A COMPARISON OF DEEP LEARNING ARCHITECTURES FOR SEMANTIC MAPPING OF VERY HIGH RESOLUTION IMAGES

Qinghui Liu[1], Arnt-Børre Salberg[1] and Robert Jenssen[2]

***Abstract*— Semantic mapping of land cover is a key, but challenging, problem in remote sensing. Recent advances in deep learning, especially deep convolutional neural networks (CNNs), have shown outstanding performance in this task. In order to develop refined deep learning pipeline for meeting the rising need for accurate semantic mapping in remote sensing images, this paper study and compare a number of advanced deep learning segmentation architectures, which have obtained state-of-the-art results on computer vision contests like the Pascal VOC. To further analyze and compare the effectiveness of some elaborate layers and underlying structures introduced by these architectures, we evaluate them by re-implementing, train and test them on ISPRS Potsdam dataset. Our results show that a promising performance with overall F1_score above 87% and mIoU of 79% can be obtained by only using the RGB images, without any post-processing such as conditional random field (CRF) smoothing. At last, we propose several possible approaches to further enhance the deep learning architectures to better deal with high-resolution aerial images. We therefore consider this work to be helpful for the remote sensing research community.**

I. INTRODUCTION

Semantic mapping has been one of the most active research topics in remote sensing in the past decades, and is the key problem for land cover mapping, object detection and change detection in high resolution aerial or satellite images. In recent years, deep learning techniques have emerged as the dominating methods for image classification and segmentation [1], and have also gained increasing interest in remote sensing [2], [3]. Many deep architectures, such as Fully Convolutional Networks (FCN) [4] , UNet[5] and SegNet [6], have been adapted in the ISPRS semantic segmentation challenge [7] and achieved outstanding performance [8], [9], [10].

In 2017, significant progress was made in the Pascal VOC[11] scene segmentation challenge. A number of new deep architectures, such as Pyramid Scene Parsing (PSP) network [12], Global Convolutional Network (GCN)[13] and Dense Upsampling Convolution (DUC) network[14], have achieved the best state of the art performance and, by far, surpassed the best scores produced by the previous FCN, UNet or SegNet based methods. Although these new network architectures have shown very impressive results on natural image datasets (VOC 2012), they have not been directly applied to remote sensing images.

[1]Norwegian Computing Center, Dept. SAMBA, P.O. Box 114 Blindern, NO-0314 OSLO, Norway

[2]Department of Physics and Technology, UiT the Arctic University of Norway, Tromsø, Norway

In this work, we analyze and compare the effectiveness of some elaborate layers and underlying structures newly introduced by these architectures. Our main contributions are: 1) We present how to implement and refine the latest deep learning architectures with PSP, GCN and DUC for the task of semantic mapping in high-resolution (6000×6000 pixels) aerial images. 2) We show how to effectively address highly imbalanced classes by utilizing median frequency balancing (MFB) weights [8] in the cross-entropy loss functions. 3) We evaluate and compare the presented models on the ISPRS Semantic Labeling Challenge datasets of Potsdam using the exact same subsets of train, evaluation and test. 4) We present discussions of the experiment results by comparing them with previous state-of-the-art models (FCN, UNet and SegNet). A number of practical approaches are proposed to further enhance the deep learning pipeline in remote sensing images.

II. METHODS

A. Architectures

In this work, we mainly make use of the most recent deep learning architectures, PSPNet[12], GCN[13] and DUC[14], which have been proven to provide outstanding results on the Pascal VOC contest.

The PSP network introduces a pyramid pooling module to aggregate the context that captures global scene categories by applying large kernel pooling layers. Dilated convolutions[15] are used to modify an inherent ResNet[16], and a pyramid pooling module is added to it. The feature maps from the ResNet are concatenated with upsampled output of the parallel pooling layers with kernels covering the whole, half of and small portions of image as shown in Figure 1.

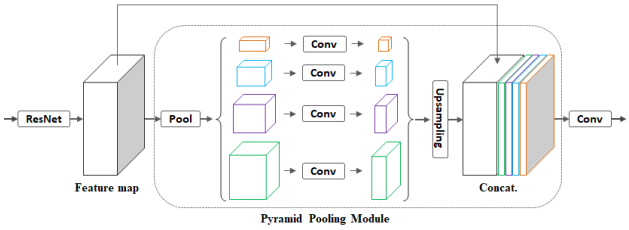

Fig. 1. PSP-ResNet architecture [12].

For the GCN architecture, a module called global convolutional network (GCN) that are based on convolutions with very larger kernels and a block called boundary refinement (BR), as shown in Figure 2, are adopted into to

a encoder-decoder pipeline. BR module utilizes a simple residual structure to refine the predictions near the object boundaries. A ResNet (without dilated convolutions) forms the encoder part, while CGNs and deconvolutions form the decoder linked with BRs as shown in Figure 3 .

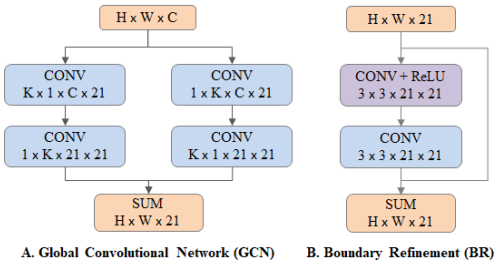

Fig. 2. The structures of GCN and BR modules [13].

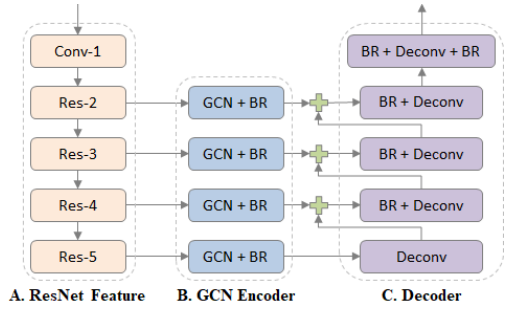

Fig. 3. Pipeline of GCN-ResNet.

For the DUC architecture, a hybrid dilated convolution (HDC) framework is adopted in the encoding stage as shown in Figure 4. HDC is used to alleviate the "gridding issue" caused by the standard dilated convolution operation [14]. The DUC module is learnable and performed on the features provided by a ResNet, which is able to capture and recover detailed information in the leaning pipeline.

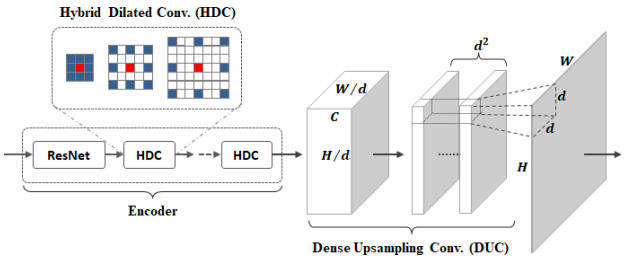

Fig. 4. DUC-ResNet architecture [14].

In addition to the three architectures described above, we also conduct primary study on the previous appealing deep learning frameworks covering FCN, U-Net and SegNet. All these network models are trained end-to-end on the same dataset (ISPRS Potsdam).

B. Data augmentation and normalization

We extract the image patches (of size 512×512) from the high resolution aerial RGB images (of size 6000×6000) with 50% overlap. We then conduct augmentation by flipping left to right and up down, rotating 90, 180 and 270 degrees which yields 8 times augmented image patches. In addition, we utilize random rotation of +10 to -10 degree and then randomly crop small patches of 256×256 for each training epoch.

We also normalize image patches with mean and standard deviation. Given mean: (R, G, B) and std: (R, G, B), we normalize each channel of the input images by the following formula, where ch corresponds to each channel ($ch \in (R, G, B)$)

$$norm_{ch} = \frac{pixel_{ch} - mean_{ch}}{std_{ch}} \tag{1}$$

C. Optimizer and weighted loss function

In our work, we choose AdaDelta [17] as the optimizer of the model, since in practical, AdaDelta seems to be "safer" because it doesn't depend so strongly on setting of learning rates, and base on our own experiments as well, it always provided the quickest convergence. In addition, Our goal is to evaluate and compare each model's performance, rather than further push the state-of-the-art results, so we perform all experiments based on the same optimizer and loss function instead of trying different ones.

As loss function, we apply a 2D cross-entropy loss function with median frequency balancing (MFB) weights as defined in the equations 2 and 3 [8]

$$W_c = \frac{\text{median}(\{f_c | c \in C\})}{f_c}, \tag{2}$$

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} {l_c}^{(n)} \log ({p_c}^{(n)}) W_c \tag{3}$$

where W_c is the weight for class c, f_c the pixel- frequency of class c, ${p_c}^{(n)}$ is the probability of sample belonging to class c, ${l_c}^{(n)}$ denotes the class label of sample n in class c.

III. EXPERIMENTS

To train and compare the presented deep architectures, we implement six different models, including U-Net, FCN8s-VGG16[18], SegNet-VGG19, GCN-ResNet50, PSP-ResNet50, and DUC-ResNet50. We train and test them on the same ISPRS Potsdam dataset.

A. Dataset

ISPRS released a benchmark dataset covering the city of Potsdam that contains 24 6000×6000 RGB images annotated by hand with six labels including impervious surfaces, buildings, tree, low vegetation, car and clutter/background. To evaluate our models, the labeled part of the dataset is divided into training validation, and test set. The training set consists of 19 images (areas: 2_10, 2_12, 3_10, 3_11, 3_12, 4_10, 4_12, 5_10, 5_12, 6_7, 6_8, 6_10, 6_11, 6_12,7_7, 7_8,

7_9, 7_10 and 7_12), the validation set of 2 images (areas: 2_11 and 4_11), and the test set contains 3 images (areas: 5_11, 6_9 and 7_11).

B. Evaluation methods

The performance is measured by both mean Intersection over Union (mIoU), and the F1-score defined by the ISPRS as the following.

$$\text{F1\_score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision $= \frac{\text{TP}}{\text{TP+FP}}$ and Recall $= \frac{\text{TP}}{\text{TP+FN}}$, TP stands for True Positive, FN- False Negative, TN- True Negative and FP- False Positive.

We train and validate these networks with 256x256 windows of data as input and batch size of 12 as well as other hyper-parameters settings, except the learning rates that we use different settings for different models. We test the trained models over a sliding window (with 33% overlapping of windows) on a high resolution aerial image and stitched the predictions together to form the whole semantic mapping picture.

C. Results

Table I shows our experiment results. The individual scores are F1-scores and the mean F1-score (mF1) is the average for the scores corresponding to all labels, except the clutter/background class. We observe that the DUC-ResNet50 model achieved the best mF1 of 88.2% and mIoU of 79.3%, though it was just slightly better than PSP-ResNet50 (mF1:87.9%, mIoU:78.9%) and SegNet-VGG19 (mF1:87.4%, mIoU:78.1%) (TableI). The PSP-ResNet50 is more accurate on small objects (car: 81.9% IoU and tree: 73.8% IoU) than other models (Table I), while DUC-ResNet50 has better predictions on big objects (building: 92.1% IoU, surface: 82.4% IoU and 68.7% IoU). Figure 5 shows a qualitative comparison of the semantic mapping results for the DUC, PSP, GCN, SegNet, FCN and UNet models on one of the test images.

We therefore believe that the use of HDC in DUC-ResNet model enlarges the receptive fields of the network, which is helpful for better recognizing multi-scale and relatively big objects (such as buildings, low-vegetation, and surfaces). As for small objects in remote sensing images such as vehicles, the PSP module seems to be an effective method to decode their global context information for better predictions, but it also leads to expensive computation cost caused by the 3x3 convolutions used to fuse different global pooling results, which could be an issue in a resource restricted environment. Comparably, GCN requires considerable less computational cost among these models and achieved good performance on most objects except on cars (IoU of 76.4%). The possible reason is that the default large kernel size (15 used in our experiments) of GCN is too big to be suitable for detecting small objects in remote sensing images.

TABLE I

SEMANTIC MAPPING RESULTS ON POTSDAM RGB TEST IMAGES (F1_SCORE AND IOU FOR EACH CLASS AND AVERAGE F1_SCORE AND MIOU FOR ALL CLASS EXCEPT THE CLUTTER CLASS). NOTE THAT FOR VGG16/19 AND RESTNET50 DEEP NETWORKS EMPLOYED IN THE PRESENTED MODELS, THEY ARE INITIALIZED BY PRE-TRAINED WEIGHTS FROM IMAGENET[19].

Models	**mF1**	**Building**	**Tree**	**Low-veg**	**Surface**	**Car**	Clutter
U-Net	0.830	0.930	0.776	0.755	0.860	0.828	0.173
FCN8s-VGG16	0.839	0.937	0.821	0.770	0.870	0.796	0.222
SegNet-VGG19	0.874	0.952	0.825	0.800	0.895	0.897	0.341
GCN-ResNet50	0.870	0.953	0.840	0.799	0.890	0.866	0.297
PSP-ResNet50	0.879	0.953	**0.848**	0.803	0.893	**0.900**	0.346
DUC-ResNet50	**0.882**	**0.959**	0.842	**0.814**	**0.903**	0.891	0.327
Models	**mIoU**	**Building**	**Tree**	**Low-veg**	**Surface**	**Car**	Clutter
U-Net	0.715	0.870	0.635	0.607	0.756	0.707	0.095
FCN8s-VGG16	0.728	0.883	0.698	0.626	0.771	0.662	0.125
SegNet-VGG19	0.781	0.909	0.702	0.668	0.810	0.814	0.206
GCN-ResNet50	0.774	0.911	0.725	0.666	0.802	0.764	0.175
PSP-ResNet50	0.789	0.910	**0.738**	0.671	0.807	**0.819**	0.209
DUC-ResNet50	**0.793**	**0.921**	0.728	**0.687**	**0.824**	0.804	0.195

IV. CONCLUSIONS

In this paper, we presented a comparative study of state of the art deep learning architectures for semantic mapping in very high-resolution aerial images. We selected and trained six different cutting-edge deep learning algorithms and evaluated and compared the performance of the models on the ISPRS Potsdam dataset. Based on the results, we conclude that the DUC model (with mF1 of 88.2% and mIou of 79.3%) is the best architecture, and we also consider that the use of HDC in DUC model is helpful for better recognizing relatively big objects (such as buildings, low-vegetation, and surfaces). We thus believe that dilated convolution based algorithms (such as HDC) are promising research direction to obtain improved results on multi-scale, variable and ambiguous objects in very high-resolution images. While global average pooling based methods (such as PSP) could be helpful for the detection of smaller objects.

We also find from the results that deeper/larger architectures (such as GCN-ResNet50) may not produce better segmentation performance than "shallower/smaller" deep networks (such as SegNet-VGG19). It may be related to that some deeper architectures are more suitable for interpreting natural imagery than parsing remote sensing images, or that we need fine-tune some key parameters (such as the kernel size) and more data to train such networks. Therefore, when using deep learning networks in remote sensing, it is important to carefully evaluate the models being applied.

It is also worth noticing that the clutter/background class got a very low score for each model, while small object such as the car and tree classes achieved much better performance. One reason for this may be that the MFB loss function re-weights each class in the cross-entropy loss according to the class-weights we set before training the networks. Thus, we consider MFB an effective method for dealing with imbalanced datasets in remote sensing. To the best of our knowledge, this is the first comparison study that focuses on remote sensing semantic mapping by using a number of leading deep learning architectures from the latest Pascal VOC contest.

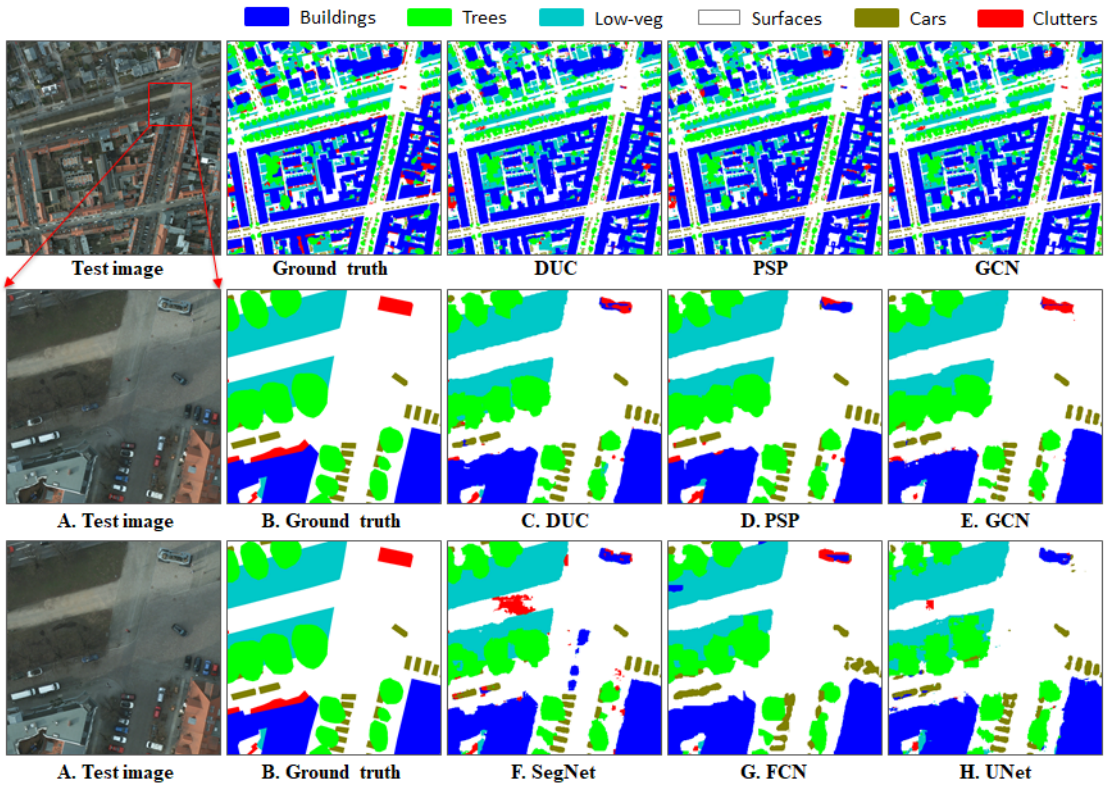

Fig. 5. The mapping results on the test image of 5_11 by using sliding window method with trained models. A. a test RBG image (6000 × 6000) of an example patch (1200 × 1200), B. ground truth, C. predictions from DUC model, D. predictions from PSP model, E. predictions from GCN model, F. predictions from SegNet model, G. prediction from FCN model, H. predictions from UNet model.

REFERENCES

[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[2] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[3] N. Audebert, A. Boulch, H. Randrianarivo, B. Le Saux, M. Ferecatu, S. Lefevre, and R. Marlet, "Deep learning for urban remote sensing," in *Urban Remote Sensing Event (JURSE), 2017 Joint*, pp. 1–4, IEEE, 2017.

[4] 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*, pp. 3431–3440, 2015.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[7] I. S. for Photogrammetry and R. S. (ISPRS), *2D Semantic Labeling Contest - Potsdam*. 2016.

[8] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–9, 2016.

[9] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Asian Conference on Computer Vision*, pp. 180–196, Springer, 2016.

[10] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, USA*, 2017.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *arXiv preprint arXiv:1612.01105*, 2016.

[13] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," *arXiv preprint arXiv:1703.02719*, 2017.

[14] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," *arXiv preprint arXiv:1702.08502*, 2017.

[15] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[17] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

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

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.