HIERARCHICAL JOINT-GUIDED NETWORKS FOR SEMANTIC IMAGE SEGMENTATION

Chien-Yao Wang, Jyun-Hong Li, Seksan Mathulaprangsan, Chin-Chin Chiang, and Jia-Ching Wang

Department of Computer Science and Information Engineering, National Central University, Taiwan

ABSTRACT

Semantic image segmentation is now an exciting area of research owing to its various useful applications in daily life. This paper introduces a hierarchical joint-guided network (HJGN) which is mainly composed of proposed hierarchical joint learning convolutional networks (HJLCNs) and proposed joint-guided and making networks (JGMNs). HJLCNs exhibit high robustness in the segmentation of unseen objects that are not contained in training categories. JGMNs are very effective in filling holes and preventing incorrect segmentation predictions. The proposed HJGNs outperform the state-of-the-art methods on the PASCAL VOC 2012 testing set, reaching a mean IU of 80.4%.

Index Terms— semantic image segmentation, hierachical joint-guided networks, hierachical joint learning convolutional networks, joint-guided and masking network

1. INTRODUCTION

Image segmentation techniques are now widely used in the field of computer vision. In particular, image semantic segmentation supports real-life applications, including photograph classification, foreground and background separation, road barrier detection, autopilot, smart monitoring, object detection, object tracking, and others. Image semantic segmentation can not only realize what objects are in an image but also locate objects in an image. In recent years, many investigations of image semantic segmentation and related topics have been published.

Most early image semantic segmentation methods firstly segment the regions of the image and then classify these regions using the classifier [1, 2]. Recently, a convolutional neural network (CNN) is used to segment and classify objects simultaneously. Hariharan *et al.* [3] use two region-based CNNs [4] to perform image segmentation and classification: one is used to classify the regions in an image and the other network identifies the foreground region. Convolutional feature masking [5] supports simultaneous detection and segmentation by exploiting the concept of spatial pyramid pooling [6] as a rectangular mask. This method uses CNN for generating irregularly shaped masks. Owing to its high effectiveness, this technique has been widely used [7–9].

Fully convolutional networks (FCN) [10] use convolutional layer to substitute the fully connected layer in a network, and utilize up-sampling to make the CNN be able to do pixel-wise classification. However, up-sampling causes a loss of spatial information. Long *et al.* [10] presented the skip pixel algorithm that combines the final prediction layer with lower layers with finer strides, which can reach the third pooling layer. Huang *et al.* [9] presented the object boundary-guided semantic segmentation network by simplifying the problem of semantic segmentation into a three-category problem by separating an image into a background, objects, and boundaries. This model makes the skip pixel algorithm can reach the second pooling layer, yielding a precise area of objects to produce masks for semantic segmentation. To solve the semantic segmentation problem, de-convolutional networks [11, 12], and dilated convolutional networks [13] have also been proposed.

Unlike methods that use features to segment objects, the segmentation approaches that involve deep neural networks (DNNs) and are based on regression typically yield rough segmentation results. Therefore, the results obtained using these methods require adjustment. One of the reasons is that a pixel in a pooling layer corresponds to a large receptive field in the original image. The lack of information between adjacent pixels causes pixel-wise CNNs to yield fragmentary results. To solve this problem, the conditional random field (CRF) [14, 15] and the domain transform [16], formulated as a recurrent neural network (RNN) to train an end-to-end system, have been proposed.

Motivated by the above studies and the object boundary-guided semantic segmentation network, this work proposes hierarchical joint-guided networks (HJGNs) that mainly compose of proposed hierarchical joint learning convolutional networks (HJLCNs) and proposed joint-guided and masking networsk (JGMNs). HJLCN has a great predictive capacity which can make skip pixel algorithm reach the first convolutional layer. To solve the problem that the FCN yields rough and fragmentary results, JGMN develops a novel object and boundary joint-guided mode. Boundary prediction will be used to predict the boundaries of target objects, and object prediction will be used to fill holes of fragmentary segmentation results. Compare with [17–20], proposed HJHN achieves state-of-the-art results on PASCAL VOC 2012 [21] dataset.

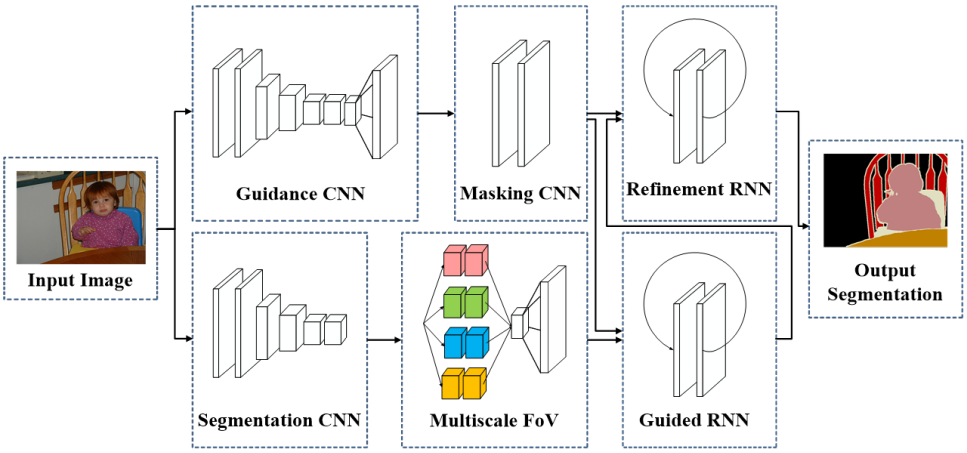

Fig. 1. Overview of proposed end-to-end semantic image segmentation system.

2. SYSTEM OVERVIEW

Figure 1 presents the proposed hierarchical joint networks, which comprise two main parts. The first part is HJLCN, which is formed using lower-level semantics, objects, boundaries, background, and higher-level semantics. The second part is JGMN, which restricts the predicted pixels of objects by boundaries from the inside and fill the predicted area of objects from the inside. The guidance CNN and the segmentation CNN use the proposed HJLCN architecture, as well as the masking CNN and the guided RNN use the JGMN with joint domain transform. The obtained outputs are adjusted using CRF-RNN [14] to generate the final segmentation results.

3. PROPOSED METHOD

3.1. Generate Object, Boundary, and Background

The proposed image semantic segmentation system requires information about objects and boundaries for guidance and masking. This paper uses morphology [22] to produce objects, boundaries, and background labels. Suppose that object labels of various classes are denoted as $L_i^{categories}$, and i = 1, 2, ... , N, where N denotes the number of categories.

$$L^{object} = Union \text{ of } L_i^{categories} \tag{1}$$

$$L^{boundary} = Dilation(L^{object}, SE) - L^{object} \tag{2}$$

$$L^{background} = U - L^{boundary} - L^{object} \tag{3}$$

where L^{object} is the label of an object; $L^{boundary}$ is the label of a boundary; $L^{background}$ is the label of the background; U is all pixels of the image, and SE is the structural element that is used in the computation of $Dilation$.

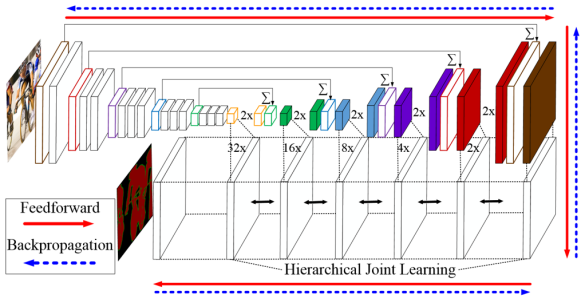

Fig. 2. Proposed hierarchical joint learning convolutional network (HJLCN).

3.2. Hierarchical Joint Learning Convolutional Network

According to observations made herein concerning the computation of the hierarchical skip pixel algorithm in the FCNs, the performance is almost dominated by the lower-level prediction layer. Figure 2 presents the proposed HJLCN. The proposed hierarchical joint learning method yields final output results by applying convex combination to predictions made by the FCN at every layer; doing so is the equivalent of performing convolution with l_1 normalization on the concatenation of all FCN prediction layers. In HJLCN, each layer can share learned attributes with each other layers, so it can be connected hierarchically to a single pixel perception field to increase the accuracy of segmentation.

3.3. Semantic Image Segmentation Network

The CNN using the atrous algorithm [17] is adopted as semantic segmentation network. While skip pixel using multiscale upsampling, which will causes a loss of spatial information, to promote accurate prediction of semantic segmentation, dilated convolution yields the prediction result of multiscale perception field without distorting spatial information.

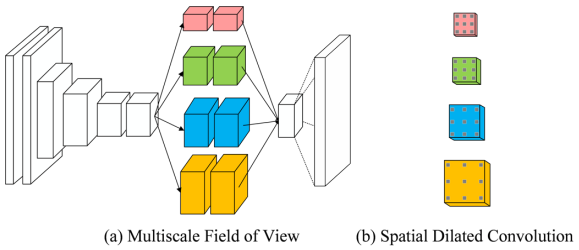

Fig. 3. Spatial dilated convolutional layer for expanding multiscale receptive field of view.

3.4. Joint-Guided and Masking Networks

In JGMN, proposed joint-guided domain transform uses the object activation status x^o as soft mask of the output of the semantic image segmentation network to obtain the initial semantic image segmentation activation status y, which is transformed into a sequential one-dimensional signal. The boundary activation status x^b and the object activation status x^o of the proposed object boundary prediction (OBP) network are used to perform joint guidance by Eq. 4.

$$y_i = \alpha_i y_{i-1} + \beta_i x_i^b + \gamma_i x_i^o \tag{4}$$

where i is an index of the sequential 1-D signal, α_i, β_i, and γ_i are the weights of y_{i-1}, x_i^b, and x_i^o, respectively, and $\alpha_i + \beta_i + \gamma_i = 1$.

Since Eq. 4 can be rewritten as Eq. 5 and Eq. 6, the JGMN can be reformulated as a Convolutional-RNN with a one-layer CNN and a one-layer gated recurrent unit RNN.

$$y_i = \alpha_i y_{i-1} + (1-\alpha_i)\left(\frac{\beta_i}{1-\alpha_i} x_i^b + \frac{\gamma_i}{1-\alpha_i} x_i^o\right) \tag{5}$$

$$\frac{\beta_i}{\beta_i+\gamma_i} x_i^b + \frac{\gamma_i}{\beta_i+\gamma_i} x_i^o = \omega_i x_i^b + (1-\omega_i) x_i^o \tag{6}$$

where $\omega_i = \frac{\beta_i}{\beta_i+\gamma_i}$. Figure 4 presents the back-propagation path of the proposed JGMN. After JGMN is used to fill holes and to eliminate noise, the dense CRF [14] is used to perform the final adjustment and obtain the segmentation result.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

The PASCAL VOC 2012 [21] which consists of 20 categories of foreground objects, and one category of background is used for evaluating proposed method. The FCN [10] was used to initialize weights and OBP-HJLCN was compared with OBP-FCN [9]. The initialization setup of Chen *et al.* [16] was used to compare the results of domain transform JGMN (DT-JGMN) with those of DT-EdgeNet [16], and the ResNet101 pre-trained model of DeepLabv2 [17] was used to initialize weights of proposed HJGN.

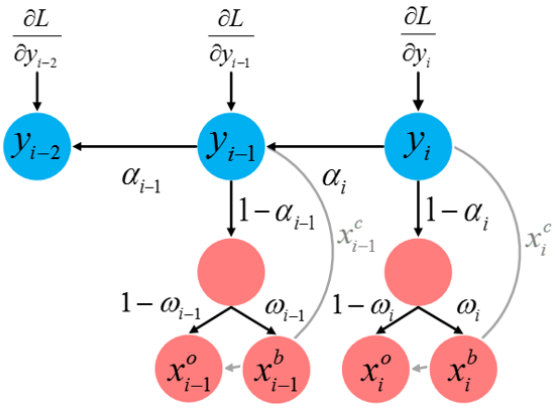

Fig. 4. Backpropagation path of JGMN.

4.2. Experimental Results

4.2.1. *Comparison of OBP-FCN and proposed OBP-HJLCN*

The first experiment was conducted to compare the performance of the proposed OBP network, OBP-HJLCN, with that of OBP-FCN [9], as presented in Fig. 5. For the proposed OBP-HJLCN, both the completeness of an object region and the continuity of the boundaries were excellent.

Observations of the OBP-HJLCN, presented in Fig. 6, demonstrates that the proposed method is powerful when applied to images which contain unseen categories of objects. Although information about birdcage and barrier were not provided while training, remarkable predictions of objects and boundaries of them were obtained in OBP-HJLCN.

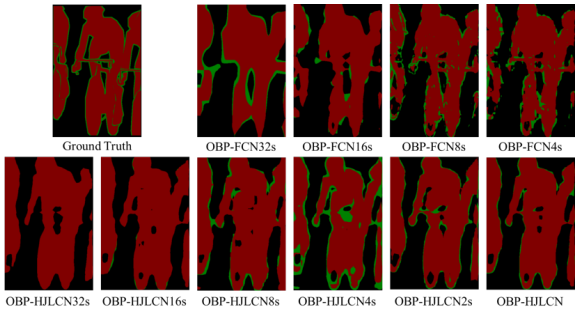

Fig. 5. Results obtained using OBP-FCN and OBP-HJLCN.

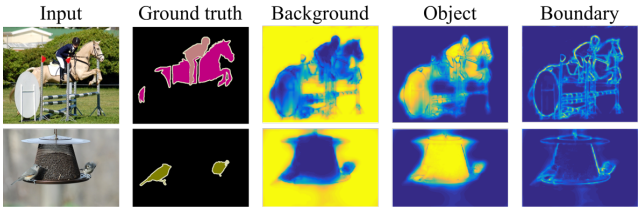

Fig. 6. Results obtained using OBP-HJLCN.

4.2.2. *Comparison of DT-EdgeNet and proposed DT-JGMN*

An experiment was conducted to compare segmentation results obtained using edge guidance [16] with those obtained by the proposed boundary-guided segmentation, which is presented in Fig. 7. DT-EdgeNet [16] used the output of edge prediction network as the input to the domain transform, it preserves the edges of objects and background. OBP-HJLCN was used herein in place of the edge prediction network of DT-EdgeNet, it includes only edges around objects. Therefore, domain transform can eliminate predictions of non-target objects in proposed DT-JGMN.

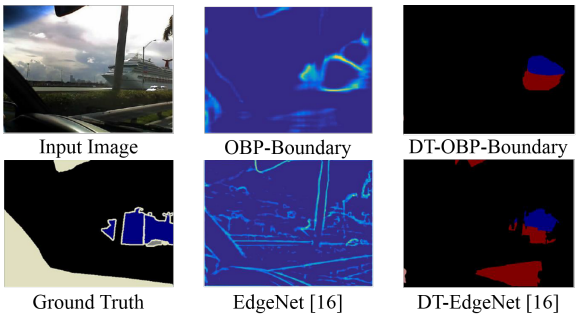

Fig. 7. Results obtained using DT-EdgeNet and DT-JGMN.

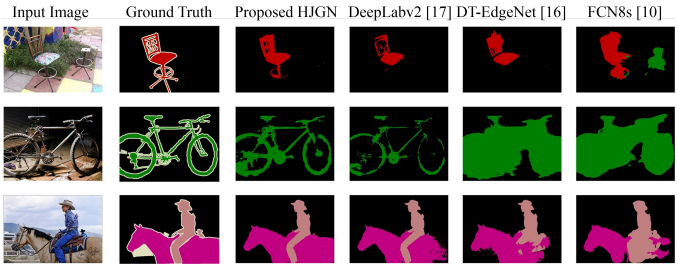

Fig. 8. Results obtained using PASCAL VOC 2012 validation set.

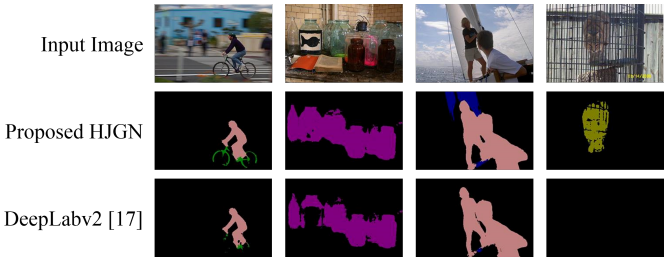

Fig. 9. Results obtained using PASCAL VOC 2012 test set.

4.2.3. *Comparison of state-of-the-art with proposed HJGN*

Figures 8 and 9 present the results of semantic image segmentation using the PASCAL VOC 2012 validation set and testing set. Figure 9 reveals that the proposed HJGN can find objects, such as a boat behind a person and a cat in a cage, while that DeepLabv2 cannot. In Table 1, the results obtained using the proposed HJGN are compared to those obtained using other approaches in the literature, based on ResNet 101, and the proposed HJGN is found to outperform all state-of-the-art networks and yields a mean IU of 80.4% on the PASCAL VOC 2012 test set.

Table 1. Comparison of state-of-the-art approaches for doing semantic image segmentation with proposed HJGN when applied to PASCAL VOC 2012.

Categories	Proposed HJGN	CentraleSupelec Deep G-CRF [18]	DeepLabv2 [17]	LRR-4x-ResNet-COCO [19]	Adelaide-VeryDeep-FCN-VOC [20]
aeroplane	92.7	**92.9**	92.6	92.4	91.9
bicycle	54.8	**61.2**	60.4	45.1	48.1
bird	91.6	91.0	91.6	**94.6**	93.4
boat	68.0	66.3	63.4	65.2	**69.3**
bottle	76.9	**77.7**	76.3	75.8	75.5
bus	**95.7**	95.3	95.0	95.1	94.2
car	**89.3**	88.9	88.4	89.1	87.5
cat	92.6	92.4	92.6	92.3	**92.8**
chair	35.2	33.8	32.7	**39.0**	36.7
cow	**89.0**	88.4	88.5	85.7	86.9
table	69.3	69.1	67.6	**70.4**	65.2
dog	89.4	**89.8**	89.6	88.6	89.1
horse	92.7	**92.9**	92.1	89.4	90.2
motorbike	**87.9**	87.7	87.0	88.6	86.5
person	**87.5**	**87.5**	87.4	86.6	87.2
plant	**66.8**	62.6	63.3	65.8	64.6
sheep	88.5	89.9	88.3	86.2	**90.1**
sofa	**62.2**	59.2	60.0	57.4	59.7
train	86.1	**87.1**	86.8	85.7	85.5
monitor	76.2	74.2	74.5	**77.3**	72.7
Mean IU	**80.4**	80.2	79.7	79.3	79.1

5. CONCLUSIONS

This paper proposes the HJGN which consists of two novel DNN architectures, HJLCN and JGMN. The HJLCN uses hierarchical joint learning to develop FCN skip architecture that can provide a pixel-wise field of view. The JGMN uses objects and boundaries to guide semantic image segmentation by joint domain transformation. The proposed OBP networks have potential for the segmentation of unseen data. Overall, the proposed HJGN achieves a mean IU of 80.4% on the PASCAL VOC 2012 test set.

6. REFERENCES

[1] P. Arbelez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik, "Semantic segmentation using regions and parts," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," *European Conference on Computer Vision (ECCV)*, 2012.

[3] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," *European Conference on Computer Vision (ECCV)*, 2014.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," *arXiv: 1412.1283*, 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2015.

[7] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] G. Papandreou, L. Chen, K. Murphy, and A. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[9] Q. Huang, C. Xia, W. Zheng, Y. Song, H. Xu, and C. Kuo, "Object boundary guided semantic segmentation," *arXiv:1603.09742v4*, 2016.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision (IJCV)*, 2013.

[12] J. Pont-Tuset, P. Arbelaez, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *arXiv:1503.00848*, 2015.

[13] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[14] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[15] A. Arnab, S. Jayasumana, S. Zheng, and P. Torr, "Higher order conditional random fields in deep neural networks," *arXiv:1511.08119v4*, 2016.

[16] L. Chen, J. Barron, G. Papandreou, K. Murphy, and A. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[18] S. Chandra and I. Kokkinos, "Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs," *arXiv:1603.08358v2*, 2016.

[19] G. Ghiasi and C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," *arXiv:1605.02264v2*, 2016.

[20] Z. Wu, C. Shen, and A. Hengel, "High-performance semantic segmentation using very deep fully convolutional networks," *arXiv:1604.04339v1*, 2016.

[21] L. Van Gool M. Everingham and, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *IJCV*, 2010.

[22] R. Haralick, S. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1987.