Deep Aggregation Net for Land Cover Classification

Tzu-Sheng Kuo[1]*, Keng-Sen Tseng[1]*, Jia-Wei Yan[2]*, Yen-Cheng Liu[2], Yu-Chiang Frank Wang[1,2]
[1]Department of Electrical Engineering, National Taiwan University
[2]Graduate Institute of Communication Engineering, National Taiwan University
{b03901032, b03901154, r06942033, r04921003, ycwang}@ntu.edu.tw

Abstract

Land cover classification aims at classifying each pixel in a satellite image into a particular land cover category, which can be regarded as a multi-class semantic segmentation task. In this paper, we propose a deep aggregation network for solving this task, which extracts and combines multi-layer features during the segmentation process. In particular, we introduce soft semantic labels and graph-based fine tuning in our proposed network for improving the segmentation performance. In our experiments, we demonstrate that our network performs favorably against state-of-the-art models on the dataset of DeepGlobe Satellite Challenge, while our ablation study further verifies the effectiveness of our proposed network architecture.

1. Introduction

Land cover information is important for various applications, such as monitoring areas of deforestation and urbanization. To recognize the type of land cover (e.g., areas of urban, agriculture, water, etc.) for each pixel on a satellite image, land cover classification can be regarded as a multi-class semantic segmentation task [11, 8, 15].

With the availability of abundant segmentation images and recent advances in deep neural networks, several CNN-based models [3, 4, 9, 2, 12, 10, 14] have demonstrated the effectiveness on semantic segmentation. For example, several works [2, 12, 9] adopt encoder-decoder structures to take a RGB image as input and predict its corresponding semantic mask. To capture global context information, Zhao *et al.* [14] incorporate multi-scale features with a pyramid pooling module [7]. Similarly, DeepLabv3 [3] exploits atrous convolution with multiple rates and image-level features to improve the prediction performance. DeepLabv3+ [4] further extend DeepLabv3 with an additional decoder to refine segmentation results along the object boundaries. A common approach adopted by the above models is to aggregate different-level features in the procedure of prediction. However, as pointed out in [13], simply applying skip connections from low- to high-level layers may not fuse the spatial and semantic information in an effective manner.

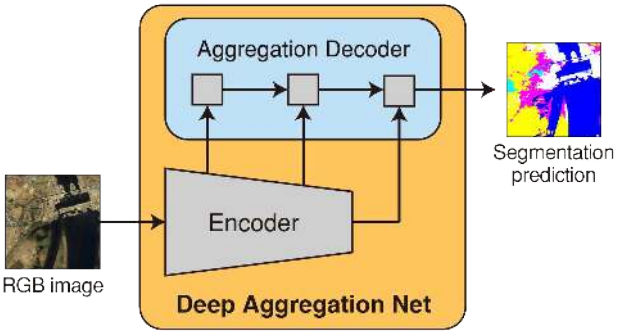


Figure 1: Illustration of deep aggregation net. Note that our model takes a RGB image as input and predicts the semantic segmentation output.

Inspired by [13], with the goal of incorporating various information across layers in the procedure of semantic segmentation, we introduce an aggregation decoder in combination with DeepLabv3 model. Specifically, our model combines different-level features progressively from the encoder for final prediction. On the other hand, we observe two properties of land cover images: 1) there are no clear boundaries across different types of land cover and 2) the area of all types of land cover are not fragmented. Based on these properties, we improve segmentation results by softening one-hot labels in ground truth masks and by removing fragmented land covers in predicted masks.

In summary, our contributions are listed as follows:

- We proposed deep aggregation net for land cover segmentation, which exploits semantic information across image scales for improved segmentation.
- We utilize soft semantic labels and graph-based fine tuning in our proposed network. Our ablation studies further verify the effectiveness of our proposed model.

*equal contribution

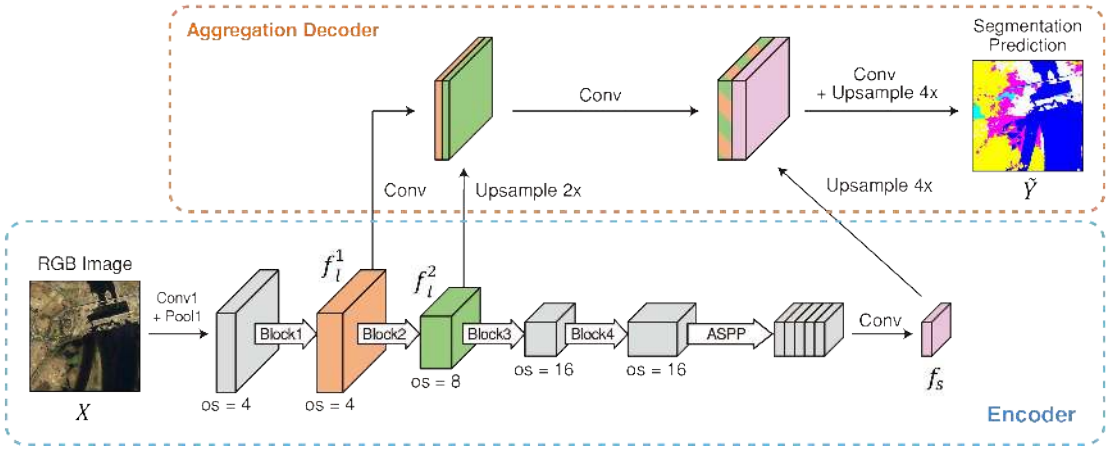


Figure 2: Architecture of our deep aggregation net. The rectangular boxes represent tensor features and the arrows denote operations. Blocks 1 to 4 are residual convolutional blocks, and ASPP indicates Atrous Spatial Pyramid Pooling. Features f_l^1 and f_l^2 are extracted before the strided convolution layer (stride = 2) in Blocks 1 and 2, respectively. Each tensor feature is specified with its output-stride (os), which denotes the ratio of input image spatial resolution to the feature resolution. During aggregation decoding, tensor features in smaller scales are bilinearly upsampled before concatenation.

2. Proposed Method

Given pairs of RGB satellite images and land cover masks $\{X, Y\}$, we aim at training a model to produce land cover segmentation prediction $\tilde{Y}$. In this section, we first describe the proposed network, and further describe the details of soft semantic labels and graph-based fine tuning.

2.1. Proposed Architecture of Deep Aggregation Net

As depicted in Figure 2, our model applies an encoder-decoder structure to perform semantic segmentation. Here we adopt DeepLabv3 [3], which applies atrous convolution to cascaded ResNet and a pyramid pooling module, as our encoder to extract multi-level features across different layers. Our model further combines these features consecutively from lower to higher levels. To be more detailed, we first concatenate two low-level features f_l^1 and f_l^2 extracted from the encoder, then feed them into convolution layers to produce a fused feature. Next, we concatenate the fused feature with the semantic feature f_s, and introduce the final convolution layers followed by up-sampling to obtain segmentation masks.

Our model takes a RGB image X as input, and produces a segmentation mask $\tilde{Y}$. The training loss function $\mathcal{L}_{seg}$ for semantic segmentation is thus defined as below:

$$\mathcal{L}_{seg} = \mathcal{H}(Y, \tilde{Y}), \tag{1}$$

where $\mathcal{H}$ denotes the cross-entropy loss and Y denotes the ground truth segmentation annotation.

2.2. Soft Semantic Labels for Segmentation

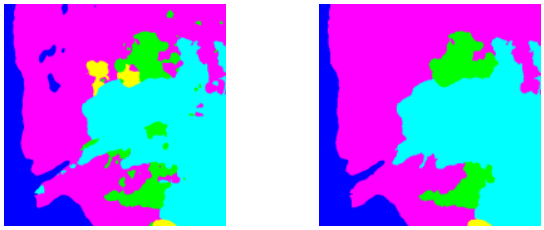
With the observation that there is no clear boundary between two different land cover regions within a RGB satellite image, we choose to smooth spatial boundaries across such regions during semantic segmentation. Toward this end, we convert one-hot label segmentation mask Y into soft label segmentation mask Y_s by applying Gaussian filtering on each channel independently. We also apply normalization to ensure class weights of each pixel sum to one. Therefore, our training loss can be modified as:

$$\begin{aligned} \mathcal{L}_{seg}^{soft} &= \mathcal{H}(Y_s, \tilde{Y}) \\ &= \mathcal{H}(g(Y), \tilde{Y}), \end{aligned} \tag{2}$$

where $g(\cdot)$ denotes two-dimensional Gaussian filtering with standard deviation σ along with pixel-wise normalization.

2.3. Graph-based Fine Tuning

To prevent fragmented segmentation prediction, we employ graph-based fine tuning to refine our final prediction. Here we consider the segmentation prediction $\tilde{Y}$ as an undirected graph, where pixels are nodes, while the edges are connected between adjacent pixels with the same class. We run breadth-first search (BFS) on segmentation prediction $\tilde{Y}$ to detect groups of connected pixels with a same class. Groups with the pixel number fewer than a threshold value T are reassigned labels of their neighbor pixels. For simplicity, the neighboring pixel is defined as the pixel next to the top-left pixel in a group. Note that our graph-based fine tuning is applied after model prediction and do not increase computation cost during training.



(a) (b)

Figure 3: Example segmentation results (a) without and (b) with graph-based fine-tuning.

3. Experiment

We first compare the proposed method to the existing models. We then conduct an ablation study to verify each module of our proposed method.

3.1. Dataset and Evaluation

Here we use the dataset provided by the organizer of CVPR 2018 DeepGlobe Satellite Challenge [5]. The training set contains 803 RGB satellite images with size 2448×2448 pixels as well as 50cm pixel resolution. Each satellite image is paired with a ground truth mask for land cover type annotation. Seven types of land covers are included: urban, agriculture, rangeland, forest, water, barren, and unknown. Models are evaluated on the validation set with 171 satellite images segmentation pairs. The evaluation metric is pixel-wise mean Intersection over Union (mIoU) score.

3.2. Implementation Details

We implemented our network on Tensorflow [1]. The backbone encoder was pre-trained on ILSVRC-2012-CLS dataset [6]. During the training process, we adopted polynomial learning rate decay with decay rate 0.9, applied batch normalization to convolution layers, randomly left-right flipped, and cropped input images size from 2448×2448 to 512×512. The standard deviation σ of the Gussian filter was set to 8 for soft semantic labels. The threshold value T for graph-based fine tuning was empirically set to 8000 pixels, which corresponded to 2000 m^2 in real world. It took roughly 10 hours to train our network for 300 epochs with batch size 10 on a single Nvidia GeForce GTX 1080Ti.

3.3. Comparison and Ablation Study

As shown in Table 1, we compare our deep aggregation net with the existing models [10, 3, 4]. Our model improves FCN [10], DeepLabv3 [3], and DeepLabv3+ [4] by 14.9%, 12.7%, and 3.4%, respectively. This demonstrates the effectiveness of the proposed deep aggregation decoder, soft semantic labels, and graph-based fine tuning.

Table 1: Performance comparisons of semantic segmentation in mIoU.

Architecture	mIoU
FCN-32s [10]	0.4588
DeepLabv3 [3]	0.4679
DeepLabv3+ [4]	0.5101
Ours	**0.5272**

Table 2: Ablation study of our deep aggregation net. Note that, for simplicity, we fixed $\sigma = 8$ for soft semantic labels and did not fine-tune the results.

Aggregation Decoder	Soft Semantic Labels	Graph-based Fine Tuning	mIoU
			0.5101
✓			0.5259
	✓		0.5187
		✓	0.5116
✓	✓		0.5261
✓		✓	0.5292
	✓	✓	0.5190
✓	✓	✓	0.5272

To further understand the advantage of our model against the others, we show some qualitative results in Fig 4. FCN shows the ability to output segmentation masks with consistency over a large area, but falls short at details such as smaller areas and boundaries. DeepLabv3 and DeepLabv3+ improve performance on these details; however, they also produce excessive fragments and fail to maintain consistency at larger areas in some cases. Beyond the models mentioned above, our model combines multi-level features effectively and produces more accurate segmentation results at both larger and detail areas.

To verify each module of the proposed method, we also present an ablation study in Table 2. First, solely applying aggregation decoder, soft semantic labels, or graph-based fine tuning is able to outperform the DeepLabv3+ model. Among these modules, we observe that the proposed aggregation decoder improves performance the most. Second, our model can further improve the performance by applying two of these modules. This indicates that these approaches are compatible with each other. Finally, we observe that the model adopting three modules (0.5272) are slightly behind the model without soft semantic labels (0.5292). According to our empirical study, one potential reason is that standard deviation of Gaussian filter for soft semantic labels is fixed for all experiments, and this hyper-parameter could be learned in the future study.

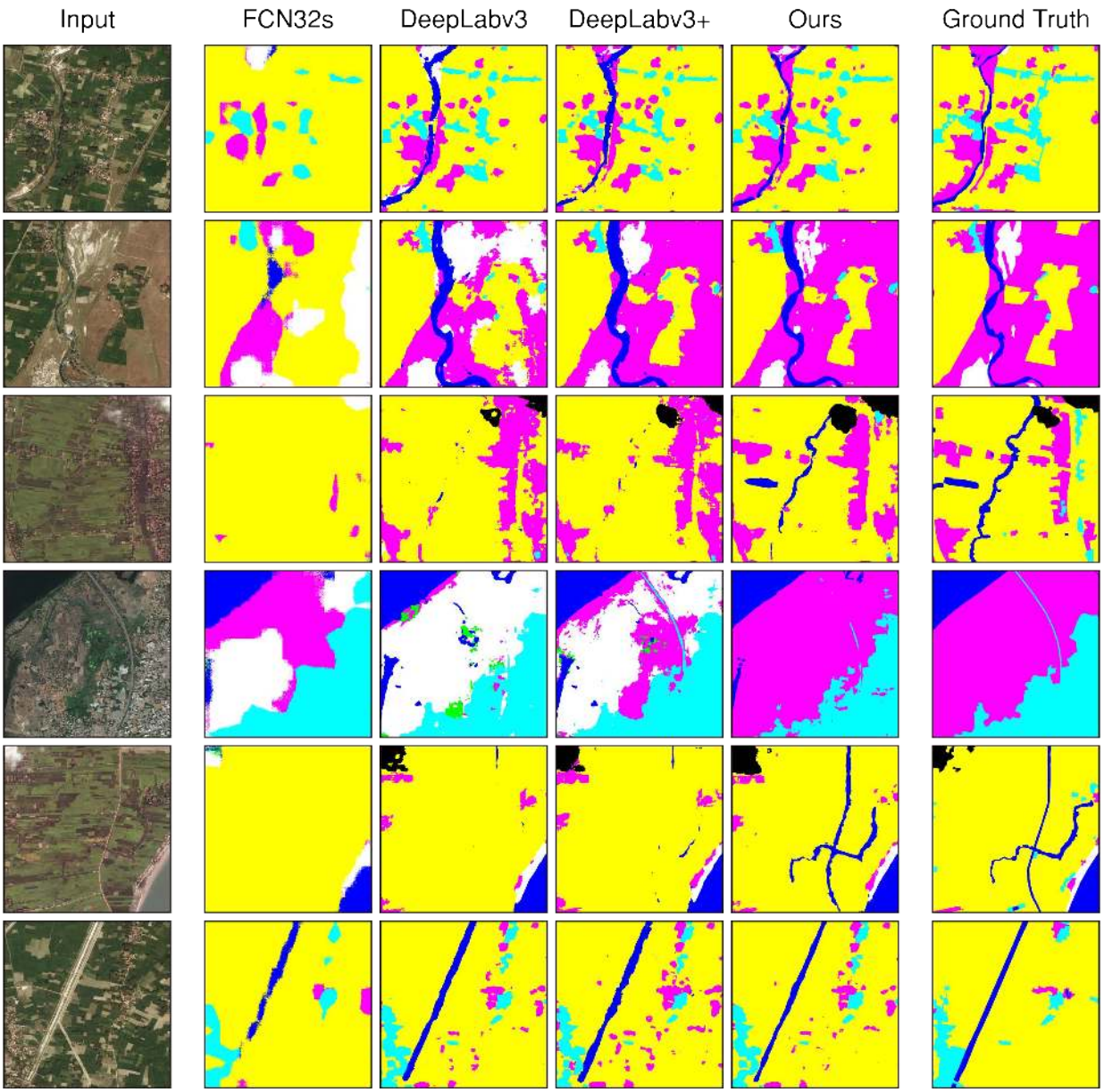


Figure 4: Example segmentation results using different models. Note that our model was able to accurately classify pixels over different scales/regions.

4. Conclusion

We presented deep aggregation net which effectively incorporates the features extracted from different layers. We also introduced soft semantic labels and graph-based fine tuning to improve the performance of our proposed model. In the experiment, we verified the effectiveness of our proposed modules and demonstrated that our model perform satisfactory result against the state-of-the-art models on the dataset of DeepGlobe Satellite Challenge.

Acknowledgments This work was supported in part by the Ministry of Science and Technology of Taiwan under grant MOST 107-2634-F-002-010.

References

[1] M. Abadi and A. Agarwal *et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 3

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image

segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 1

[3] L.-C. Chen, G. Papandreou, F. Schoff, and H. Adam. Rethinking atrous convolution for semantic image segmentaion. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2, 3

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018. 1, 3

[5] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018. 3

[6] J. Deng, W. Dong, R. Socher, L. Li-Jia, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 3

[7] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014. 1

[8] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler. Learning aerial image segmentation from online maps. volume 55, pages 6054–6068. IEEE, 2017. 1

[9] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[10] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 3

[11] D. Marmais, J. Wdgner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla. Semantic segmentation of aerial images with an ensemble of cnns. In *International Society for Photogrammetry and Remote Sensing (ISPRS)*, 2016. 1

[12] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Interventions (MICCAI)*, 2015. 1

[13] F. Yu, D. Wang, E. Shelhamer, and T. Darrell. Deep layer aggregation. *arXiv preprint arXiv:1707.06484*, 2018. 1

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[15] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. volume 5, pages 8–36. IEEE, 2017. 1