# Light-Weight Semantic Segmentation for Compound Images

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

***Abstract*— The eye structure of insects, which is called a compound eye, has interesting advantages. It has a large field of view, low aberrations, compact size, short image processing time, and an infinite depth of field. If we can design a compound eye camera which mimics the compound eye structure of insects, compound images with these interesting advantages can be obtained. In this paper, we consider the design of a compound camera prototype and low complexity semantic segmentation scheme for compound images. The prototype has a hemisphere shape and consists of several synchronized single-lens reflex camera modules. Images captured from camera modules are mapped to compound images using multi-view geometry to emulate a compound eye. In this way, we can simulate various configurations of compound eye structures, which is useful for developing high-level applications. After that, a low complexity semantic segmentation scheme for compound images based on a convolutional neural network is proposed. The experimental result shows that compound images are more suitable for semantic segmentation than typical RGB images.**

## I. Introduction

Many inspirations of nature have been used to design new technologies. Especially, bio-inspired structures are often used as is to resolve challenging problems. The most popular example is "Shinkansen", which is the fastest train in the world made in Japan [1]. The combination of sudden air flow changes and the high speed of the train makes thunder clap when the train emerges from a tunnel, which is a key problem for making high-speed transportations. It was solved by designing the train mimicking the kingfisher's head. Kingfishers can go through the air and hunt fishes in the water without splash owing to its unique shape of head. The eye structure of insects has many interesting features. The eye of insects, which is called a compound eye, is a remarkably sophisticated structure with a number of advantages. It has a large field of view (FOV), low aberrations, compact size, short image processing time, and an infinite depth of field [2]–[5]. If we can design a compound camera mimicking the structure of the compound eye, we can obtain images with these interesting advantages.

In the meantime, it can be a key issue to make conventional computer vision algorithms suitable for compound images. Among the useful algorithms, we focus on semantic segmentation, which is one of active research topics in computer vision. Semantic segmentation is the problem that not only segments an input image into several coherent regions but also understands the class of each region. It has many applications such as autonomous driving of vehicles and robot visions. On the one hand, semantic segmentation for compound images might have higher value if it can be applied to low-cost mobile robots. It is suitable for reconnaissance robots whose main purpose is to obtain as much information as possible. However, most of semantic segmentation works do not consider the complexity of the algorithm which can be one of the key issues for mobile robots.

In this paper, we consider the design of a compound camera prototype and a light-weight semantic segmentation scheme for compound images. We emulate a compound eye by incorporating several single-lens reflex cameras. Single-lens camera modules are on the hemisphere shaped frame, and they are synchronized to capture synchronized large FOV images. Images captured by camera modules are mapped to the compound eye structure using multi-view geometry to generate compound images consists of hundreds of single-eye images. In this way, we can simulate various configurations of compound eye structures, which is useful for developing high-level applications. We propose a semantic segmentation algorithm for compound images based on convolutional neural networks (CNNs). In designing the proposed CNN network, we have also considered the complexity of the algorithm, which is a key issue for mobile robots. The number of parameters of the proposed CNN network is much smaller than the conventional CNN networks such as FCN-VGG16 and FCN-GoogLeNet [6], and it has a frame rate of 120fps on a GPU. In the experiments, we show the advantage of compound images compared to single images in the application of semantic segmentation. Especially, the fact that neighboring eyes have some common sight helps the semantic segmentation process. We have also tested the proposed framework with various sizes of single images, which can give an insight into the future design of a compound eye image sensor.

The remainder of this paper is organized as follows. In Section II, related work is introduced. The proposed compound camera hardware prototype and the compound mapping scheme are explained in Section III and in Section IV, respectively. The proposed low complexity semantic segmentation network is introduced in Section V. The experimental results are described in Section VI.

## II. Related Work

### A. Compound Eye

There have been several studies that fabricate devices that mimic the compound eye structure [7]–[9]. An almost complete hemispherical form of arthropod-inspired camera was introduced in [7]. They presented materials, mechanics

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

and schemes of integrating arthropod-inspired camera. In [8], a compact compound-eye camera based on a freeform microlens array was introduced. They fabricated on a flat surface, resulting in a narrower field of view than a hemisphere shape. [9] also developed an artificial compound eye with a theoretical analysis. However, they only focused on the manufacturing process of compound eye devices, and no application was proposed.

On the other hand, some applications using compound images have been introduced [10], [11]. [10] developed high-resolution image reconstruction method from several low-resolution images captured by a compound camera. They rearranged pixels in all single-eye images in a virtual image plane consisting of fine pixels. A 3D ego motion estimation method was proposed in [11]. They showed that the geometry of the compound eye is optimal for 3D ego motion estimation, and a linear camera motion estimation algorithm was proposed. However, they focused on low-level applications that are difficult to use for general recognition problems.

### B. Semantic Segmentation

Semantic segmentation has been actively researched [6], [12]–[16]. In the early-stage, conditional random field model [17] was used to classify classes of each pixel [12], [13]. The input images were over-segmented into super-pixels, and the super-pixels were merged using several scene-specific conditional random fields models [12]. The features were extracted with texture-layout filters and they were classified with conditional random fields in [13].

The performance of many computer vision problems including semantic segmentation has been improved based on CNNs [6], [15], [16]. In the CNN framework, features and classifiers are jointly optimized, resulting in improved performance. Deconvolution network based on VGG net [18] was proposed in [15]. They utilized the deconvolution and unpooling scheme for semantic segmentation. However, the input image size of the deconvolution network was fixed, and this problem was solved in fully convolutional networks [6]. The concept of fully convolutional network can be easily incorporated in other network structures, and applied to ResNet [19] and VGG Net [18].

## III. Compound Camera Prototype

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 $1280 \times 960$ size images at 24.6 fps. The blueprint of the proposed device and the implemented hardware are visualized in Figure 1.

We denote each camera module as $C_1, C_2, \cdots, C_6$, and denote the center of the hemisphere as $O$. These notations are shown in Figure 1. The first camera module $C_1$ is at the center, and the other camera modules $C_2, \cdots, C_6$ are uniformly deployed around $C_1$. All camera modules are synchronized, allowing to capture large FOV images. The captured images from the prototype are manipulated to emulate a compound

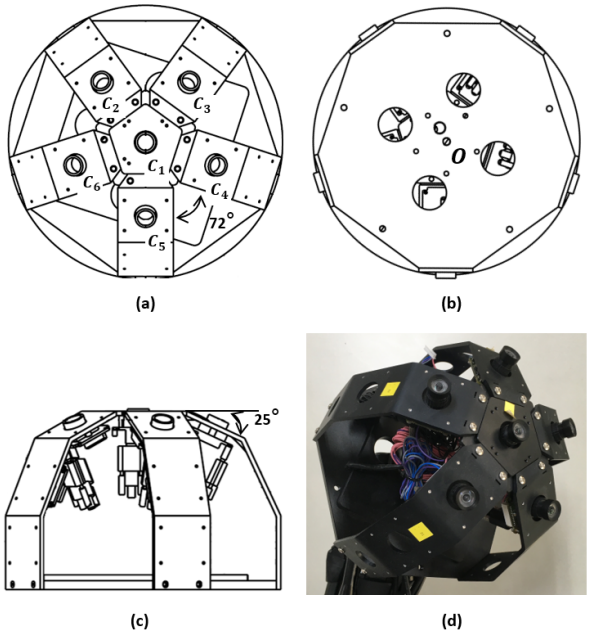


Fig. 1. The blueprints of the proposed device are visualized in (a) top view, (b) bottom view, and (c) side view. An image of the implemented hardware is shown in (d). It consists of six single-lens reflex camera modules, and they are synchronized.

eye. Specifically, images captured by camera modules are mapped to the compound eye structure. To achieve this, we propose a mapping scheme that transforms the images captured from camera modules to compound image space using homography. In this way, we can simulate various configurations of compound eye structures easily compared to fabricated compound eye hardwares. Also, this aspect can be useful for developing high-level applications.

## IV. Compound Image Mapping

We introduce compound image mapping procedure in this section. The core of the scheme is to transform an image captured from a camera module to a single eye image. It is assumed that single eyes are distributed on the hemisphere surface and the captured object is far enough to consider a single image as a plane. Here, the size of a single image is $S \times S$. Let us consider a spherical coordinate system of which the origin, $x$-axis, and $z$-axis are $O$, $\overrightarrow{OC_5'}$, and $\overrightarrow{OC_1}$, respectively, where $C_5'$ is the projected point of $C_5$ onto the floor surface of the prototype. Here, the view of a camera is represented with a coordinate in the spherical coordinate system as shown in Figure 2 (a). We consider the transformation of an image at $a = (r, \theta_a, \phi_a)$ to an image at $b = (r, \theta_b, \phi_b)$, where $r$ is the radius of the hemisphere, $\theta$ is the polar angle, and $\phi$ is the azimuthal angle. For this procedure, the homography $H_{ab}$ between the images at $a$ and $b$ is needed. For easy of computation, we calculate $H_{ab}$ by a product of $H_{ac}$ and $H_{cb}$, where the view $c$ is $(r, 0, 0)$. $H_{ac}$ can be obtained as [20]

$$H_{ac} = R_{ac} - \frac{t_{ac} n^T}{d}, \tag{1}$$

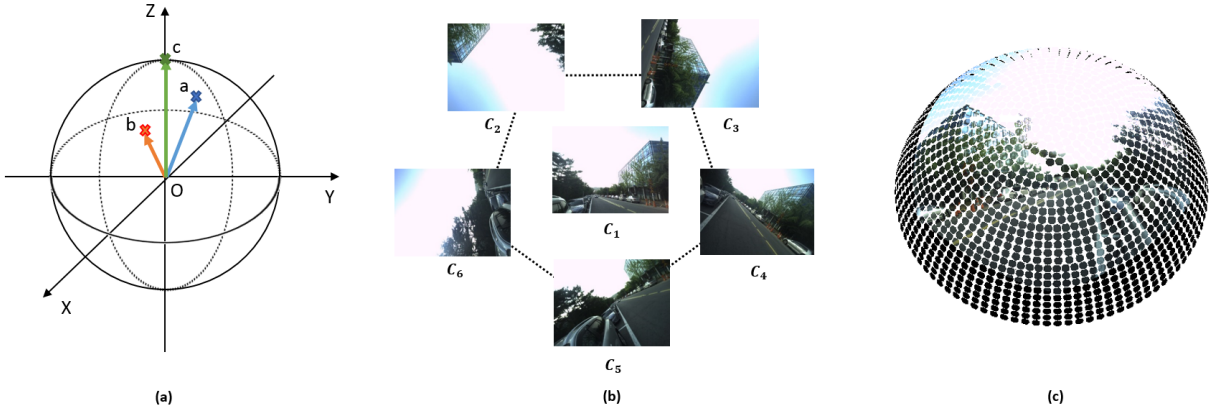


Fig. 2. A visualization of the compound eye's spherical coordinates, and an example of the proposed compound mapping scheme. (a) Three views of single eyes are visualized, (b) the images captured by the compound eye prototype, and (c) corresponding compound mapping results.

where $n$ is the normal vector of the image plane at view $c$, $d$ is the distance between the camera and the image plane at view $c$, $t_{ac}$ is the translation vector from $c$ to $a$, and $R_{ac}$ is the rotation matrix by which $a$ is rotated in relation to $c$. Here, $R_{ac}$ can be obtained as follows:

$$R_{ac} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_a) & -\sin(\theta_a) \\ 0 & \sin(\theta_a) & \cos(\theta_a) \end{bmatrix} \begin{bmatrix} \cos(\phi_a) & \sin(\phi_a) & 0 \\ -\sin(\phi_a) & \cos(\phi_a) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (2)$$

and $t_{ac}$ can be obtained as

$$t_{ac} = (r\sin(\theta_a)\cos(\phi_a), r\sin(\theta_a)\sin(\phi_a), r\cos(\theta_a)). \quad (3)$$

Similarly, $H_{bc}$ can be obtained as

$$H_{bc} = R_{bc} - \frac{t_{bc}n^T}{d}. \quad (4)$$

Finally, $H_{ab}$ can be obtained as

$$H_{ab} = H_{ac}H_{bc}^{-1}. \quad (5)$$

A pixel $p_a$ on the image at view $a$ is transformed to a pixel $p_b$ on the image at view $b$ with the following relation:

$$p_b = K_b H_{ab} K_a^{-1} p_a, \quad (6)$$

where $K_a$ and $K_b$ are intrinsic camera parameter matrices, and both $p_a$ and $p_b$ are in the homogeneous coordinates. For a single eye, we assume that the intrinsic parameters are the same as the ones of the camera module. Note here that only valid pixels within the size of the single image are selected. An example of the proposed compound mapping is shown in Figure 2 (c).

## V. Semantic Segmentation for Compound Images

In this section, the proposed semantic segmentation network for compound images is introduced. To leverage the conventional CNN scheme, it is convenient to transform compound images into a shape which is suitable for convolution. To achieve this, we put some constraints on compound image configuration. We consider a discrete level on the hemisphere surface and each level has uniformly increasing

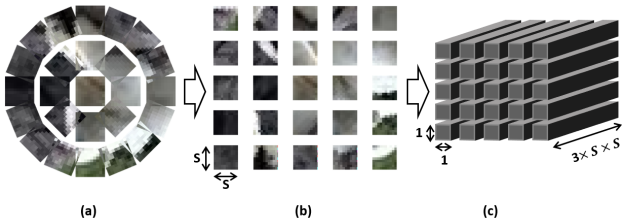


Fig. 4. (a) An example of the compound image configuration with the constraints. This example shows the case of $n = 3$, (b) The single images on the hemisphere surface are considered to be on a flat square surface, (c) An illustration showing that each single image is vectorized to make it suitable for CNN.

polar angle. The number of single eyes at the $n$th level is $(8n-8)$ except the first level. At the first level, there is only one single eye. With these constraints, we can transform the compound image into a tensor representation by vectorizing each image of single eye. Note that in this transformation, we can preserve the spatial neighboring relations of each single eye. An example of this discrete level compound image configuration and corresponding tensor representation is shown in Figure 4.

After that, we can take advantage of the conventional CNN scheme by using the tensor represented compound images. We design the network with two aspects: (i) the number of parameters is small, and (ii) the network structure is simple for easy implementation for a mobile robot. The basic structure of the proposed network follows the fully convolutional networks [6]. The proposed network structure is shown in Figure 3. The network consists of four convolutional layers, and a leaky ReLU activation layer is followed after each convolutional layer. In all convolutional layers, 3x3 filters are used. Note that in the last layer, no activation layer is applied, but a pixel-wise softmax layer is applied. The output of the network is pixel-wise confidence distribution which shows how much the single image is classified to different classes. The number of parameters of the proposed network (in the case of $S = 10$) is 0.33 millions, which is much smaller than the other networks (134 millions for FCN-VGG16, 6 millions

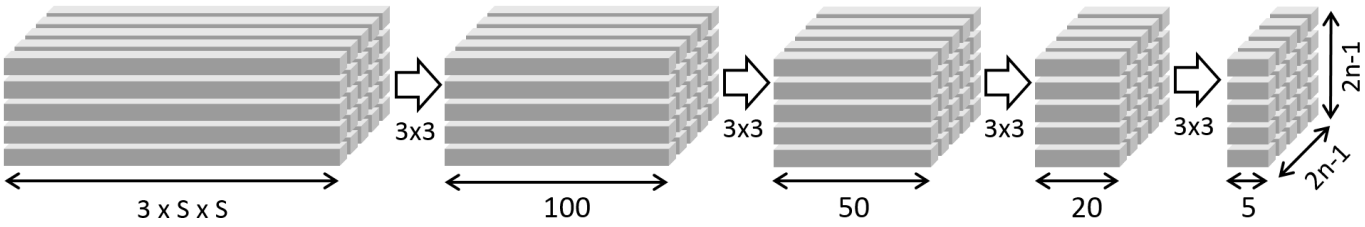


Fig. 3. A visualization of the proposed semantic segmentation network for compound images. It consists of four convolutional layers and a leaky ReLU activation layer is followed after each convolutional layer except the last layer. The output of the last layer is fed to the softmax layer to predict classes of each single image. In all convolutions, we use 3x3 size filters.

TABLE I
PERFORMANCE OF THE PROPOSED SCHEME WITH VARIOUS SIZES OF SINGLE IMAGES

| S (Pixel) | 1 | 3 | 5 | 10 | 20 | 30 |
|---|---|---|---|---|---|---|
| Mean IU | 0.411 | 0.421 | 0.427 | **0.432** | 0.425 | 0.416 |

for FCN-GoogLeNet [6]).

We choose the mean squared error loss between the network output and the ground truth. The network is trained with RMSprop [21] with a learning rate of 0.0001. We use the batch size of 256, and the network is trained for 200 epochs.

## VI. EXPERIMENTAL RESULTS

### A. Training Set

A data set of compound images is needed to train the proposed network. However, to the best of our knowledge, publicly available data set of compound images does not exist. Furthermore, making a new data set incorporating the proposed prototype needs pixel-wise class annotations, which is a very harsh task. Hence, we simulated compound images with a public semantic segmentation data set using the mapping scheme introduced in Section IV. For this procedure, we assumed that the image was captured from the first camera $C_1$, and input images and corresponding segmentation maps were transformed to $21 \times 21 \times (3 \times S \times S)$ size tensor, *i.e.*, $n = 11$. We used the COCO-Stuff 10K data set [22] for the data set generation. This data set is composed of $10,000$ complex images from COCO data set [23], and each image has dense pixel-level class annotations. There are total 182 classes of 91 thing classes and 91 stuff classes. However, inferencing all detailed classes might have little utility improvement compared to the increased complexity. Therefore, we selected four classes that is suitable for mobile robots, and they are *things*, *ground*, *sky*, *structure* and the other classes were considered as *background*. $9,000$ images were used as the training data set, and the other images were used as the test data set. For this procedure, we set $r = 110$mm and $d = 1500$mm. We note that a single eye image has more than one classes after the compound image mapping. We chose the most frequent class as the ground truth class of each single eye image.

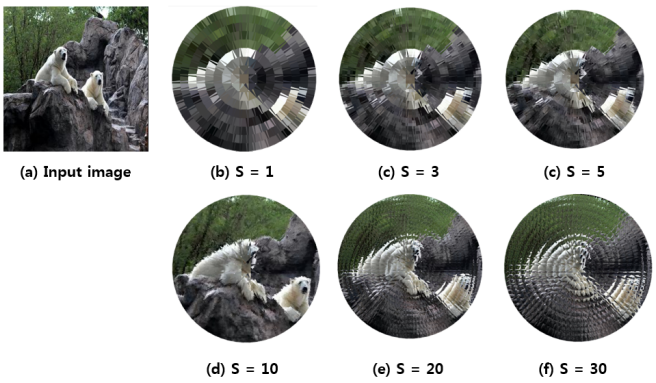


Fig. 5. Visualizations of compound mapped images with various sizes of single eye image.

### B. Evaluation Metric

The performance was evaluated with mean intersection over union (IU) metric following the practice in [6]. The mean IU metric is defined in each test sample, and the final measure is the average among all test samples. Let $k_{ij}$ be the number of pixels whose ground-truth class is $i$ and classified as class $j$. Then, the mean IU metric of a sample is defined as

$$\frac{1}{K}\sum_i \frac{k_{ii}}{\sum_j k_{ij} + \sum_j k_{ji} - k_{ii}}, \tag{7}$$

where $K$ is the total number of classes.

### C. Quantitative Evaluation

We evaluated the proposed network with various single eye image sizes. In each case, the network structure is the same except the channel depth of the first layer which is $3 \times S \times S$. The comparison result is shown in Table I, and some visualized results are shown in Figure 6. We can see that the best performance was achieved when the size of the single eye image is $10 \times 10$. To analyze the result, we visualized the compound mapped images with various sizes of the single eye image, which is shown in Figure 5. When the single eye size is too small ($S = 1$), the compound images cannot capture enough information compared to the cases of larger single eye sizes. On the other hand, when the single eye size is too large ($S = 30$), there are too many common pixels, which make it difficult to define clear semantic class. In case of a moderate single image size ($S = 10$), the

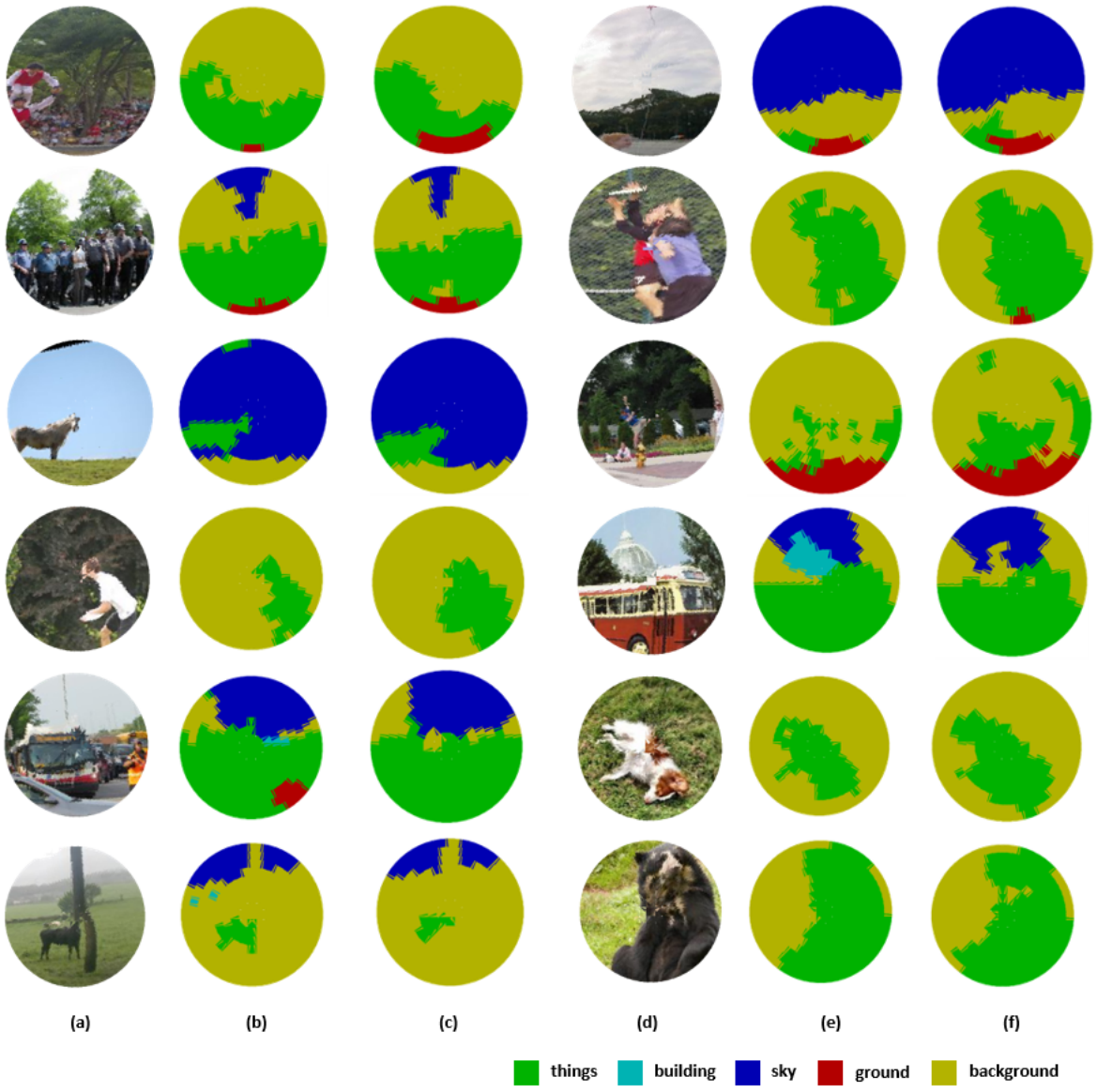


Fig. 6. Some examples of the semantic segmentation results. (a), (d) : compound mapped images, (b), (e) : ground truth semantic labels, (c), (f) : test results of the proposed network. Here, different colors mean different classes.

common pixel information can help to determine neighboring semantic classes. To verify this merit of compound images compared to typical RGB images, we applied the proposed network to RGB images. For the fair comparison, we roughly cropped the RGB images to have the same visible regions compared to the compound images. After that, the cropped images were resized to the size of $210 \times 210$, which make an RGB image to have the same number of pixels compared to a compound image at $S = 10$, and every non-overlapping $10 \times 10$ patches were vectorized to make it suitable for the proposed network. The mean IU for this case was 0.364 which is much worse than the compound image cases.

Interestingly, the performance for the case of RGB images was worse than the case of $S = 1$ which has no common pixel information as RGB images. However, with the pixel-wise accuracy measure, the case of RGB images showed slightly better performance (61.7%) than the case of $S = 1$ (61.0%). This difference was came the increased *false positive* in the case of RGB images. The compound mapped images had a tendency to have fewer *false positive*.

### D. Qualitative evaluation

We qualitatively evaluated the proposed network on the real images captured from the proposed prototype. We simulated images captured in outdoor to make the compound images of $n = 25$. Note that all images taken from $C_1, \cdots, C_6$ were used and the size of a single eye image was set to $S = 10$. For the test, we used the network trained on the COCO-Stuff 10K training set. Some results are visualized in Figure 7. From these figures, we can see that the proposed network can find out rough class of each single eye although it was tested on the images which are taken under different

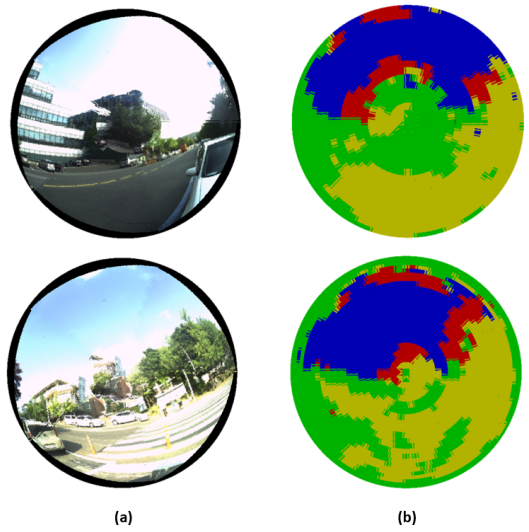


Fig. 7. Some examples of the semantic segmentation results on the real images. (a) compound mapped images, (b) inferred classes of each single image. Here, different colors mean different classes.

conditions compared to the COCO-Stuff 10K data set.

### E. Computation time

We measured the computation time for inferencing $1,000$ test images. It took 8.2 seconds on a 12 GB NVIDIA TitanX GPU. It is about 120 fps, and is suitable for real-time applications. We can expect that a low-cost customized hardware can be developed to process necessary processing of the proposed scheme to be applied to low-cost mobile robots.

## VII. Conclusions

In this work, we have considered the design of a compound camera prototype and a light-weight semantic segmentation scheme which is suitable for mobile robots. The experimental results have shown that compound images have merits for the semantic segmentation application compared to typical RGB images. We have also tested the proposed scheme with various sizes of a single eye image. The best performance have been achieved when the size of a single eye image is $10 \times 10$. In this configuration, the $n$th discrete level has the polar angle of $(3n-3)^\circ$, and each single eye has a uniformly increasing azimuthal angle whose interval is $360/(8n-8)^\circ$. In addition, each single eye image has approximately 53% of overlapping region. This experimental result can provide a guideline for the future design of a compound eye image sensor. Designing a flexible convolutional neural network structure that can process the general image structure, such as compound images, is an interesting research topic, which is left as future work.

## 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 (UD130070ID).

## References

[1] P.-E. Fayemi, N. Maranzana, A. Aoussat, and G. Bersano, "Bio-inspired design characterisation and its links with problem solving tools," in *DS 77: Proceedings of the DESIGN 2014 13th International Design Conference*, May 2014.

[2] E. Warrant and D.-E. Nilsson, *Invertebrate vision*. Cambridge University Press, 2006.

[3] R. Dudley, *The biomechanics of insect flight: form, function, evolution*. Princeton University Press, 2002.

[4] D. Floreano, J.-C. Zufferey, M. V. Srinivasan, and C. Ellington, *Flying insects and robots*. Springer, 2010.

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

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2015.

[7] 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.

[8] 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.

[9] 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.

[10] Y. Kitamura, R. Shogenji, K. Yamada, S. Miyatake, M. Miyamoto, T. Morimoto, Y. Masaki, N. Kondou, D. Miyazaki, J. Tanida, *et al.*, "Reconstruction of a high-resolution image on a compound-eye image-capturing system," *Applied Optics*, vol. 43, no. 8, pp. 1719–1727, 2004.

[11] J. Neumann, C. Fermuller, Y. Aloimonos, and V. Brajovic, "Compound eye sensor for 3d ego motion estimation," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 2004.

[12] X. He, R. Zemel, and D. Ray, "Learning and incorporating top-down cues in image segmentation," *Proc. European Conference on Computer Vision*, May 2006.

[13] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int'l J. Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009.

[14] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. IEEE Int'l Conf. Computer Vision*, September 2009.

[15] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int'l Conf. Computer Vision*, December 2015.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[17] J. Lafferty, A. McCallum, F. Pereira, *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. of International Conference on Machine Learning*, June 2001.

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

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016.

[20] O. Chum, T. Pajdla, and P. Sturm, "The geometric error for homographies," *Computer Vision and Image Understanding*, vol. 97, no. 1, pp. 86–102, 2005.

[21] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, 2012.

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

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. European Conference on Computer Vision*, September 2014.