https://doi.org/10.2352/ISSN.2470-1173.2019.7.IRIACV-457

Exploring Variants of Fully Convolutional Networks with Local and Global Contexts in Semantic Segmentation Problem

Dong-Won Shin, Jun-Yong Park, Chan-Young Sohn, and Yo-Sung Ho; Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005, South Korea

Abstract

Recently, the semantic inference from images is widely used for various applications, such as augmented reality, autonomous robots, and indoor navigation. As a pioneering work for semantic segmentation, the fully convolutional networks (FCN) was introduced and outperformed traditional methods. However, since FCN only takes account of the local contextual dependency, it does not reflect the global contextual dependency. In this paper, we explore variants of FCN with local and global contextual dependencies in the semantic segmentation problem. In addition, we tried to improve the performance of semantic segmentation with extra depth information from a commercial RGBD camera. Our experiment result indicates that exploiting the global contextual dependencies and the additional depth information improves the quality of semantic segmentation

Introduction

Owing to the visual recognition system in the human brain, people can easily segment the entire image into meaningful sub-regions and exploit the semantic information to complete specific tasks, such as navigation, exploration, and grasping. Interestingly enough, this semantic segmentation problem is very simple to human beings, but it is very difficult to robotic agents. There are various attempts to solve the semantic segmentation problem in the computer and robotic vision fields. By virtue of the recent advancement in deep learning algorithms, semantic segmentation results have been improved significantly.

As a pioneer work for semantic segmentation using the deep learning approach, the fully convolutional networks (FCN) was introduced and outperformed traditional methods [1]. On top of the VGG-16 networks, the fully connected layer at the end was replaced by the convolutional operation to maintain a spatial context and the skip connections were constructed to preserve the feature maps from the intermediate layers. However, this fully convolutional architecture misses the global contextual dependencies through the entire image.

With the advent of the commercial RGB-D cameras, such as Microsoft Kinect, Asus Xtion, Intel Realsense, and Orbbec Astra, the complementary depth information can be easily obtained; thus, employing the depth value in semantic segmentation has now become very beneficial.

In this paper, in order to overcome the drawback of FCN over the global contextual dependencies, we have explored its variants with a variety of input feature encoding. In addition, we have tested the variants extensively with the publically available RGB-D datasets (NYUDv2 and SUNRGBD) based on popular evaluation measures (mean accuracy, overall accuracy, and mean intersection-over-union).

Related Works

A fully convolutional network (FCN) architecture has been introduced by [3]. This combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentation by applying end-to-end training. However, this approach only considers the local contextual dependencies due to the convolutional operation, which limits the performance improvement.

The global contextual dependencies indicate the correlation among spatially distant semantic labels in the image. For example, if the sky appears on the top of the image, it is less probable that the microwave oven shows in the bottom of the image. In order to reflect the global contextual dependencies, the ReSeg algorithm employs recurrent neural networks in the 2D spatial domain along the vertical and horizontal axes [2]. This achieves a better result than the conventional method.

As the depth information is easily available from commercial RGB-D cameras, the LSTM-CF approach made use of the depth information to estimate the semantic meaning correctly from the image [3]. Moreover, the memorized context layer and memorized fusion layer learn the global context to make a better quality.

On top of the encoder-decoder type network, the FuseNet algorithm incorporated the depth information in the sparse and dense fusion block. The sparse fusion inserts the fusion layer before each pooling, and the dense fusion adds after each activation. They mathematically proved that the proposed fusion technique can produce a stronger signal for training [4].

The masked convolution is a kind of the auto-regressive model that estimates future values by using a joint probability distribution of previous sequences [5]. It is originally used for the image generation. However, we applied it to the semantic segmentation problem since the masked convolution can learn the joint probability distribution over the input feature encodings.

U-Net is convolutional neural networks for biomedical image segmentation [8]. The networks consist of a contracting path and an expansive path. A property of this networks is that it performs copy and crop between the two paths. Concatenate the result values before the max pooling in the expansive path in each layer. Also, there is no fully connected layer. Each input image is cut in each patch, and an overlap-tile strategy is used. Data augmentation makes a great contribution to improving the performance of this network.

In several later papers, CNN is more accurate with input and each layer closer together. Densely connected convolutional network is, some layers are grouped together into one block and dense connections are formed between layers within one block [9].

Proposed Method

A. FCN baseline

FCN consists of a series of convolutional blocks containing the convolution layer, the ReLU activation function, and the max-pooling layer, followed by the upsampling operation at the end. In order to respect the global structure, FCN combines the fine and coarse layers with an elementwise fusion. In our experiment, we take the FCN-8s model as the baseline since it shows better performance than the others (FCN-16s and FCN 32s). Fig. 1 illustrates the network structure of FCN.

Even though it respects the global structure by the layer fusion, this is not enough to get the consistent global contextual dependencies from the input. Thus, we present several FCN variants considering the global contextual dependencies and exploiting additional depth information.

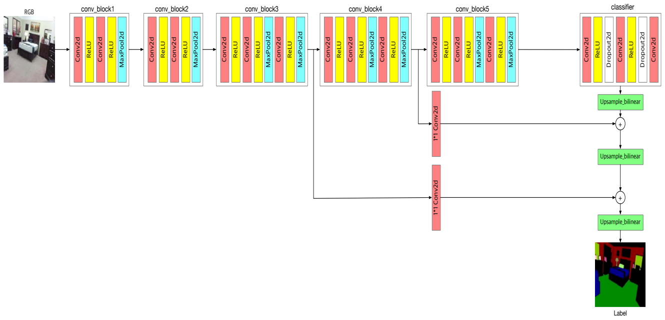


FIG. 1 NETWORK STRUCTURE OF FCN

B. FCN with ReNet layer

One of the methods capturing the global contextual dependencies is ReNet layer applying the recurrent neural network (RNN) to a 2D spatial domain [6]. After dividing the input feature into a grid structure, each information in a grid cell is fed into RNNs node in a vertical and horizontal direction and the two feature maps from each direction are concatenated. By the recurrent structure through the vertical and horizontal direction, the ReNet layer can learn about the global contextual dependencies. We finally combined it with FCN as shown Fig. 2.

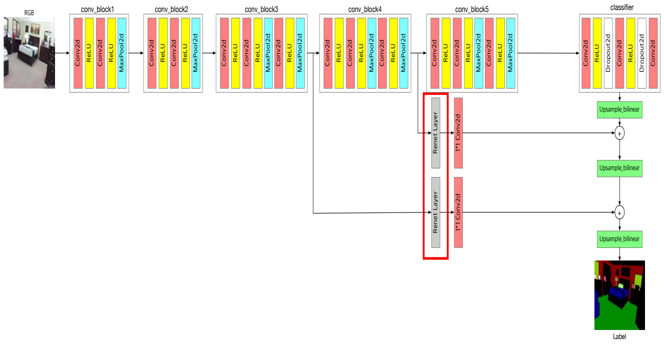


FIG. 2 THE NETWORK STRUCTURE OF FCN WITH RENET LAYER

C. FCN with masked convolution

Another method to capture the global contextual dependencies is masked convolution from Google Deepmind[5]. Masked convolution is one of the autoregressive models that predicts the output variable depends linearly on its own previous values. By using the convolution operation masked on the lower-right from the center, it can learn the global contextual dependencies of the dataset. Fig. 3 shows the network structure of FCN with masked convolution.

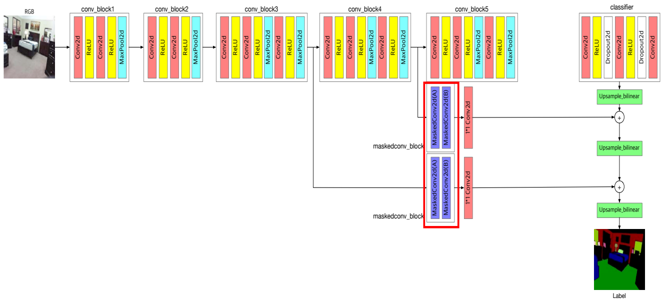


FIG. 3 THE NETWORK STRUCTURE OF FCN WITH MASKED CONVOLUTION

D. FCN with RGB-D input

An additional depth information can improve the performance of semantic segmentation. In order to consider the depth information, we construct Siamese networks for one channel depth input. The additional depth branch is same as the original FCN baseline but the channel of input layer should be changed to one channel since the channel of the input depth is one. The output feature maps from the color and depth branch are summed up by the elementwise operation. Fig. 4 represents the network structure of FCN with RGB-D input.

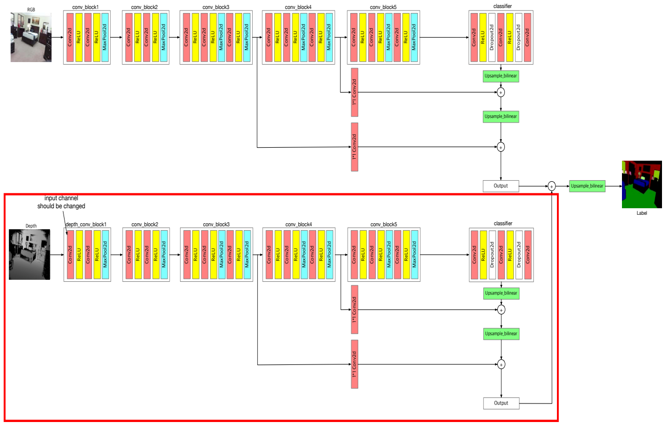


FIG. 4. THE NETWORK STRUCTURE OF FCN WITH RGB-D INPUT

E. FCN with RGB-HHA input

The HHA representation is a geocentric feature encoding converted from the depth map [7]. It consists of three channels and each channel represents the height above the ground, the horizontal disparity and the angle with the gravity vector, respectively. This expressive feature encoding helps the network to learn the distinguishable representation from the ordinary depth information for the semantic segmentation. Fig. 5 shows the network structure of FCN with RGB-HHA input. As similar as the network structure of FCN with RGB-D input, it is composed of the Siamese networks and the late fusion at the end of the network.

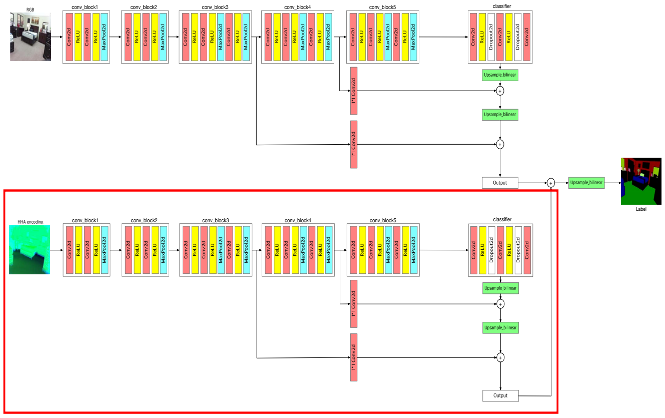
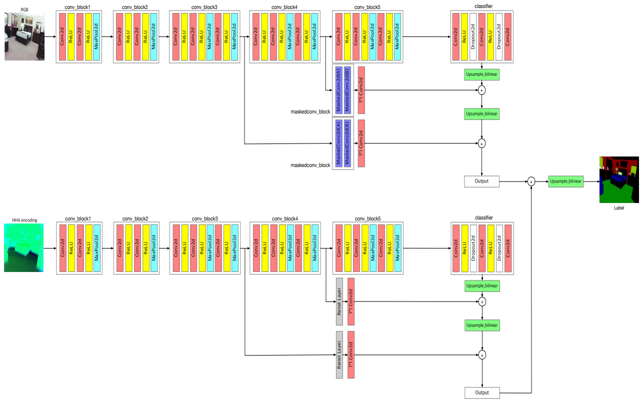


FIG. 5 THE NETWORK STRUCTURE OF FCN WITH RGB-HHA INPUT

F. Hybrid method

The hybrid method is a combination of the masked convolution and ReNet layer. For the color branch, we exploited the masked convolution block to capture the global contextual dependencies. For the depth branch, we employed the ReNet layer to capture the global contextual dependencies.



Experiment Results

A. Evaluation dataset

We explored the FCN variants with the publically available RGB-D datasets: NYUDv2 and SUNRGBD [7]. NYUDv2 dataset contains 795 images for the training and 654 images for the test from a variety of indoor scenes with 13 class labels. This dataset is relatively smaller then SUNRGBD dataset, therefore it is suitable for checking the tendency of the network performance.

Next, SUNRGBD dataset contains 5285 images for the training and 5050 images for the test. This dataset was captured by four different RGB-D sensors and includes more images from diverse indoor environments with 37 class labels, therefore it is suitable for the extensive experiments for the semantic segmentation.

Both datasets provide the RGB, depth images and the corresponding semantic label sets produced by the manual endeavor from annotation tools. Especially for SUNRGBD dataset, since it offers a toolbox for computing the HHA representation, we aggressively exploited the toolbox.

B. Evaluation metric

We evaluated the performance of the semantic segmentation by three different measures: overall accuracy, mean accuracy and mean intersection over union.

The overall accuracy is the percentage of the correctly classified pixels, defined by

$$overall\ accuracy = \frac{1}{N}\sum_{c} TP_c\,, c \in \{1, \dots, K\} \qquad (1)$$

Next, the mean accuracy is the average of classwise accuracy, defined by

$$mean\ accuracy = \frac{1}{K}\sum_{c} \frac{TP_c}{TP_c + FP_c} \qquad (2)$$

Lastly, the mean intersection-over-union (IoU) is an average value of the intersection of the prediction and ground truth regions over the union of them, defined by

$$mean\ IoU = \frac{1}{K}\sum_{c} \frac{TP_c}{TP_c + FP_c + FN_c} \qquad (3)$$

C. Hardware specification and hyperparameters

In our experiments, we employed NVIDIA Geforce GTX 1080 Ti with 11GB RAM. We used the stochastic gradient descent. The learning rate, momentum, and weight decay are 10^{-5}, 0.99 and 5^{-4} respectively. The number of epoch was 50 and the total training time was approximately one day for RGB variants and two days for RGB-D, RGB-HHA variants.

D. Qualitative results

Through Table 1 to 8, those show the exploring results from FCN variants with the different evaluation measures and the datasets.

From the experiment, we can reach several conclusions. First, FCN baseline shows the better results when it exploits the depth or HHA encoding for the network. FCN with ReNet layer produces the growing trend like FCN baseline, this, however, indicates a slightly worse result than each of them.

When we only compare the RGB input case, FCN with masked convolution shows the best result among the others but represents the worse result when it considers the depth branch. This shows the masked convolution is beneficial on the color branch but not the depth branch.

Conclusion

In this paper, we explored the variants of the fully convolutional networks with the local and global contexts and diverse input encodings. For the global contextual dependencies, we experimented the ReNet layer and the masked convolution. For the input feature encodings, we experimented RGB, RGB-D, and RGB-HHA. The test results indicate the masked convolution is good for the color branch but not the depth branch. Also, ReNet layer is compatible with the color and depth branch but not outperform the baseline network. We hope this inspiration is going to be beneficial future network models for semantic segmentation problem.

TABLE 1 VISUALIZATION OF THE RESULT FROM NYUDV2

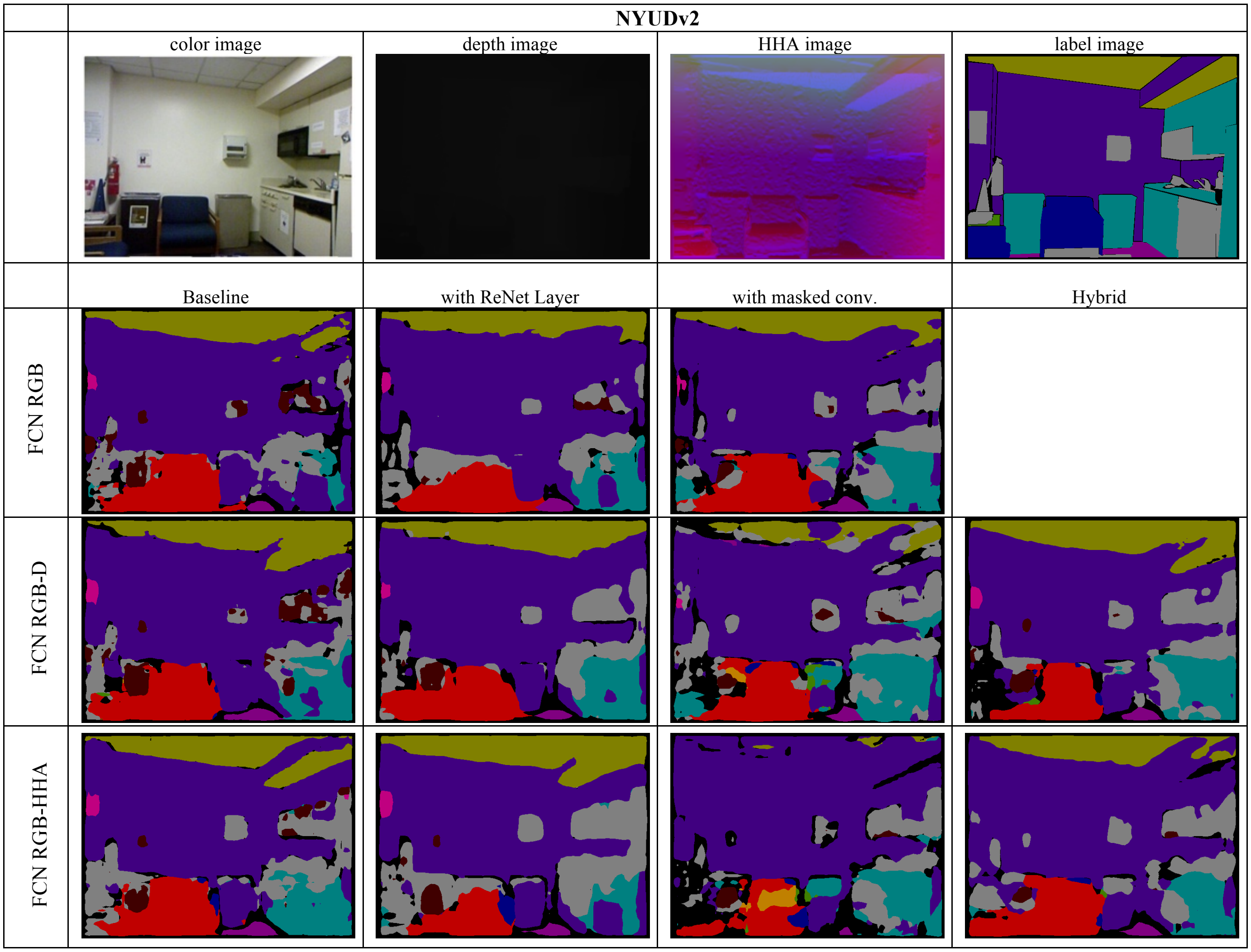


TABLE 2 NYUDV2 MEAN IOU				
	Baseline	with ReNet layer	with masked convolution	Hybrid method
FCN RGB	0.33013	0.31877	0.36638	
FCN RGB-D	0.34947	0.33803	0.34644	0.37582
FCN RGB-HHA	0.38284	0.37668	0.32541	0.40218

TABLE 3 NYUDV2 MEAN ACCURACY				
	Baseline	with ReNet layer	with masked convolution	Hybrid method
FCN RGB	0.46821	0.45994	0.51028	
FCN RGB-D	0.48753	0.47480	0.48485	0.51520
FCN RGB-HHA	0.52087	0.51159	0.45208	0.54824

TABLE 4 NYUDV2 OVERALL ACCURACY				
	Baseline	with ReNet layer	with masked convolution	Hybrid method
FCN RGB	0.57633	0.56781	0.60167	
FCN RGB-D	0.59422	0.58880	0.56407	0.61135
FCN RGB-HHA	0.61602	0.61288	0.56141	0.58121

Table 5 Visualization of the result from SUNRGBD

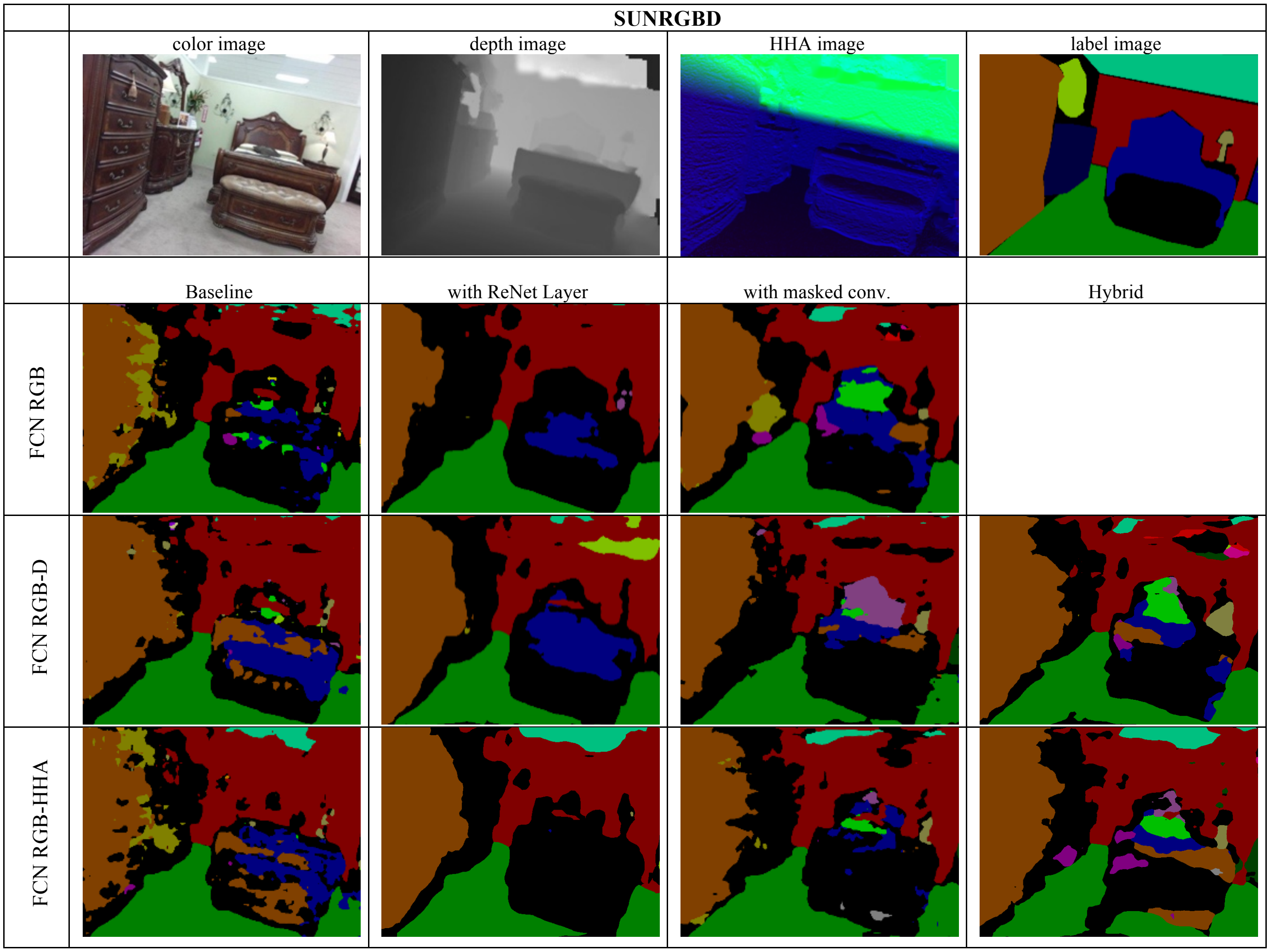


Table 6 SUNRGBD mean IoU				
	Baseline	with ReNet layer	with masked convolution	Hybrid method
FCN RGB	0.18282	0.18140	0.20674	
FCN RGB-D	0.21219	0.20585	0.21623	0.22468
FCN RGB-HHA	0.19939	0.21550	0.20861	0.22995

Table 7 SUNRGBD mean accuracy				
	Baseline	with ReNet layer	with masked convolution	Hybrid method
FCN RGB	0.24742	0.24845	0.28679	
FCN RGB-D	0.28543	0.28500	0.31475	0.31371
FCN RGB-HHA	0.26992	0.30084	0.29126	0.32084

Table 8 SUNRGBD overall accuracy				
	Baseline	with ReNet layer	with masked convolution	Hybrid method
FCN RGB	0.57705	0.57918	0.59035	
FCN RGB-D	0.59416	0.59669	0.56568	0.59358
FCN RGB-HHA	0.59615	0.59898	0.58167	0.59600

Acknowledgement

This research was partially supported by the ‘Brain Korea 21 Plus Project’ of the Ministry of Education & Human Resources Development, Republic of Korea (ROK) [F16SN26T2205], and partially supported by the ‘Civil-Military Technology Cooperation Program’ grant funded by the Korea government.

References

[1] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651, 2017.

[2] F. Visin et al., “ReSeg : A Recurrent Neural Network-based Model for Semantic Segmentation,” IEEE Conf. Comput. Vis. Pattern Recognit. Workshop, pp. 1–13, 2016.

[3] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, “LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling,” European Conference on Computer Vision (ECCV), vol. 9906, pp. 541–557, 2016.

[4] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, “FuseNet : Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture,” Asian Conference on Computer Vision (ACCV), 2016.

[5] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel Recurrent Neural Networks,” Int’l Conf. on Machine Learning (ICML), vol. 48, pp. 1747–1756, 2016.

[6] F. V. Politecnico et al., “ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks.”, Arxiv, 2015.

[7] S. Song, S. P. Lichtenberg, and J. Xiao, “SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 567–576, 2015.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," MICCAI 2015

[9] G. Huang, Z. Liu, L. Maaten, “Densely Connected Convolutional Networks,” Computer Vision and Pattern Recognition(CVPR), 2017

Author Biography

Dong-Won Shin received his B.S. in computer engineering from the Kumoh National Institute of Technology, Gumi, Korea (2013) and his M.S. in School of Information and Communications from Gwangju Institute of Science and Technology, Gwangju, Korea (2015). He is currently a Ph. D student. His research interests include 3D computer vision and machine learning.

Chan-Young Sohn received his B.S. in computer engineering from the Sejong University, Seoul, Korea (2015). He is currently an M.S. student at the Gwangju Institute of Science and Technology, Gwangju, Korea. His research interests include computer vision and machine learning.

Jun-Young Park received his B.S. in information and telecommunicaitons engineering from the Suwon University, Suwon, Korea (2018). He is currently an M.S. student at the Gwangju Institute of Science and Technology, Gwangju, Korea. His research interests include computer vision and machine learning.

Yo-Sung Ho received his B.S. in electronic engineering from the Seoul National University, Seoul, Korea (1981) and his Ph.D. in electrical and computer engineering from the University of California, Santa Barbara (1990). He worked in Philips Laboratories from 1990 to 1993. Since 1995, he has been with the Gwangju Institute of Science and Technology, Gwangju, Korea, where he is currently a professor. His research interests include image analysis, 3D television, and digital video broadcasting.

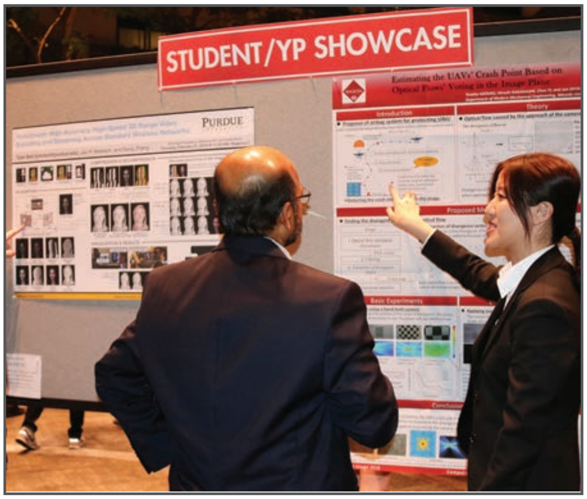
JOIN US AT THE NEXT EI!

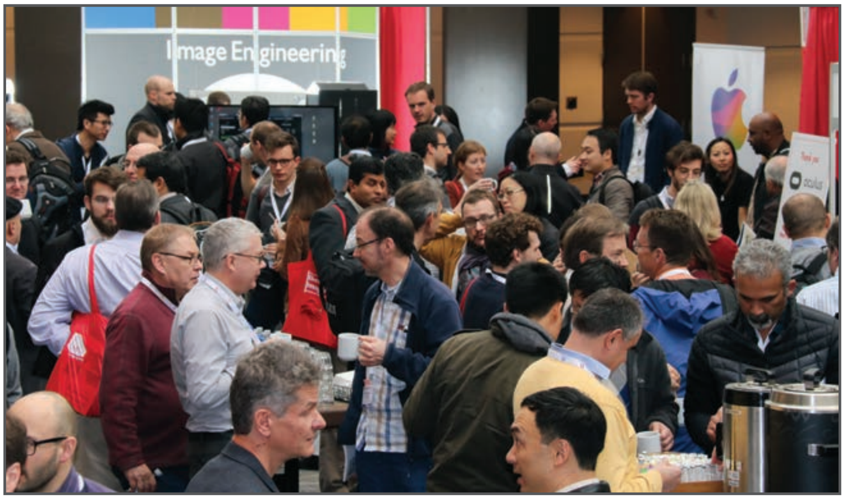
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!


STUDENT/YP SHOWCASE
Estimating the UAVs' Crash Point Based on Optical Flows' Voting in the Image Plane
PURDUE



AR is a fundamental and transformative interface and platform.

20+ years from my initial working demo to today's AR displays

Optimistic that it won't take another 20 to reach ubiquitous consumer AR!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org