ICIC Express Letters
Part B: Applications
Volume **11**, Number **1**, January **2020**

ICIC International ©2020 ISSN 2185-2766
pp. **25–32**

DEEP HIERARCHICAL SEMANTIC SEGMENTATION ALGORITHM BASED ON IMAGE INFORMATION ENTROPY

Qing Li, Hongjian Wang*, Juan Li, Yao Xiao and Wenyue Hu

College of Automation
Harbin Engineering University
No. 145, Nantong Street, Nangang District, Harbin 150001, P. R. China
373512591@qq.com; *Corresponding author: cctime99@163.com
{ lijuan041; xiaoyao9; huwenyue }@hrbeu.edu.cn

Received July 2019; accepted October 2019

Abstract. *Computer vision and machine learning researchers are increasingly interested in image semantic segmentation. More and more application scenarios need precise and efficient segmentation techniques, such as autopilot, indoor navigation, and even virtual reality and augmented reality. Currently, semantic segmentation algorithm has low segmentation accuracy for small-scale targets. In this paper, DHSS (Deep Hierarchical Semantic Segmentation) algorithm based on image information entropy is proposed, which can effectively excavate DHSF (Deep Hierarchical Semantic Features) by abstracting image layers. The algorithm is based on the feature maps generated by VGG16, selecting the output feature map of pooling layers to construct the image hierarchy, and using image information entropy to describe low-level semantic feature maps which constitutes DHSF with strong expressive ability. Then connecting DHSF with ASPP (Atrous Spatial Pyramid Pooling) generates DHSS algorithm. Compared with the existing algorithms, the proposed DHSS algorithm in this paper can improve the semantic segmentation accuracy of small-scale objects, which effectively prevents the loss of small targets in the process of semantic segmentation.*

Keywords: VGG16, Deep hierarchical semantic features, Feature map, Deep hierarchical semantic segmentation

1. **Introduction.** Nowadays semantic segmentation is one of the key problems in computer vision. In the macro sense, semantic segmentation is a high-level task that paves the way for scene understanding. As a core element of computer vision, the scene understanding is becoming more and more important because more and more application scenarios need to infer relevant knowledge or semantics (i.e., from concrete to abstract) from images. These applications include unmanned vehicle [1, 2, 3], human-computer interaction [4, 5], computational photography [6], image search engine [7], augmented reality [8, 9], etc. These problems have been solved by applying various traditional computer vision and machine learning techniques. Although these methods are extremely popular, the deep learning revolution has brought about tremendous changes in related realms. Consequently, many computer vision problems, including semantic segmentation, begin to use depth architecture, usually Convolutional Neural Networks (CNNs) [10, 11, 12], which far exceed traditional methods on accuracy and even efficiency.

In recent years, deep learning has shown unique advantages in the fields of semantic segmentation [13], semantic recognition [14] and classification [15]. In addition, human perception system is a clear hierarchical structure [16]. The processing of visual information by human brain is a process of transferring and abstracting from layer to layer. Therefore, the image features constructed by simulating the hierarchical structure of the human brain have stronger expressive ability, and can describe the essential information

DOI: 10.24507/icicelb.11.01.25

of the image more efficiently from the source. Because the object's semantic information remains unchanged no matter what form it presents in the image, this paper proposed the deep hierarchical semantic segmentation algorithm based on image information entropy. First, VGG16 [17] is used to construct image hierarchical structure, and then construct the image features with strong expressive ability. Further, the image Deep Hierarchical Semantic Features (DHSF) based on image information entropy is extracted, combining ASPP [18] to construct image deep hierarchical semantic segmentation algorithm network. The experimental results show that the semantic segmentation network based on DHSF performs perfect in the accuracy of semantic segmentation and avoids the loss of small-scale targets.

Besides the development of research and the conclusion, this paper can be divided into three parts as the following: in Section 2, the image information entropy is introduced briefly and the structure of DHSF for hierarchical features is described; Section 3 expounds the framework of DHSS algorithm in detail; the experiments and analysis are given in Section 4.

2. Deep Hierarchical Semantic Features Based on Image Information.

2.1. Image information entropy.

Information entropy is a concept used to measure information quantity in information theory. In 1948, Shannon [19], father of information theory, pointed out in the paper *A Mathematical Theory of Communication* that any information has redundancy, and its quantity is related to the probability or uncertainty of the occurrence of each symbol (number, letter or word) in the information. Shannon drew on the concept of thermodynamics and called the average amount of information excluded from redundancy "information entropy".

The basic unit of digital image is the pixel. Each image data stored in a computer is essentially a matrix of pixels. Because different gray-scale pixels fill different spatial regions with different probability distributions, different images show different shape features.

Let an image with k gray levels take the value of k as 255, where i ($i \in 1, \ldots, k$) level gray level is p_i, then the entropy (information content) is:

$$H(p_i) = p_i \log(1/p_i) \tag{1}$$

where $0 \leq i \leq k$. If the accumulation of the entropy of different gray levels is defined as the information entropy of the image, then for the gray level image, the information entropy of the whole image is:

$$H = \sum_{i=0}^{k} H(p_i) = -\sum_{i=0}^{k} p_i \log(p_i) \tag{2}$$

where $k = 255$, p_i represents the probability of the occurrence of gray-scale pixels in the whole image. When $p_i = 0$, $p_i \log(p_i) = 0$. p_i is calculated by gray histogram, that is, the number of pixels equal to gray level i divided by the total number of pixels.

It can be seen from the above that the information entropy is calculated according to the gray-scale color of the image, so it is also called color information entropy.

2.2. Deep hierarchical semantic feature extraction algorithm.

Each level of VGG16 hierarchy contains several feature maps, each of which is expressed in different scales and levels. Figure 2 is a visualization result of some feature maps contained in the pooling layer.

Figure 2 shows that pool1, pool2 and pool3 layers have abundant and clear characteristic information, which can be understood directly through visualization. They can capture the edge of the object and the rapidly changing pixel information in the image. The detailed texture structure of the original image is still retained in the feature map. The pool4 layer also has some clear features, but the visualization results of the

Image Thermal map

FIGURE 1. Image information entropy thermal map

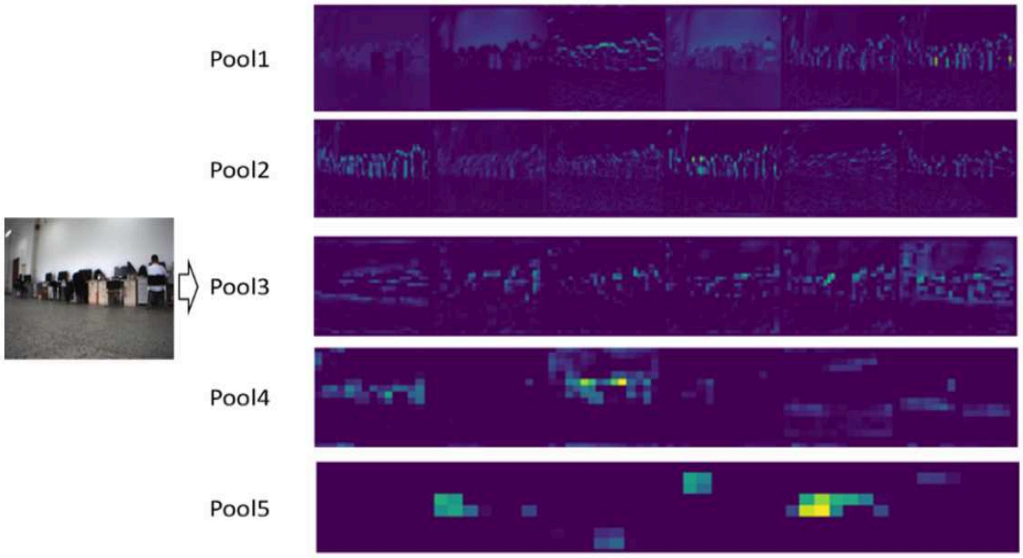

FIGURE 2. Partial feature maps of each pooling layer of image

pool5 layer are not easy to understand. The features express the semantic information of the image through continuous iteration and abstraction, which is difficult to understand through visualization. Therefore, the algorithm chooses the features of pool1, pool2, pool3 and pool4 to process the image information entropy to form the method of image deep hierarchical semantic feature extraction, as shown in Figure 3.

In VGG16 hierarchical structure, each layer of feature map expresses image information from different aspects. The algorithm makes full use of the information of feature maps to describe images, which can make the constructed image features more expressive.

In the paper, bilinear interpolation is used to sample the features at different levels to make them have the same size as the original image, which builds a 3D matrix $F \in R^{N \times H \times W}$ by stacking all feature maps, where N is the number of characteristic maps, H is image height and W is image width. F is expressed as:

$$F = [up(F_1), up(F_2), \ldots, up(F_L)] \tag{3}$$

where, up is sampling operation $up(F_1) \in R^{N_l \times H \times W}$, N_l is the number of characteristic maps of l-level. For any region Q on the image, its descriptor can be expressed as $D_Q \in R^N$ by N dimensional vectors, and the i-dimensional d_i of the descriptor represents the result of describing the corresponding region pixels on the i-level feature map.

Figure 3 is a schematic diagram of image DHSF extraction. 64 feature maps from pool1 layer are sampled twice, 128 feature maps from pool2 layer are sampled 4 times, 256 feature maps from pool3 layer are sampled 8 times, and 512 feature maps from pool4 layer are sampled 16 times. Then, the information entropy of the feature maps is obtained from the up-sampling images of each pooling layer.

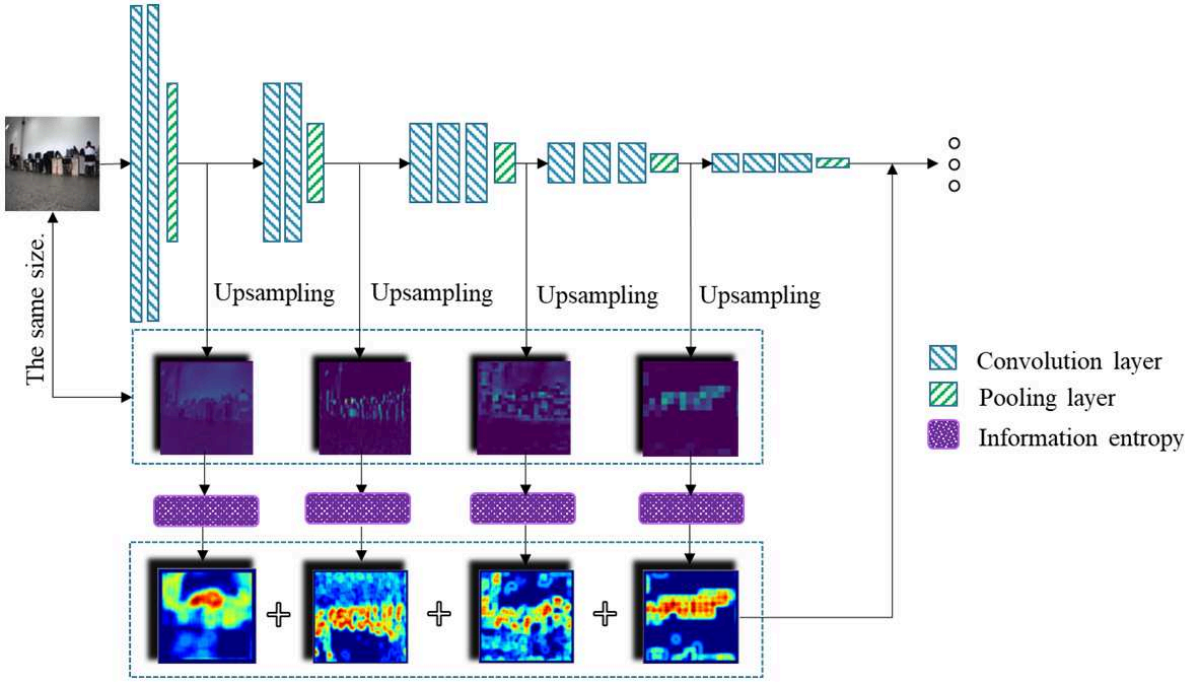

FIGURE 3. Deep hierarchical semantic feature extraction

3. **Deep Hierarchical Semantic Segmentation Algorithm.** The traditional semantic segmentation method is to resize the image to the same size, but this will cause some features to distort or disappear. SPP (Spatial Pyramid Pooling) [20] associates image pyramids used in SIFT feature extraction, scales images to different scales, and then extracts SIFT feature points with rotation and scaling invariance. Therefore, SPP is also used to achieve image size and different aspect ratio processing. ASPP drew on the structure of SPP. In fact, it is through different atrous convolution to zoom the image to different degrees and get different size of input feature maps. Because ASPP has filters with different rates (6, 12, 18, 24), and then pooling the features of the sub-windows generates a fixed-length representation, which is shown as Figure 4.

The ASPP layer is connected behind the last convolution layer, pooled the features, generating a fixed size output, which is then sent to the first full connection layer. It avoids requiring the same size at the network's input ports. This paper combines deep hierarchical semantic features with ASPP for image deep hierarchical semantic segmentation (DHSS) algorithm. The structure of DHSS algorithm is shown in Figure 5.

This paper uses VGG16 which is trained on ImageNet, adopting the deep hierarchical semantic feature extraction structure proposed in the previous section to obtain the information entropy thermo gram of the pooling layers (pool1, pool2, pool3, and pool4), combining with the results of ASPP to obtain semantic segmentation map. This method

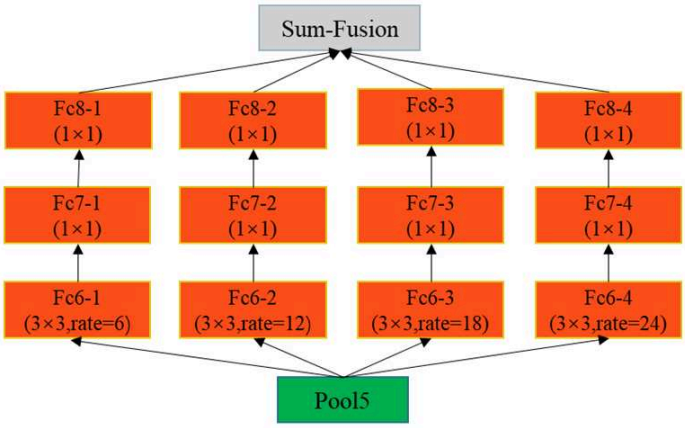

FIGURE 4. ASPP structure

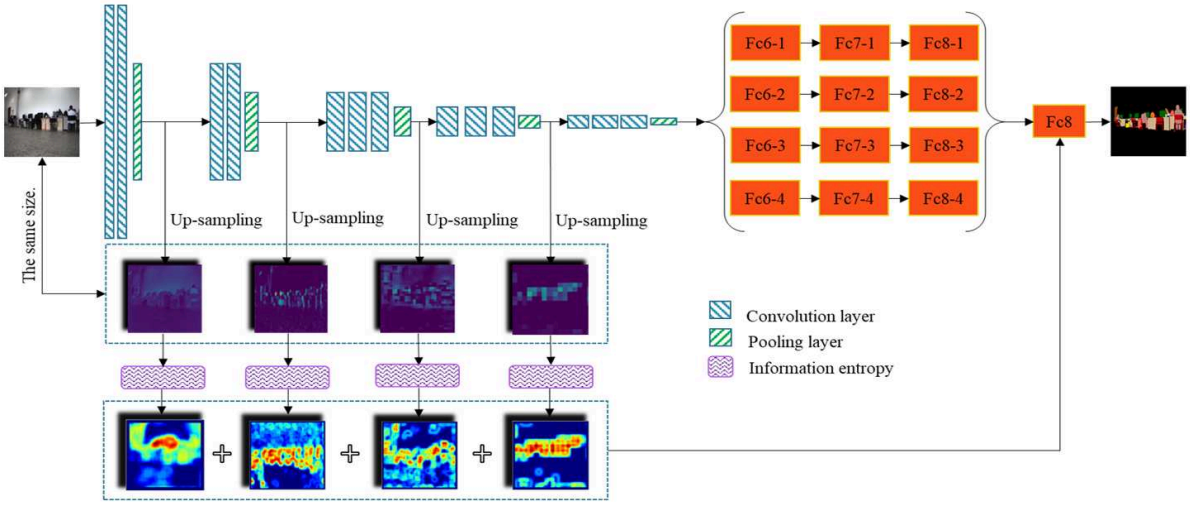

FIGURE 5. Network structure of DHSS algorithm

eliminates the CRF iteration process of Deeplab and achieves the end-to-end semantic segmentation network structure.

4. **Results and Analysis.** The semantic segmentation results of Figure 6 show that Deeplab-v2 tests Exp-VOC2012 dataset cannot predict the results of extended tags. Exp-Deeplab-v2 tests Exp-VOC2012 dataset can predict the results for new tags, but the segmentation accuracy is low. As shown in Figure 6, the DHSS algorithm proposed in this paper improves the segmentation accuracy of small-scale targets effectively and avoids the loss of small-scale targets. As shown in Table 1, the DHSS algorithm improves the segmentation accuracy of Exp-VOC2012.

The number of the 27 classes expanded in this paper is less in training set and test set, so the segmentation accuracy of Deeplab-v2 and Exp-Deeplab-v2 is awfully bad. However, the DHSS algorithm proposed in this paper can effectively enhance the segmentation accuracy by deep feature extraction. Pascal VOC2012 dataset is used for semantic segmentation with 20 kinds of semantic tags: aircraft, bicycle, boat, bus, car, motorcycle, train, bottle, chair, dining table, potted plant, sofa, TV/monitor, bird, cat, cattle, dog, horse, sheep and human. The dataset is divided into two subsets, using 1464 and 1449 images for training and testing respectively [13].

In this paper, Pascal VOC2012 is extended to 47 tags, which is called Exp-VOC2012 later. The Exp-VOC2012 adds 27 tags (box, bed, bag, ball, book, building, billboard, bicman, cup, column, curtain, cabinet, desk, fan, fence, hat, heating, hotpot, light, object,

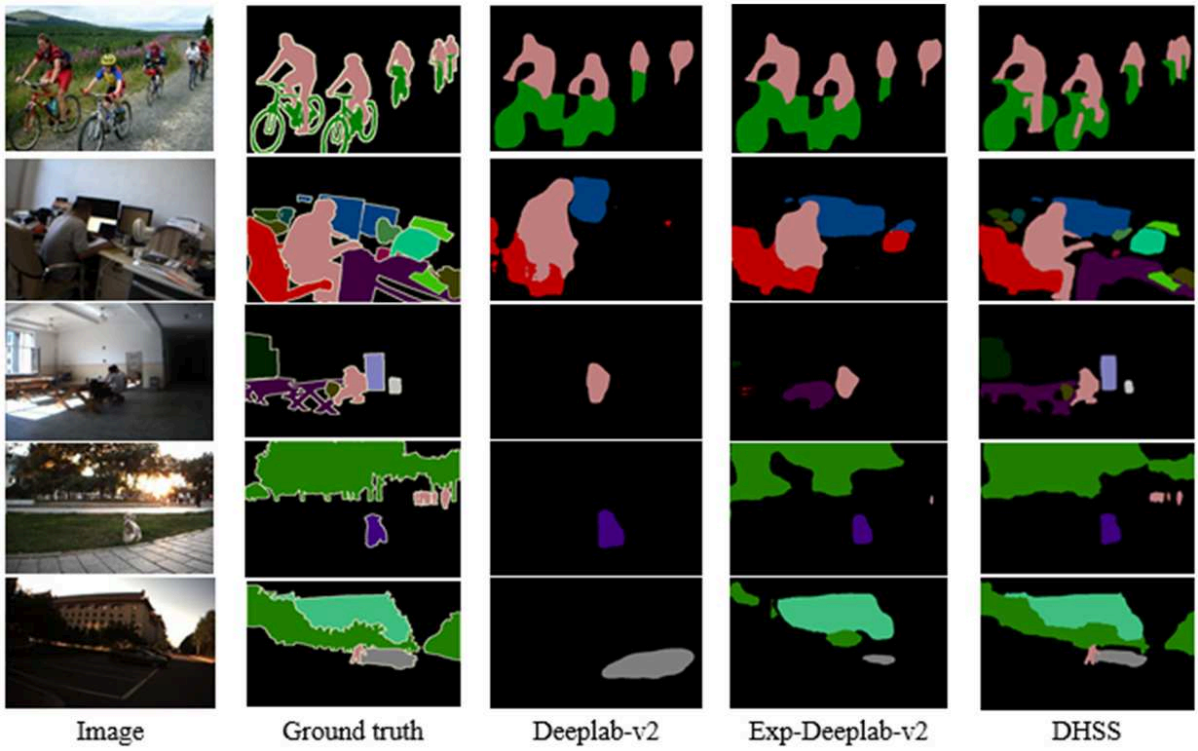

FIGURE 6. Visualization semantic segmentation results

TABLE 1. Results comparison experiments on Pascal VOC2012/Exp-VOC2012 test

Model	MIoU
Deeplab-v2/Pascal VOC2012	70.9
Deeplab-v2/Exp-VOC2012	31.3
Exp-Deeplab-v2/Exp-VOC2012	51.9
DHSS/Exp-VOC2012	79.8

pram, quilt, sign, stairs, tree, trash-can and umbrella) based on Pascal VOC2012. We use Deeplab-v2 to test Exp-VOC2012 (Deeplab-v2/Exp-VOC2012), Exp-Deeplab-v2 to test Exp-VOC2012 (Exp-Deeplab-v2/Exp-VOC2012), and DHSS algorithm to test Exp-VOC2012 (DHSS/Exp-VOC2012). The test results are shown in Figure 6. The Exp-Deeplab-v2 is a model that Deeplab-v2 trained in Exp-VOC2012 dataset.

Mean Intersection over Union (MIoU) [13]: this is the standard metric for segmentation purposes. It computes a ratio between the intersection and the union of two sets, in our case the ground truth and our predicted segmentation. That ratio can be reformulated as the number of true positives (intersection) p_{ii}, over the sum of true positives, false negatives and false positives (union) p_{ij} and p_{ji}. That IoU is computed on a per-class basis and then averaged, which is shown as:

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{4}$$

5. **Conclusions.** In this paper, image deep hierarchical semantic segmentation algorithm is proposed. Firstly, the algorithm is based on the image features generated by VGG16, selecting the output feature map of pooling layers to construct the image hierarchy, choosing the best hierarchical combination based on the task of semantic segmentation, using image information entropy to describe low-level semantic feature maps which constitutes

deep hierarchical semantic features with strong expressive ability, and describing the essential information of the image better. Ultimately, connecting the algorithm with ASPP generates image deep hierarchical semantic segmentation algorithm. The experimental results show that, compared with the existing algorithms, the DHSS algorithm can improve the accuracy of small-scale target semantic segmentation.

On the basis of the existing research, the next stage will replace the basic VGG16 model with more advanced model for specific semantic segmentation tasks, train feature extraction model, and construct image hierarchy structure, so that the extracted DHSF has more obvious advantages and can effectively and accurately complete the task of semantic segmentation.

Acknowledgment. This research work is supported by Natural Science Foundation of China (No. 61633008, No. 51609046). The authors also gratefully acknowledge the helpful support of the open source of Deeplab and Pascal VOC2012 dataset.

REFERENCES

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele, The cityscapes dataset for semantic urban scene understanding, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3213-3223, 2016.

[2] A. Sallab, M. Abdou, E. Perot et al., Deep reinforcement learning framework for autonomous driving, *Electronic Imaging*, vol.2017, no.19, pp.70-76, 2017.

[3] L. Deng, M. Yang, Y. Qian et al., CNN based semantic segmentation for urban traffic scenes using fisheye camera, *IEEE Intelligent Vehicles Symposium (IV)*, 2017.

[4] M. Oberweger, P. Wohlhart and V. Lepetit, Hands deep in deep learning for hand pose estimation, *arXiv Preprint*, arXiv:1502.06807, 2015.

[5] D. Kollias, G. Marandianos, A. Raouzaiou et al., Interweaving deep learning and semantic techniques for emotion analysis in human-machine interaction, *IEEE the 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, 2015.

[6] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee and I. S. Kweon, Learning a deep convolutional network for light-field image super resolution, *Proc. of the IEEE International Conference on Computer Vision Workshops*, pp.24-32, 2015.

[7] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang and J. Li, Deep learning for content-based image retrieval: A comprehensive study, *Proc. of the 22nd ACM International Conference on Multimedia*, pp.157-166, 2014.

[8] L. Abdi and A. Meddeb, Driver information system: A combination of augmented reality and deep learning, *The Symposium*, 2017.

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell, DeCAF: A deep convolutional activation feature for generic visual recognition, *ICML*, pp.1-2, 2014.

[10] C. Farabet, C. Couprie, L. Najman and Y. LeCun, Learning hierarchical features for scene labeling, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.35, no.8, pp.1915-1929, 2013.

[11] H. C. Shin, H. R. Roth, M. Gao et al., Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, *IEEE Trans. Medical Imaging*, p.1, 2016.

[12] F. Liang, C. Shen and F. Wu, An iterative BP-CNN architecture for channel decoding, *IEEE Journal of Selected Topics in Signal Processing*, p.1, 2018.

[13] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea et al., A review on deep learning techniques applied to semantic segmentation, *arXiv Preprint*, arXiv:1704.06857, 2017.

[14] Y. Hu, Z. Chen and W. Lin, RGB-D semantic segmentation: A review, *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2018.

[15] J. Sun and J. Ponce, Learning discriminative part detectors for image classification and co-segmentation, *Proc. of the IEEE International Conference on Computer, Vision*, pp.3400-3407, 2013.

[16] Q. Li, K. Li, X. G. You et al., Place recognition based on deep feature and adaptive weighting of similarity matrix, *Neurocomputing*, vol.199, pp.114-127, 2016.

[17] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *Computer Science*, 2014.

[18] L. C. Chen, G. Papandreou, I. Kokkinos et al., DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol.40, no.4, pp.834-848, 2016.

[19] C. E. Shannon, A mathematical theory of communication, *Bell Labs Technical Journal*, vol.27, no.4, pp.379-423, 1948.

[20] K. He, X. Zhang, S. Ren et al., Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol.37, no.9, pp.1904-1916, 2014.