Estimating Objectness Using a Compound Eye Camera

Hwiyeon Yoo, Donghoon Lee, Geonho Cha, and Songhwai Oh

***Abstract*— In this paper, we introduce a new hardware platform that mimics a compound eye of an insect and propose an algorithm to detect objects using it. The compound eye camera has a wide viewing angle and simulates a number of single eyes on its hemisphere. Each single eye is an elementary unit to acquire visual inputs. Visual information from single eyes is hierarchically merged to estimate objectness. We achieve the accuracy of 77.14% on a combined dataset of PASCAL VOC 2012 and COCO-Stuff 10K databases.**

I. Introduction

Detection is a problem of localizing objects in an image. It is one of the most important problems in computer vision since it has valuable applications by itself, e.g., pedestrian detection, and helps to address other problems such as scene understanding, action recognition, and object tracking. Therefore, it has been widely studied and recent methods based on deep neural networks show remarkable results [1]–[3].

In the last few years, cameras with a wide viewing angle are increasingly popular, e.g., omni-directional cameras and 360-degree cameras that capture spherical images instead of planar images [4], [5]. Therefore, conventional object detection algorithms are not directly applicable. In addition, it is more challenging to understand this type of data due to the larger search area on the curved surface.

We address this problem inspired by the mechanism of a compound eye of an insect. A compound eye has a bottom-up structure which merges visual information of single eyes. Single eyes are spread over the spherical surface and each of them independently accepts a small amount of visual inputs. These inputs are then combined to understand the surrounding scene [6], [7].

We develop a hardware, a compound eye camera, which mimics the compound eye of an insect. It emulates a large number of single eyes and each of them captures a two-dimensional low-resolution image, i.e., we approximate the spherical input by a set of piecewise flat images. We assume that each single eye can independently perform simple calculations.

The purpose of this work is to localize objects in an image taken by the compound eye camera as shown in Figure 1. For this objective, we estimate objectness which quantifies how likely it is for a region to cover an object [8]. We propose a compound eye objectness network (CEOnet), which is tailored to estimate objectness on spherical inputs of a compound eye camera.

H. Yoo, D. Lee, G. Cha and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul, Korea (e-mail: {hwiyeon.yoo, donghoon.lee, hwiyeon.yoo}@cpslab.snu.ac.kr, songhwai@snu.ac.kr).

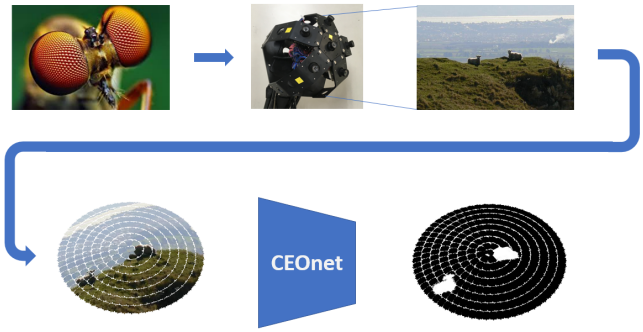


Fig. 1: An overview of the proposed objectness estimation algorithm using a compound eye camera. The developed compound eye camera platform is shown in the middle of the top row. The compound eye camera captures a spherical image of the scene (bottom left). The CEOnet detects objectness on this sphere.

Experimental results demonstrate that the spatial arrangement of single eyes and the size of receptive fields at each layer are important factors for effective objectness estimation. With the configuration of a single eye with 10×10 pixels, we achieve the highest accuracy of 77.14% on PASCAL VOC 2012 [9] and COCO-Stuff 10K [10] datasets.

The remainder of this paper is organized as follows. In Section II, we discuss related work. In Section III, the structure of the compound eye camera is described. In Section IV, the network architecture of the proposed algorithm is discussed. The performance of the our method is evaluated in Section V.

II. Related Work

Most of the recent object detection algorithms are based on deep neural networks, especially convolutional neural networks (CNNs). For example, [1] proposes the faster region-based convolutional neural network (Faster R-CNN), which classifies objects in candidate regions given by a region proposal network (RPN). The RPN predicts locations and shapes of bounding boxes based on predefined anchor boxes. You only look once (YOLO) [2] and the single shot multibox detector (SSD) [3] adopt the same idea that a single network can propose locations of objects and classify them at once. The SSD uses similar structure to the RPN to propose and classify candidate regions. YOLO divides an image into a 7×7 grid cells and each cell proposes and classifies candidate regions. However, these algorithms are designed to deal with only two-dimensional images. Since we take a

spherical image as an input, these methods cannot be used directly to our problem.

There have been several works to make a camera system mimicking a compound eye of an insect [11]–[14]. Most of these works focus on developing a hardware system consisting of a series of small image sensing devices with low power consumption. In addition, applications considered in those works are limited to generate recognizable images from visual information captured by many single eyes. Our algorithm extends the application of these hardwares to a higher-level vision problem, objectness estimation, with an emulated compound eye camera.

Solving high-level vision problems based on a wide view camera, especially using omni-directional cameras is also studied in previous works. Marković [4] applied an omni-directional camera on a mobile robot to perform object detection, tracking, and following a target. Yang [5] attached an omni-directional camera to a smartphone to enable a peripheral vision, e.g., recognizing the device's environment, user's hands and activities.

III. Compound Eye Camera

We have designed a compound camera prototype which consists of six single-lens reflex cameras. The camera modules are on the hemisphere-shaped metal frame. Each camera module can capture 1280×960 pixels images at 24.6Hz. The structure of the proposed camera system is shown in Figure 2 and constructed platform is shown in Figure 1 (top middle). Table I shows the detail specification of a camera used in our compound eye camera system. Due to the practical difficulty of deploying a large number of small cameras, e.g., a dragonfly has about 23 thousands of single eyes [15], there are six cameras that cover hemispherical field of view. Time stamps of all cameras are synchronized with respect to the master camera in the center. These synchronized cameras emulate densely distributed single eyes on the hemisphere based on multi-view geometry as shown in Figure 3.

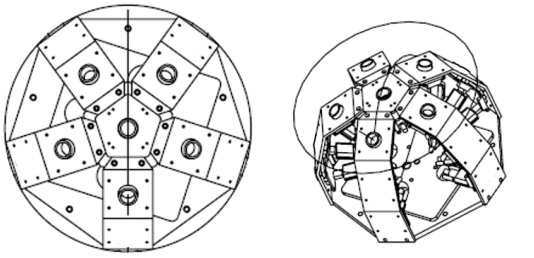
Each emulated single eye captures a rectangular low-resoltuion image, e.g., 10×10 pixels. Emulated single eyes with the same polar angle, θ in Figure 3, are evenly distributed along the surface line of their latitude. Also, the compound eye data has constant angular stride of the latitudes of single eyes. We study the effect of different single eye distributions in the experimental section.

IV. Compound Eye Objectness Network

In this section, we describe each step of the CEOnet. Figure 4 shows an overview of the CEOnet. First, we discuss how to handle spherical input images. Then, a feature encoding method for single eyes is described. Finally, we explain a network for hierarchically merging neighboring single eye information.

A. Compound Eye Data

As a compound eye of an insect consists of a finite number of single eyes, we assume that the spherical input image can be approximated as a locally flat image. By doing so, a single eye corresponds to each flat region. Compound eye data is represented as a tensor $\mathbb{R}^{N \times h \times w \times c}$, where N is a number of single eyes and h, w, and c are height, width, and channel depth of a single eye image, respectively.



(a) Top view (b) Side view

Fig. 2: The structure of the compound eye hardware platform.

TABLE I: Hardware specification of a camera used in the compound eye camera.

Attribute	Value
Diameter	189.98 mm
Angle between the master and slave cameras	25°
Angle between the slave cameras	72°
Resolution of each camera	1280×960 pixels
Joint field of view (FOV)	67°
Unsynchronized frame rate	24.6 Hz

B. Feature Encoding of a Single Eye

Since single eyes capture two-dimensional images, we can apply typical convolutions for encoding. Figure 5 shows the feature encoding network for a single eye. Each single eye is encoded into a d-dimensional vector. The network consists of two convolutional layers with a filter size of 3×3 pixels followed by a fully connected layer. We can boost this calculation by parallelizing computation of d-dimensional vectors of all single eyes.

C. Region Proposals

Conventional detection methods use bounding boxes for region proposals [1]–[3]. However, a bounding box cannot conserve its shape on the hemispherical surface in our problem. Therefore, rather than using rectangular bounding boxes, we define candidate regions based on a set of neighboring single eyes.

The neighbors are determined by k-nearest neighbors (k-NN) algorithm. We measure the Euclidean distance between locations of single eyes for k-NN. By doing so, we can handle irregular distribution of single eyes on the hemispherical surface of the compound eye camera.

D. Region Convolutional Network

The flow of objectness estimation in the CEOnet is represented in Figure 5 and 6. Each layer of the region convolutional network takes N_r regions as an input and each region contains n single eye features.

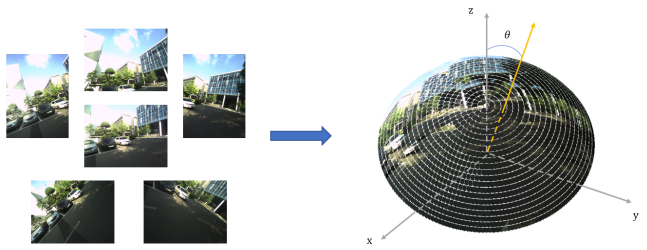


Fig. 3: Six images taken by the compound eye camera are converted on a compound eye with a large number of emulated single eyes.

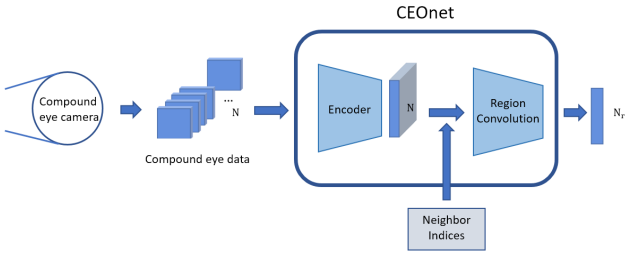


Fig. 4: An overview of the CEOnet. Compound eye data and neighbor indices of each single eyes are fed to the CEOnet as inputs. The CEOnet consist of two networks: an encoding network and a region convolutional network. The encoding network learns a feature embedding of single eyes. The region convolutional network learns to predict objectness based on neighboring single eyes.

The mechanism of the region convolutional network is described in Figure 6(b). As we hierarchically merge information of neighbors based on convolutions, the receptive field gets larger as the network gets deeper.

The objective function of the CEOnet is designed as follows:

$$Loss = \|G - D\|_1 + (0.5 - G)^T D,$$

where G is the ground-truth and D is the estimated objectness score. The first term is a regression loss. The second term penalizes the difference between the binary object masks of G and D after thresholded by 0.5. In addition, it makes D far from 0.5 to reduce the ambiguity of the inferred class.

V. Experiments

A. Dataset

We train and evaluate the proposed network with the PASCAL VOC 2012 and the COCO-Stuff 10K datasets. We merge 2,913 images from the PASCAL VOC 2012 dataset, and 10,000 images from the COCO-Stuff dataset. Each of these images has a ground-truth object segmentation map.

To convert the ground-truth map to the compound eye format, we calculate the ratio of an object mask in each single eye. Therefore, a single eye has an objectness value between $[0, 1]$ that the single eye is more likely to contain an object when the value is close to 1. Objectness of a region is an average of objectness values of single eyes in the region. The ground-truth is defined as an N_r-dimensional vector which represents objectness of N_r regions. Figure 7 shows some examples of the ground-truth conversion.

TABLE II: Results of CEOnet on various configurations.

Single eye size	Accuracy
3×3 pixels	73.98%
5×5 pixels	74.46%
10×10 pixels	**77.14%**
20×20 pixels	75.06%
30×30 pixels	71.87%
Baseline	72.32%
261 single eyes	75.62%
10×10 pixels with L1	73.31%
10×10 pixels with Loss-L1	73.26%
10×10 pixels with L2	72.71%
10×10 pixels with L2+Loss-L1	72.85%

B. Implementation Details

In the CEOnet, all single eyes and regions share network parameters. Therefore, the network parameters which are trained in a certain viewing angle could be applied directly to other viewing angles. Considering this point, we train the CEOnet in a narrow field of view, 30 degrees, that can cover only an image of the master camera. By doing so, we can use existing datasets for training, as if they are taken by the master camera.

When emulating single eye images, there is a flexibility to modulate the size of single eyes. We have tested five different single eye sizes, $3 \times 3, 5 \times 5, 10 \times 10, 20 \times 20$ and 30×30 pixels. Figure 8 shows an example of compound eye data with various single eye sizes.

We can also customize the total number of emulated single eyes and the distance (or angle) between them. Two different configurations of single eyes described in Figure 9 are tested in the experiments. Figure 9(a) has $(2i+1)^2-(2i-1)^2$ single eyes and 9(b) has $4i$ single eyes on the i-th latitude. The latitude index i increases as the polar angle gets larger. In the field of view of the master camera, there are 441 single eyes in Figure 9(a) and 261 single eyes in 9(b).

To evaluate the output of the CEOnet, we measure classification accuracy. We apply a threshold of 0.5 at each region to obtain a binary objectness value.

C. Results

Table II shows results of different eye sizes with 441 single eyes. It shows that it is important to find an appropriate overlap ratio between single eyes by varying their sizes. The configuration with a small single eye cannot cover the entire area of the original scene as shown in Figure 8. Therefore, the accuracy is low when the size of the single eye is too small. On the other hand, as shown in Figure 8(e) and 8(f), too large single eyes are not effective since objects do not have smooth boundaries on the compound eye. In our experiments, a 10×10 pixels single eye configuration has the highest

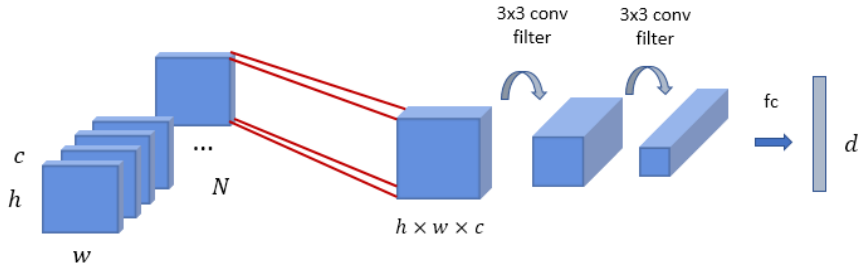
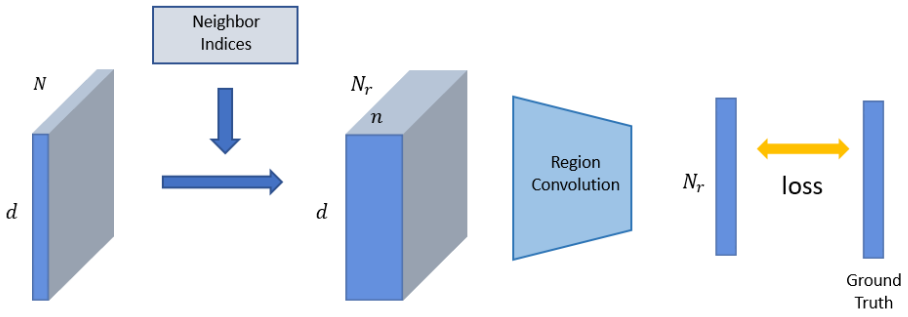
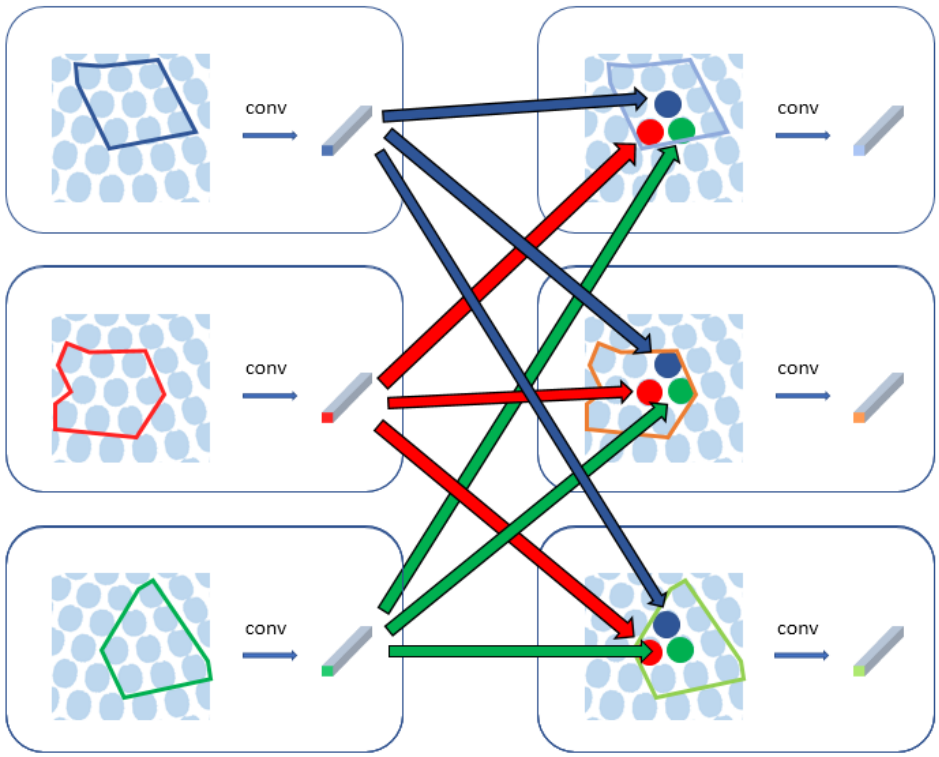


Fig. 5: Compound eye data and the single eye encoding network. The shape of compound eye data is $\mathbb{R}^{N\times h\times w\times c}$. Since every single eyes share parameters of the encoding network, we achieve less number of training parameters. Output of the network contains encoded features of all single eyes with a shape of $\mathbb{R}^{N\times d}$.

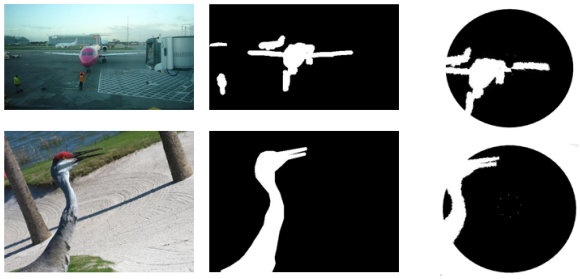


(a)



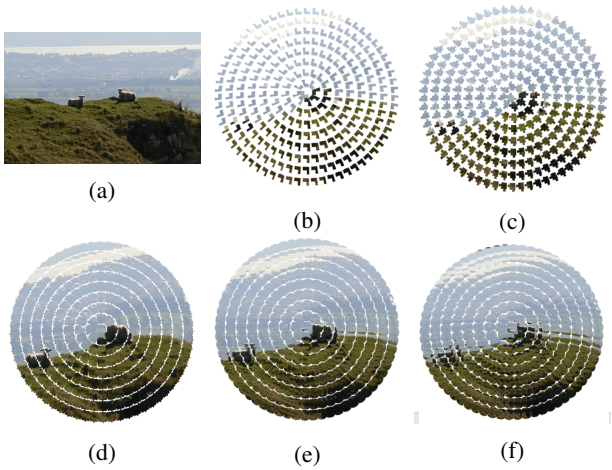
(b)

Fig. 6: (a) Given encoded features of all single eyes, we collect n neighboring features in each region. N_r is a total number of regions. (b) The region convolutional network. Information in each region is merged by a convolution. Note that the shape of regions may irregular unlike conventional CNNs. The final output is objectness estimate at each region.



(a) (b) (c)

Fig. 7: (a) Samples from the PASCAL VOC 2012 dataset. (b) Corresponding ground-truth maps. (c) Converted ground-truth maps on the compound eye.



(a) (b) (c)

(d) (e) (f)

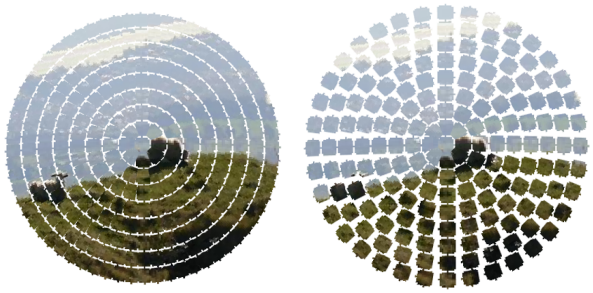
Fig. 8: Compound eye data in various single eye sizes. (b) to (g) represents different single eye sizes of $3 \times 3, 5 \times 5, 10 \times 10, 20 \times 20$ and 30×30 pixels, respectively.

accuracy of 77.14% by covering the entire scene smoothly without missing patches. In this configuration, the angular stride of the latitudes of single eyes is 3 degrees, and 53% of a single eye region overlaps with neighboring single eyes on average. Figure 10 shows some examples of compound eye images from the 10×10 pixels single eye configuration.

For a baseline experiment, we divide two-dimensional 210×210 pixels image into patches that have the same number and size of single eyes, i.e., 21×21 patches with 10×10 pixels. As shown in Table II, the CEOnet achieves better accuracy by using overlapping single eyes.

The density of single eyes also affects the performances. With a coarser configuration, e.g., 261 single eyes of 10×10 pixels, CEOnet achieves a lower accuracy of 75.62% than that of 441 single eyes.

We also do ablation experiments for the proposed loss function. We measure the accuracy of 10×10 pixels single eye model with only the first or the second term of the proposed loss. Moreover, we compare the performance of L1-norm based losses and L2-norm based losses. Through the ablation study, the proposed loss function shows the highest accuracy.



(a) (b)

Fig. 9: Dense and coarse distributions of single eyes. Total number of single eyes in (a) and (b) are 441 and 261, respectively.

(a) (b) (c)

Fig. 10: Some examples of the CEOnet results. 441 single eyes of 10×10 pixels are used. (a) PASCAL VOC 2012 images on compound eye camera. (b) Ground-truth maps of (a) in the compound eye data format. (c) Inferred objectness by the CEOnet.

Table III shows that the accuracy is proportional to the depth of the region convolutional network. This is attributed to the fact that a deeper network has a bigger receptive field.

We measure the computation time of the CEONet on a NVIDIA TITAN X (Pascal) GPU machine with 12GB memory. With the configuration of 441 single eyes of 10×10 pixels, the CEONet takes 13.3ms for inferencing 128 test images. We expect that the CEONet is suitable for real-time applications with this speed.

(a) (b)

Fig. 11: Compound eye data from real world images with six cameras and corresponding outputs of the CEOnet.

TABLE III: Results of the CEOnet on varying network architecture.

Region convolutional network	Accuracy
1 layer	71.56%
2 layer	73.04%
3 layer	74.28%
4 layer	77.14%

Finally, we apply the trained network to images captured by the developed compound eye camera platform. All six cameras are used to generate a compound eye image. The field of view of the generated compound eye data is 67 degrees. Figure 11 shows results of the CEOnet on compound eye images.

VI. Conclusion

We have developed a compound eye camera platform which emulates a compound eye of an insect and proposed a new objectness estimation algorithm, CEOnet, for the compound eye. Through comparative experiments, we have discovered that the 10×10 pixels single eye configuration achieves the best performance with an accuracy of 77.14%. We have also studied the effect of the distribution of single eyes and the number of region convolutional layers. In addition, we have successfully applied the CEOnet to compound eye images captured by the compound eye camera platform. Based on our work, we plan to solve other high-level vision problems using the compound eye camera. It can provide a good light-weight alternative to solve recognition problems for low-cost micro robots.

ACKNOWLEDGMENT

This research was supported by a grant to Bio-Mimetic Robot Research Center funded by Defense Acquisition Program Administration and by Agency for Defense Development (UD160027ID).

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2015.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[4] I. Marković, F. Chaumette, and I. Petrović, "Moving object detection, tracking and following using an omnidirectional camera on a mobile robot," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014.

[5] X.-D. Yang, K. Hasan, N. Bruce, and P. Irani, "Surround-see: enabling peripheral vision on smartphones during active use," in *Proceedings of the ACM symposium on User Interface Software and Technology (UIST)*. ACM, 2013.

[6] M. F. Land, "The optics of animal eyes," *Contemporary Physics*, vol. 29, no. 5, pp. 435–455, 1988.

[7] J. Duparré and F. Wippermann, "Micro-optical artificial compound eyes," *Bioinspiration & biomimetics*, vol. 1, no. 1, p. R1, 2006.

[8] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 34, no. 11, pp. 2189–2202, 2012.

[9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[10] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," *arXiv preprint arXiv:1612.03716*, 2016.

[11] J. Duparré, P. Dannberg, P. Schreiber, A. Bräuer, and A. Tünnermann, "Thin compound-eye camera," *Applied optics*, vol. 44, no. 15, pp. 2949–2956, 2005.

[12] Y. M. Song, Y. Xie, V. Malyarchuk, J. Xiao, I. Jung, K.-J. Choi, Z. Liu, H. Park, C. Lu, R.-H. Kim, *et al.*, "Digital cameras with designs inspired by the arthropod eye," *Nature*, vol. 497, no. 7447, pp. 95–99, 2013.

[13] L. Li and Y. Y. Allen, "Design and fabrication of a freeform microlens array for a compact large-field-of-view compound-eye camera," *Applied optics*, vol. 51, no. 12, pp. 1843–1852, 2012.

[14] J. Duparré, P. Dannberg, P. Schreiber, A. Bräuer, and A. Tünnermann, "Artificial apposition compound eye fabricated by micro-optics technology," *Applied Optics*, vol. 43, no. 22, pp. 4303–4310, 2004.

[15] G. Pritchard, "On the morphology of the compound eyes of dragonflies (odonata: Anisoptera), with special reference to their role in prey capture," *Physiological Entomology*, vol. 41, no. 1-3, pp. 1–8, 1966.