for still image salient object detection when comparing with other benchmarks and performing the ablation study. During training and testing, the images are all resized to 512*512 through zero padding before feeding into the network. We train our framework in an end-to-end manner using RMSPropOptimizer with both decay and momentum set to 0.9. The learning rate is initially set to 0.045 and decayed by 0.9 per epoch. Batch normalization is adopted after each convolution and before activation. The loss function is an image-level cross-entropy loss. Experiments are performed on a desktop with a GeForce GTX TITAN Black GPU and a

3.60GHz Intel processor. Limiting by the memory, the batch size is default set to 6 in our experiment.

B. Comparison with the State of the Art

We compare our LMNet against 7 state-of-the-art salient object detection methods, including MDF [15], MC [17], RFCN [44], DS [45], DCL [21], DSS [18], and PaTS [14]. For fair comparison, we use either the implementations or the saliency maps provided by the authors. However, none is is provided for PaTS [14] except for F-measure values on MSRA, DUTOMRON, and ECSSD datasets. Therefore, PaTS was just compared with others on F-measure.

TABLE II
COMPARISON OF THE SIZE AND THE COMPUTATIONAL COST BETWEEN DIFFERENT NETWORKS. GPU(S) AND CPU(S) RESPECTIVELY DENOTE THE RUNNING TIME (IN SECONDS) ON GPU AND CPU.

Methods	Params	MADD	GPU(s)	CPU(s)
MDF [15]	56.87M	21.68G	29.141	750.768
MC [17]	116.56M	194.95G	2.949	71.545
RFCN [44]	137.70M	181.65G	4.863	40.259
DS [45]	134.27M	180.88G	0.191	4.652
DCL [21]	66.25M	447.91G	0.490	7.692
DSS [18]	62.24M	250.73G	0.737	7.221
OURS	**2.10M**	**5.60G**	**0.024**	**0.362**

We focus on devising a more computationally efficient network for salient object detection through replacing the regular convolutions with the lightweight module, i.e., depth-wise separable convolutions and linear bottlenecks with inverted residuals. Comparisons of the size, the computational cost and running time between above-mentioned different networks are listed in Table II. As can be seen, our model significantly reduces the number of parameters, which is only 1/30 of the MDF [15], and takes a considerably less computational cost of 5.60G multiply-adds. What's more, under lower-configured hardware, our model spends the least running time: 0.024 seconds on GPU and 0.362 seconds on CPU. Thus, benefiting from the lightweight modules, our model yields the dramatic efficiency: significantly fewer parameters, smaller computational complexity, and less running time. Therefore, our model meets the requirement of robot system on efficiency.

Furthermore, our model maintains competitive accuracy. We use PR curves, F-measure and MAE for the quantitative evaluation on accuracy. As shown in Fig. 3, our model gets a higher PR curve than all the other algorithms. The comparison results of F-measure and MAE are illustrated in Table I. Compared to the second best approach DSS, our model increases 1.7% and reduces 0.008 on the average of five datasets for maximum F-measure and MAE respectively. Comprehensively, by virtue of the refinement effect of multi-scale features and lightweight decoding, our proposed model achieves higher maximum F-measure value and lower MAE on all the five datasets at a cost of least memory resource.

The comparison results of accuracy are visualized in Fig. 4. It can be observed that our model generates considerable accurate saliency maps in various challenging cases, e.g., low contrast between saliency and background, multiple disconnected salient objects, and multi-scale salient objects. It is also worth mentioning that thanks to the multi-scale contrast module and lightweight refinement module, our model produces sharper boundaries besides right salient region. These advantages make our results very close to the ground truth and even better than other methods on many items. Therefore, with high efficiency and accuracy our model is qualified to be employed on robot system.

C. Ablation Studies

Our LMNet consists of two important components: multi-scale contrast module and lightweight refinement module. In this section, we show the effectiveness and necessity of these two components.

Effectiveness of Multi-Scale Contrast Module. We compare the saliency map M_1 generated from the baseline model (BS), and the saliency map M_2 from the model, indicated as MSC, that integrates the multi-scale contrast module into the BS using testing images in the MSRA-B dataset. As shown in Fig. 5, our proposed multi-scale contrast module is capable of discovering and understanding subtle visual contrast among multi-scale feature maps. Besides, instead of the common dimensionality reduction method using 1×1, our multi-scale contrast module utilizes depth-wise separable convolutions to reduce channels of each scale feature maps while retaining the most representative features, and yields about 0.93% boost to the F-measure. On the other hand, we also compare the efficiency, i.e., the size, the computational cost and running time, between the BS and MSC. As listed in Table III, compared with the baseline, the multi-scale contrast module increases very small number of extra parameters and a negligible extra time computation cost while considerably improving the performance.

TABLE III
COMPARISON OF THE SIZE AND THE COMPUTATIONAL COST BETWEEN DIFFERENT DESIGN OPTIONS.

Methods	Params	MADD	GPU(s)	CPU(s)
BS	1.83M	4.14G	0.022	0.333
MSC	2.09M	5.50G	0.023	0.340
OURS	**2.10M**	**5.60G**	**0.024**	**0.362**

Effectiveness of Lightweight Refinement Module. The encoder features from our feature representation are computed with output stride =16, thus the BS and MSC both perform upsampling using in-network bilinear interpolation by a factor of 16, whereas our final model with lightweight refinement module adapts two-step upsampling: first bilinearly upsampling the encoder features by a factor of 4 and then concatenating them with the corresponding low-level features from the network backbone that have the same spatial resolution; and finally, upsampling the concatenated features by a factor of 4 again. To better show the strength of our proposed lightweight refinement module, we compare the aforementioned saliency map M_2 directly upsampling with output stride =16 and the saliency map M_3 generated from our final model integrated with the lightweight refinement module using the testing images in the MSRA-B dataset. The results are also shown in Fig. 5 and Table

TABLE I

ACCURACY PERFORMANCE COMPARISON BETWEEN DIFFERENT NETWORKS ON FIVE PUBLIC DATASETS. THE BEST THREE RESULTS ON EACH DATASET ARE SHOWN IN RED, BLUE, AND GREEN , RESPECTIVELY.

Methods	MSRA-B		DUTOMRON		HKU-IS		ECSSD		PASCAL-S	
	maxF	MAE	maxF	MAE	maxF	MAE	maxF	MAE	maxF	MAE
MDF [15]	0.885	0.104	0.677	0.095	0.860	0.129	0.833	0.108	0.764	0.145
MC [17]	0.872	0.062	0.701	0.089	0.781	0.098	0.822	0.107	0.721	0.147
RFCN [44]	0.926	0.062	0.747	0.072	0.895	0.079	0.898	0.097	0.827	0.118
DS [45]	0.856	0.061	0.765	0.070	0.808	0.071	0.810	0.160	0.818	0.170
DCL [21]	0.916	0.047	0.733	0.084	0.892	0.054	0.898	0.071	0.822	0.108
DSS [18]	0.927	0.028	0.760	0.072	0.913	0.039	0.915	0.052	0.830	0.080
PaTS [14]	0.905	-	0.691	-	-	-	0.821	-	-	-
OURS	0.931	0.027	0.798	0.067	0.927	0.034	0.913	0.065	0.862	0.074

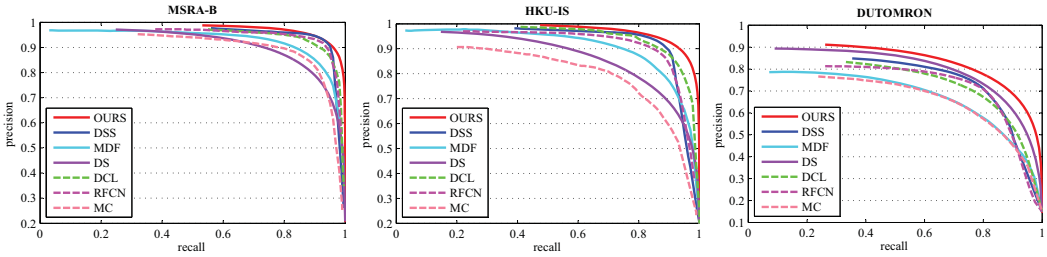

Fig. 3. Comparison of precision-recall curves of 7 salient object detection methods on three popular datasets. Our LMNet consistently outperforms other methods across all the testing datasets.

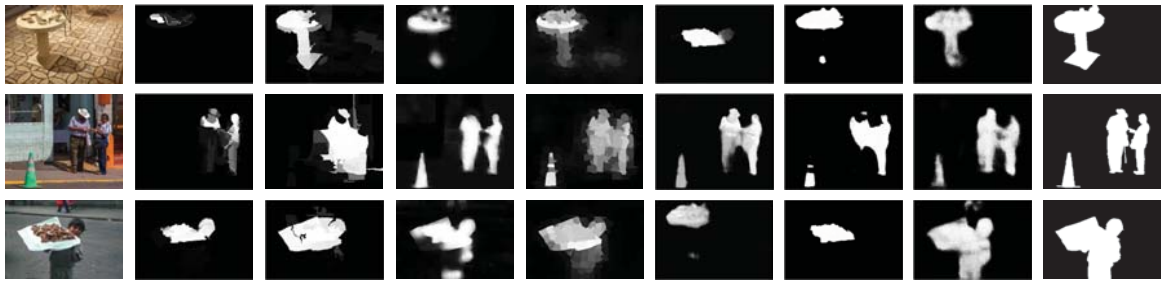

Fig. 4. Visual comparison of saliency maps generated from state-of-the-art methods, including our LMNet. The ground truth (GT) is shown in the last column. Our model consistently produces saliency maps closest to the ground truth.

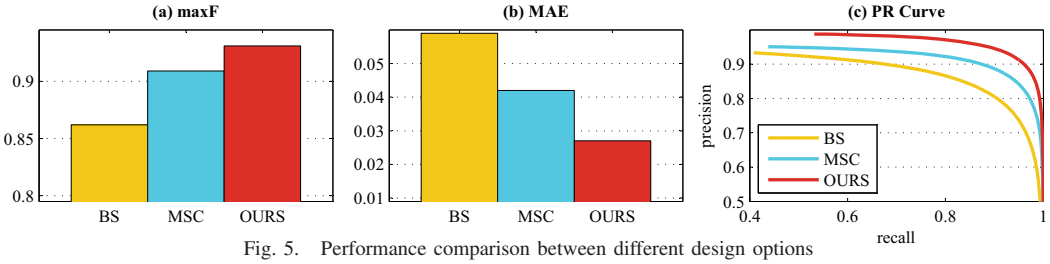

Fig. 5. Performance comparison between different design options

III. They are evident that the lightweight refinement module improves the accuracy of our model by successfully restoring object saliency details at the fairly low additional overhead. Moreover, we adopt a few depth-wise separable convolutions, rather than 1×1 convolutions to reduce channels of feature maps after concatenations and boost F-measure by about 1.3% in comparison with the latter.

V. CONCLUSION

In this paper, we have presented a novel lightweight multi-scale framework for visual localization that is directly applied to mobile robots. Our proposed approach introduces lightweight bottlenecks to significantly reduce the number of parameters and accelerate the process of training and inference. To alleviate the limitation of contrast learning in contemporary CNN, we develop a multi-scale contrast module to rapidly and sufficiently capture low-level and high-level visual contrast. Besides, a lightweight refinement module is incorporated to restore object saliency details with negligible extra cost. Extensive experiments demonstrate the effectiveness of the proposed framework, and has prominent potentials for robots working in real time.

REFERENCES

[1] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[2] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[4] W. Einhäuser and P. König, "Does luminance-contrast contribute to a saliency map for overt visual attention?" *European Journal of Neuroscience*, vol. 17, no. 5, pp. 1089–1097, 2003.

[5] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision research*, vol. 42, no. 1, pp. 107–123, 2002.

[6] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[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] 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.

[9] 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.

[10] L. Jiang, A. Koch, and A. Zell, "Salient regions detection for indoor robots using rgb-d data," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1323–1328.

[11] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1761–1768.

[12] 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.

[13] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[14] D. A. Klein, B. Illing, B. Gaspers, D. Schulz, and A. B. Cremers, "Hierarchical salient object detection for assisted grasping," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2230–2237.

[15] 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.

[16] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[17] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[18] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5300–5309.

[19] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 660–668.

[20] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 247–256.

[21] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," in *Advances in neural information processing systems*, 1993, pp. 164–171.

[27] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in neural information processing systems*, 2015, pp. 1135–1143.

[28] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient dnns," in *Advances In Neural Information Processing Systems*, 2016, pp. 1379–1387.

[29] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[30] K. Ahmed and L. Torresani, "Connectivity learning in multi-branch networks," *arXiv preprint arXiv:1709.09582*, 2017.

[31] S. Changpinyo, M. Sandler, and A. Zhmoginov, "The power of sparsity in convolutional neural networks," *arXiv preprint arXiv:1702.06257*, 2017.

[32] M. Wang, B. Liu, and H. Foroosh, "Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning and spatial" bottleneck" structure," *arXiv preprint arXiv:1608.04337*, 2016.

[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European conference on computer vision*. Springer, 2014, pp. 346–361.

[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[37] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[38] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[39] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended cssd," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 717–729, 2016.

[40] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by ufo: Uniqueness, focusness and objectness," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1976–1983.

[41] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*. IEEE, 2009, pp. 1597–1604.

[42] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[44] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 825–841.

[45] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.