Article

Semantic Segmentation on Remotely-Sensed Images Using Enhanced Global Convolutional Network with Channel Attention and Domain Specific Transfer Learning

Teerapong Panboonyuen [1]*, Kulsawasd Jitkajornwanich [2], Siam Lawawirojwong [3], Panu Srestasathiern [3], and Peerapon Vateekul [1]*

[1] Chulalongkorn University Big Data Analytics and IoT Center (CUBIC), Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Phayathai Rd, Pathumwan, Bangkok 10330, Thailand; teerapong.panboonyuen@gmail.com, peerapon.v@chula.ac.th

[2] Data Science and Computational Intelligence (DSCI) Laboratory, Department of Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Rd, Ladkrabang, Bangkok 10520, Thailand; kulsawasd.ji@kmitl.ac.th

[3] Geo-Informatics and Space Technology Development Agency (Public Organization), 120, The Government Complex, Chaeng Wattana Rd, Lak Si, Bangkok 10210, Thailand; siam@gistda.or.th, panu@gistda.or.th

* Correspondence: teerapong.panboonyuen@gmail.com; peerapon.v@chula.ac.th;

† Current address: Affiliation 3

‡ These authors contributed equally to this work.

Version December 7, 2018 submitted to Preprints

Abstract: In remote sensing domain, it is crucial to automatically annotate semantics, e.g., river, building, forest, etc, on the raster images. Deep Convolutional Encoder Decoder (DCED) network is the state-of-the-art semantic segmentation for remotely-sensed images. However, the accuracy is still limited, since the network is not designed for remotely sensed images and the training data in this domain is deficient. In this paper, we aim to propose a novel CNN network for semantic segmentation particularly for remote sensing corpora with three main contributions. First, we propose to apply a recent CNN network call "Global Convolutional Network (GCN)", since it can capture different resolutions by extracting multi-scale features from different stages of the network. Also, we further enhance the network by improving its backbone using larger numbers of layers, which is suitable for medium resolution remotely sensed images. Second, "Channel Attention" is presented into our network in order to select most discriminative filters (features). Third, "Domain Specific Transfer Learning" is introduced to alleviate the scarcity issue by utilizing other remotely sensed corpora with different resolutions as pre-trained data. The experiment was then conducted on two given data sets: (*i*) medium resolution data collected from Landsat-8 satellite and (*ii*) very high resolution data called "ISPRS Vaihingen Challenge Data Set". The results show that our networks outperformed DCED in terms of $F1$ for 17.48% and 2.49% on medium and very high resolution corpora, respectively.

Keywords: Deep Convolutional Neural Networks; Multi-Class Segmentation; Global Convolution Network; Channel Attention; Transfer Learning; ISPRS Vaihingen, Landsat-8

1. Introduction

Semantic segmentation of earthly objects such as agriculture fields, forests, roads, urban and water areas, from remotely-sensed images has been manipulated in many applications in various domains, e.g., urban planning, map updates, route optimization, and navigation [1–5], allowing us to better understand the domain's images and create important real-world applications.

Deep convolutional neural network (CNN) is a well-known technique for automatic feature learning. It can mechanically learn features in different levels and abstractions from raw images by multiple hierarchical stacking convolution and pooling layers [4–14]. To accomplish such a challenging

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

task, features at different levels are required. Specifically, abstract high-level features are more suitable for the recognition of confusing manmade objects, while labeling of fine-structured objects could benefit from detailed low-level features [1]. Therefore, different numbers of layers will effect the performance of deep learning model.

In the past few years, the modern CNNs have been extensively proposed including Global Convolutional Network (GCN) [15] in which the large kernel and effective receptive field play an important role in performing classification and localization tasks simultaneously. The GCN is proposed to address the classification and localization issues for semantic segmentation and to suggest a residual-based boundary refinement for further refining object boundaries. However, this type of architecture ignores the global context such as weights of the features in each stage. Furthermore, most methods of this type are just summed up the features of adjacent stages without considering their diverse representations. This leads to some inconsistent results that suffer from accuracy performance. The primary challenge of this remote sensing task is a lack of training data. This, in fact, has become a motivation of this work.

In this paper, we present a novel Global Convolutional Network for segmenting multi-objects from aerial and satellite images. To this end, it is focused on three aspects: (i) varying backbones using ResNet50, ResNet101, and ResNet152; (ii) applying "Channel Attention Block" [16,17] to assign weights for the feature maps in each stage of backbone architecture, and (iii) "Domain Specific Transfer Learning" [18–20] is employed to relieve the scarcity issue. The experiments were conducted using satellite imagery (from the Landsat-8 satellite) which is provided by a government organization in Thailand and well-known aerial imagery, ISPRS Vaihingen Challenge corpus [34], which is publicly available. The results showed that our method outperforms the baseline including Deep Convolutional Encoder-Decoder (DCED) in terms of $F1$ and mean of class-wise Intersection over Union (*Mean IoU*).

The remainder of this paper is arranged as follows. Related work is discussed in Section 2. Section 3 describes our proposed methodology. Experimental data sets and evaluations are described in Section 4. Experimental results and discussions are presented in Section 5. Finally, we conclude our work and discuss future work in Section 6.

2. Related Work

Deep learning has been successfully applied for remotely-sensed data analysis, notably land cover mapping on urban areas [1–3] and has increasingly become a promising tool for accelerating image recognition process with high accuracy results [4–14,21–29], and is a fast-growing field, and new architectures appear every few days. This related work is divided into three subsections: we first discuss deep learning concepts for semantic segmentation, followed by a set of multi-objects segmentation techniques using modern deep learning architecture, and finally; modern technique of deep learning are discussed.

2.1. Deep learning concepts for semantic segmentation

Semantic segmentation algorithms are often formulated to solve structured pixel-wise labeling problems based on the deep convolutional neural network (CNN). Noh et al. [13] proposed a novel semantic segmentation technique utilizing a deconvolutional neural network (DeCNN) and the top layer from DCNN adopted from VGG16 [4,8]. DeCNN structure is composed of upsampling layers and deconvolution layers, describing pixel-wise class labels and predicting segmentation masks, respectively. Their proposed deep learning methods yield high performance in PASCAL VOC 2012 corpus, with the 72.5% accuracy in the best case scenario (the highest accuracy—as of the time of writing this paper—compared to other methods that were trained without requiring additional or external data). Long et al. [12] proposed an adapted contemporary classification networks incorporating Alex, VGG and GoogLe networks into fully CNN. In this method, some of the pooling layers were skipped: layer 3 (FCN-8s), layer 4 (FCN-16s), and layer 5 (FCN-32s). The skip architecture reduces the potential over-fitting problem and has showed improvements in performance, ranging from 20%

 doi:10.20944/preprints201812.0090.v1

to 62.2% in the experiments tested on PASCAL VOC 2012 data. Ronneberger et al. [14] proposed U-Net, a DCNN for biomedical image segmentation. The architecture consists of a contracting path and a symmetric expanding path that capture context and consequently, enable precise localization. The proposed network claimed to be capable to learn despite the limited number of training images, and performed better than the prior best method (a sliding-window DCNN) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. Vijay Badrinarayanan [30–32] proposed Deep Convolutional Encoder-Decoder network (DCED), namely "SegNet", consists of two main networks encoder and decoder, and some outer layers. The two outer layers of the decoder network are responsible for feature extraction task, the results of which are transmitted to the next layer adjacent to the last layer of the decoder network. This layer is responsible for pixel-wise classification (determining which pixel belongs to which class). There is no fully connected layer in between feature extraction layers. In the upsampling layer of decoder, pool indices from encoder are distributed to the decoder where kernel will be trained in each epoch (training round) at convolution layer. In the last layer (classification), softmax is used as a classifier for pixel-wise classification. DCED is one of the deep learning model that exceeds the state-of-the-art on many remote sensing corpus.

In this work, DCED method is selected as one of our baseline since it is the most popular architecture used in various networks for semantic segmentation.

2.2. Modern deep learning architecture for semantic segmentation

Recently, lots of approaches based on DCED have achieved high performance on different benchmarks [16,30–32]. However, most of them are still suffer from accuracy performance issues. Therefore, many works of modern deep learning architectures were proposed such as instance-aware semantic segmentation [33] which is slightly different from "semantic segmentation". Instead of labeling all pixels, it focuses on the target objects and labels only pixels of those objects. FCIS [27] is based techniques based on fully convolutional networks (FCN). Mask R-CNN [9] is also built around FCN and incorporates with a proposed joint formulation. Peng [15] presents concept of large kernel matters to improve semantic segmentation by global convolutional network (GCN). They proposes a GCN to address both the classification and localization issues for the semantic segmentation. Uses large separable kernels to expand the receptive field, also added a boundary refinement block to further improve localization performance near boundaries. From the Cityscapes challenge, GCN outperforms all the previous publications (all modern deep learning baselines) and reaches the new state-of-art. Therefore, GCN is selected to be the one of our proposed method and selected to be the main model on our work.

2.3. Modern technique of deep learning

Modern technique of deep learning is an important factor for an accuracy of CNN. While the most popular modern ideas tick for semantic segmentation tasks such as Global Context, Attention Module, Semantic Boundary Detection has been used for boosting accuracy.

Global Context [16] is some modern methods have proven the effectiveness of global average pooling in the semantic segmentation task. For example, PSPNet [29] and Deeplab v3 [5] respectively extend it to the Spatial Pyramid Pooling [29] and Atrous Spatial Pyramid Pooling [5], resulting in great performance in different benchmarks. However, to take advantage of the pyramid pooling module sufficiently, these two methods adopt the base feature network to 8 times downsample with atrous convolution [5] which is time-consuming and memory intensive.

Attention Module [16]: Attention is helpful to focus on what we want. Recently, the attention module becomes increasingly a powerful tool for deep neural networks [16,17]. The method in[16,17] pays attention to different scale information. In this work, we utilize channel attention block to select the features similar to Learning a Discriminative Feature Network [16].

Refinement residual block [16]: The feature maps of each stage in feature network all go through the Refinement Residual Block. For our work, we use Boundary Refinement Block (BR) to be concept

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

of "Refinement residual block" from [15]. The first component of the block is a 1×1 convolution layer. We use it to unify the number of channels to 21. Meanwhile, it can combine the information across all channels. Then the following is a basic residual block, which can refine the feature map. Furthermore, this block can strengthen the recognition ability of each stage, inspired from the architecture of ResNet [7].

3. Proposed Method

In this section, the details of our proposed network are explained (shown in Figure 2). The network is based on GCN with three aspects of improvements: (*i*) modification of backbone architecture (shown in P1 in Figure 2), (*ii*) applying the "Channel Attention Block" (shown in P2 in Figure 2), and (*iii*) using concept of domain specific "Transfer Learning" (shown in P3 in Figure 2).

3.1. Data Preprocessing

In this paper, there are two benchmark corpus including (*i*) ISPRS Vaihingen Challenge corpus and (*ii*) Landsat-8 data set. They are very high and medium resolution images, consecutively. More details of the data sets will be explained in Section 4.1 and Section 4.2. Before a discussion about the model, it is worth to explain our data preprocessing procedure, since it is required when working with neural network and deep learning models. Thus, the mean subtraction is executed.

In addition, data augmentation is often required on more complex object recognition tasks. Therefore, a random horizontal flip is generated to increase the training data. For the ISPRS corpus, all images are standardized and cropped into 512×512 pixels with a resolution of 9 cm^2/pixel. For the Landsat-8 corpus, each image is also flipped horizontally and scaled to 512×512 with a resolution of 30 m^2/pixel from original images ($16,800 \times 15,800$ pixels).

3.2. Global Convolutional Network (GCN) with variations of backbones

GCN [15] as shown in Figure 1 is a modern architecture that surpasses the drawbacks of traditional semantic segmentation network, such as, Deep Convolutional Encoder Decoder Networks (DCED). A traditional network usually cascades convolutional layers in order to generate sophisticated features; they can be considered as local features that is specialized only for a specific task. However, it is not necessary to employ only specialized features, but the general features are also important. Thus, GCN overcomes this issue by introducing a multi-level architecture that each level aims to capture different resolution of features, so both local and global features are considered into the model.

As in Figure 1, there are two main blocks in GCN: localization block and classification block. First from the localization view in the left block, the structure is a stack of classical fully-convolutional layer called "level". Each level aims to construct features with different resolutions. Second from the classification view, there are two modules: GCN and Boundary Refinement (BR). For the GCN module, the kernel size of the convolutional structure should be as large as possible, which is motivated by the densely-connected structure of classification models. Specially, if the kernel size increases to the spatial size of feature map (named global convolution), the network will share the same benefit with pure classification models. The BR module is added to further improve localization performance near boundaries.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

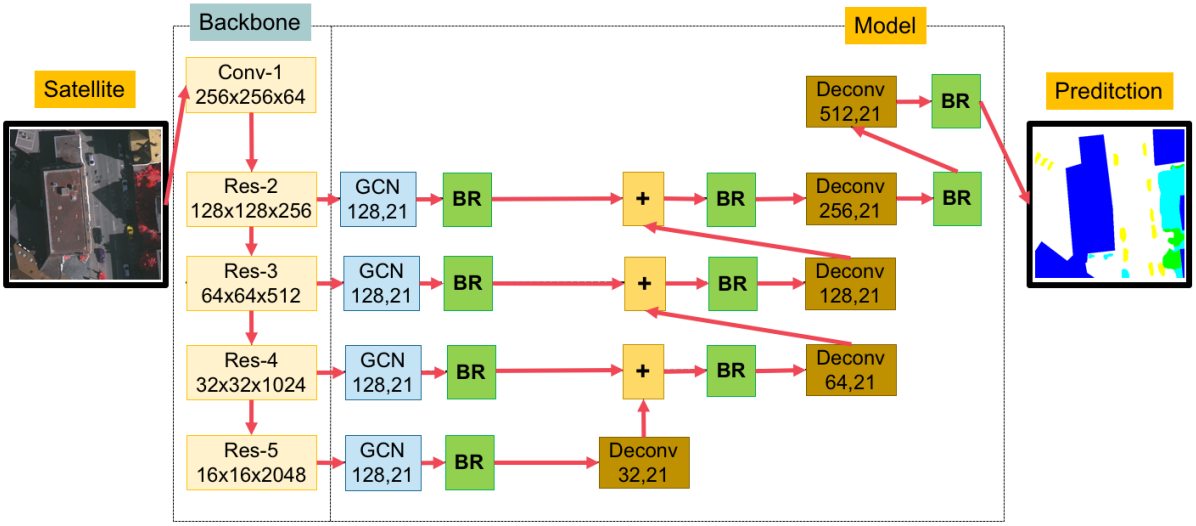


Figure 1. An overview of original Global Convolutional Network [15].

Although the GCN architecture has shown promising prediction performance, it can still be possible to further improved by varying backbones using ResNet [7] with different numbers of layers as ResNet50, ResNet101, and ResNet152 as shown in Figure 3. Also, GCN is suggested to work on large kernel size. In this paper, we set the large kernel size as 9 (this previous work [15]).

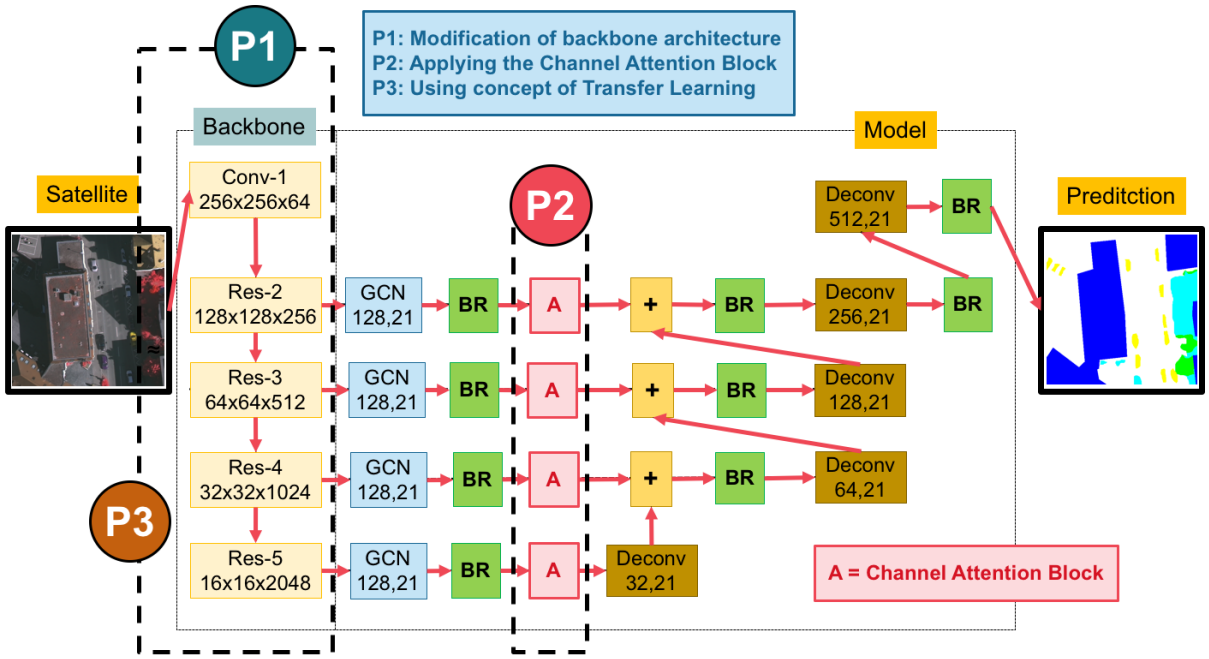


Figure 2. An overview of our proposed network.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

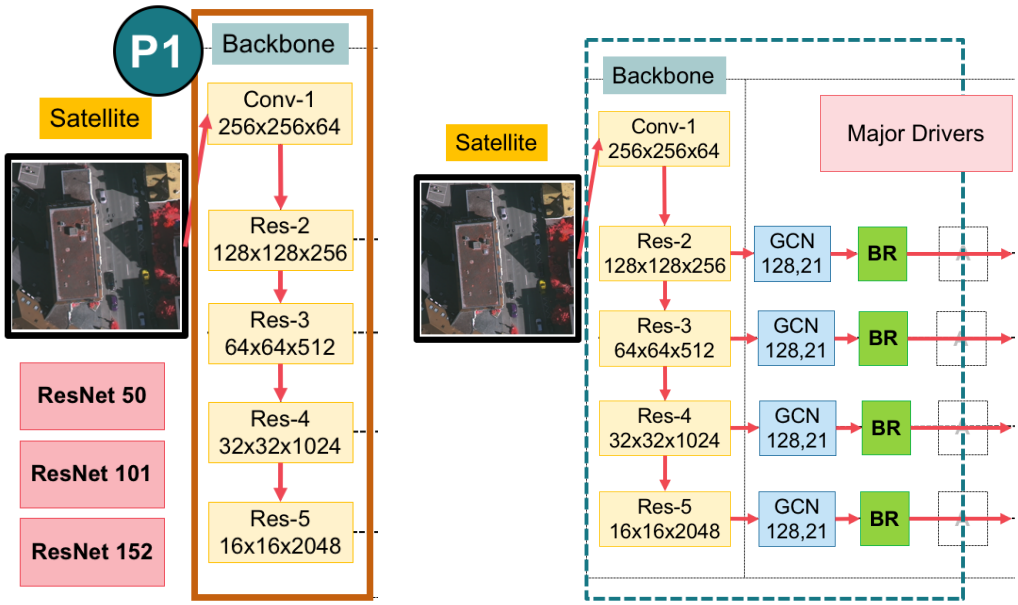


Figure 3. An overview of the whole backbone pipeline in $(left)$ the main backbone with varying by ResNet50, ResNet101, and ResNet152; and $(right)$ the major drivers of our main classification network (composed of of Global Convolutional Network (GCN) and Boundary Refinement (BR) block [15]).

3.3. Channel Attention Block

Attention Mechanisms [16,17] in Neural Networks are very loosely based on the visual attention mechanism found in humans. Human visual attention is well-studied and while there exist different models, all of them essentially come down to be able to focus on a certain region of an image with "very high resolution", while perceiving the surrounding image in "medium resolution", and then adjusting the focal point over time.

To apply this atttentional layer to our network, the channel attention block is shown in Block "A" in Figure 2 and its detailed architecture is shown in Figure 4. It is designed to change the weights of the remote sensing features on each stage (level), so that the weights are assigned more values on important features adaptively.

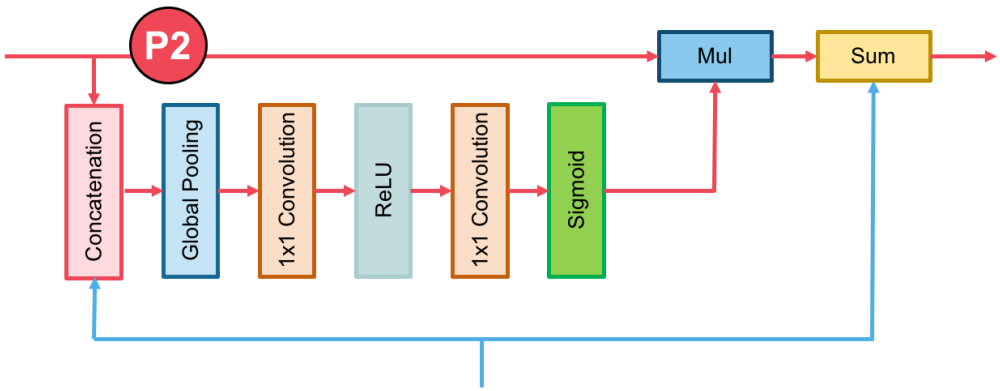


Figure 4. Components of Channel Attention Block. The blue and red lines represent the upsample and downsample operators, respectively. The red line cannot change the size of feature maps, just a path of information passing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

3.4. Domain Specific Transfer Learning

Although the deep learning approach often performs promising prediction performance, it requires a large amount of training data. Since it is difficult to obtain annotated satellite images, the perform in prior works should be limited.

Fortunately, there is a recent concept called "Domain Specific Transfer Learning" [18–20] that allows to reuse the weights obtaining from other domains' inputs. It is currently very popular in the field of Deep Learning because it enables you to train Deep Neural Networks with comparatively insufficient data. This is very useful since most real-world problems typically do not have millions of labeled data points to train such complex models.

From the inadequacy issue, we propose an effective Transfer Deep Neural Network to perform knowledge transfer between Very High Resolution (VHR) corpus and Medium Resolution (MR) corpus. It is shown in Figure 5.

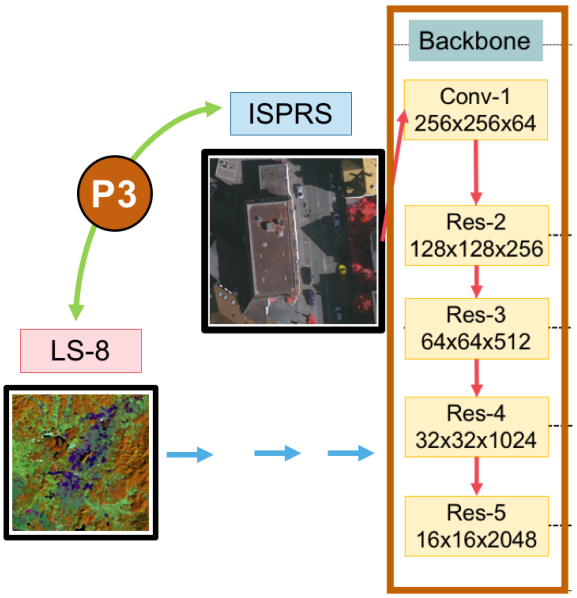


Figure 5. Domain Specific Transfer Learning strategy reuses pre-trained weights of models between two data sets – very high (ISPRS) and medium (Landsat-8; LS-8) resolution images.

4. Experimental Data Sets and Evaluation

In our experiments, two types of data sets are used: (i) medium resolution imagery (satellite images; Landsat-8 data set) made by the government organization in Thailand, name as GISTDA (Geo-Informatics and Space Technology Development Agency (Public Organization)) and (ii) very high resolution imagery (aerial images; ISPRS Vaihingen data set). All experiments are evaluated based on major metrics, such as *Average Accuracy*, *F1 Score* and *Mean IoU Score*.

4.1. Landsat-8 Data Set

In this type of data, the satellite images are from Nan, province in Thailand. The data set is obtained from Landsat-8 satellite consisting of 1,012 satellite images as shown some samples in Figure 6.

This corpus is comprised of a large, diverse set of medium resolution $(16,800 \times 15,800)$ pixels, where 1,012 of these images have high quality pixel-level labels of 5 classes: Agriculture, Forest, Miscellaneous, Urban, and Water Class. The 1,012 images are split into 800 training and 112 validation images with publicly available annotation, as well as 100 test images with annotations withheld and comparison to other methods are performed via a dedicated evaluation server. For quantitative evaluation, mean of class-wise Intersection over Union (*Mean IoU*) and *F*1 score are used.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

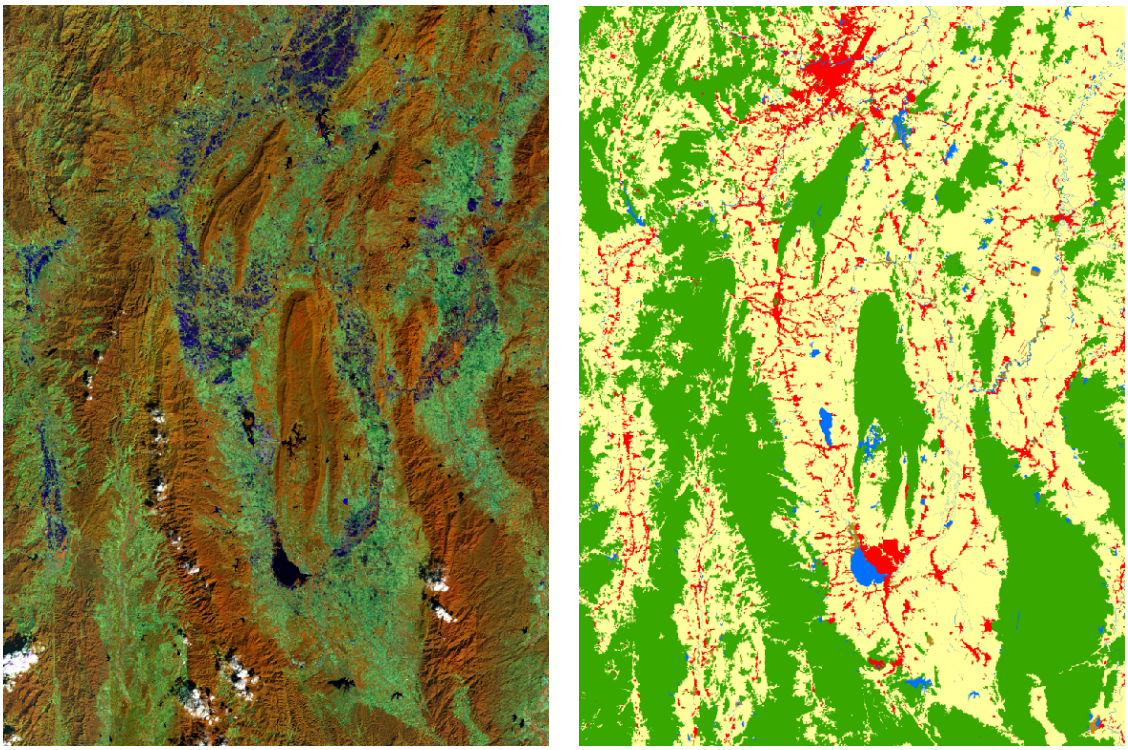


Figure 6. Sample satellite images from Nan, a province in Thailand (left) and corresponding ground truth (right). The label of medium resolution data set includes five categories: impervious surface (Agriculture, yellow), Forest (green), Miscellaneous (Misc, brown), Urban (red) and Water (blue).

4.2. ISPRS Vaihingen Data Set

One of the major challenges in remote sensing is the automated extraction of urban objects from data acquired by airborne sensors. Semantic Labeling Contest provides two state-of-the-art airborne image corpora, consisting of (i) Vaihingen corpus is a relatively small village with many detached buildings and small multi-story buildings, and (ii) Potsdam corpus shows a typical historic city with large building blocks, narrow streets and dense settlement structure. In our experiments, Vaihingen corpus was selected and used.

ISPRS 2D Semantic labeling challenge in Vaihingen [34] (Figure 7 and Figure 8) is used to be our benchmark data set. It consists of three spectral bands (i.e., red, green and near-infrared bands), corresponding DSM (Digital Surface Model) and NDSM (Normalized Digital Surface Model) data. Overall, there are 33 images of about $2,500 \times 2,000$ pixels at a Ground sampling distance (GSD) of about 9 cm in image data. Among them, the ground truth of only 16 images are available, and those of the remaining 17 images are withheld by the challenge organizer for online test. For offline validation, we randomly split the 16 images with ground truth available into a training set of 10 images, and a validation set of 6 images. For this work, DSM and NDSM data in all the experiments on this data set are not used. Following other methods, 4 tiles (image numbers 5, 7, 23, 30) are removed from the training set as a validation set. Experimental results are reported on the validation set if not specified.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

Figure 7. Overview of the ISPRS 2D Vaihingen Labeling corpus. There are 33 tiles. Numbers in the figure refer to the individual tile flag.

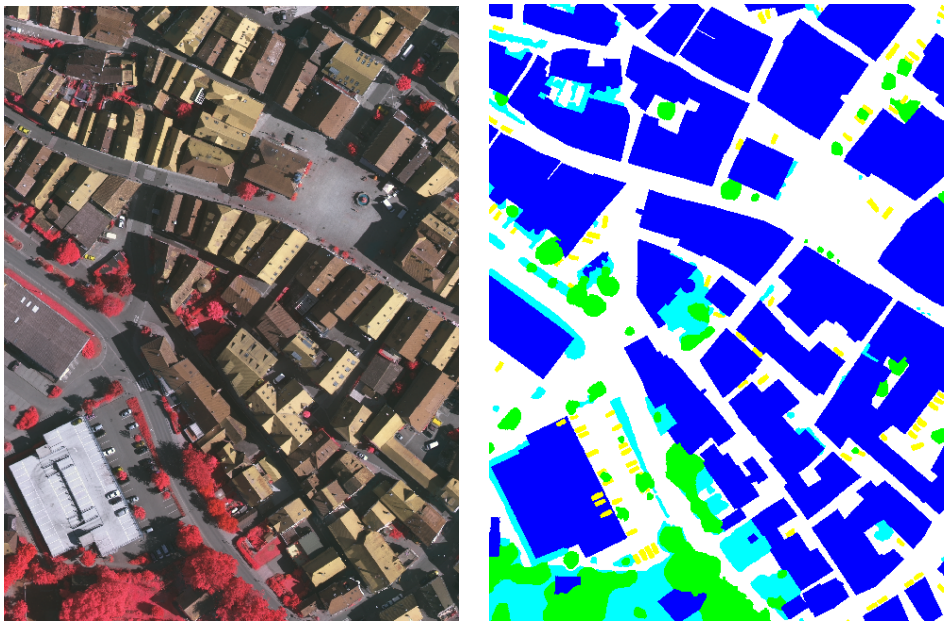


Figure 8. The sample input tile from Figure7 (left) and corresponding ground truth (right). The label of Vaihingen challenge includes six categories: impervious surface (imp surf, white), building (blue), low vegetation (low veg, cyan), tree (green), car (yellow) and clutter/background (red).

4.3. Evaluation

The multi-class classification task can be considered as multi-segmentation, where class pixels are positives and the remaining non-spotlight pixels are negatives. Let TP denotes the number of true positives, TN denotes the number of true negatives, FP denotes the number of false positives, and FN denotes the number of false negatives.

Precision, *recall*, $F1$, and *Mean IoU* are shown in equations (1-5). Precision is the percentage of correctly classified main pixels among all predicted pixels by the classifier. Recall is the percentage of correctly classified main pixels among all actual main pixels. $F1$ is a combination of *precision* and *recall*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

To evaluate the performance of different comparing deep models, we will discuss the above two major metrics ($F1$) and mean of class-wise Intersection over Union ($Mean\ IoU$)) on each category, and the mean value of metrics to assess the average performance.

$$Accuracy = \frac{TP + TN}{TP + +FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precison + Recall} \tag{4}$$

$$Mean\ IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

5. Experimental Results and Discussions

The implementation is based on a deep learning framework, called "Tensorflow-Slim" [35], which is extended from Tensorflow. All experiments were conducted on servers with Intel® Xeon® Processor E5-2660 v3 (25M Cache, 2.60 GHz), 32 GB of memory (RAM), Nvidia GeForce GTX 1070 (8 GB), Nvidia GeForce GTX 1080 (8 GB) and Nvidia GeForce GTX 1080 Ti (11 GB). In stead of using the whole image (1,500×1,500 pixels) to train the network, we randomly crop all images to be 512×512 as inputs of each epoch.

For training, Adam optimizer [11] is chosen with an initial learning rate of 0.004 and weight decay of 0.00001. Batch normalization [10] is used before each convolutional layer in our implementation to ease the training and make it be able to concatenate feature maps from different layers. To avoid overfitting, common data augmentations are used as details in Section 3.1. For the measurements, we use the mean pixel intersection-over-union ($mean\ IoU$) and $F1$ score as the metric.

Inspired by [16,26,36], we use the "poly" learning rate policy where the learning rate is multiplied by Eq. 6 with power 0.9 and initial learning rate as $4e^{-3}$. The learning rate is scheduled by multiplying the initial as seen in Eq. 6.

$$learning\ rate = (1 - \frac{epoch}{MaxEpoch})^{0.9} \tag{6}$$

All models are trained for 50 epochs with mini-batch size of 4, and each batch contains the cropped images that are randomly selected from training patches. These patches are resized to 521 × 521 pixels. . The statistics of batch normalization is updated on the whole mini-batch.

This section illustrates details of our experiments. The proposed deep learning network is based on GCN with three improvements: (*i*) varying backbones using ResNet, (*ii*) Channel Attention and Global Average Pooling, and (*iii*) Domain Specific Transfer Learning. From all proposed strategies, there are six acronyms of strategies as shown in Table 1.

Table 1. Abbreviations on our proposed deep learning methods

Abbreviation	Description
A	Channel Attention Block
GCN	Global Convolutional Network
GCN50	Global Convolutional Network with ResNet50
GCN101	Global Convolutional Network with ResNet101
GCN152	Global Convolutional Network with ResNet52
TL	Domain Specific Transfer Learning

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

For the experimental setup, there are three experiments on two remotely-sensed data sets: Landsat-8 data set and ISPRS Vaihingen challenge data set (details in Section 4.1 and Section 4.2). The experiments aim to illustrate that each proposed strategy can really improve the performance. First, "GCN152" method is compared to "GCN50" method and "GCN101" method for the varying backbones using ResNet with different numbers of layers on GCN networks strategy. Second, "GCN152-A" method is compared to "GCN152" method for the "Channel Attention" strategy. Third, the full proposed technique "GCN152-TL-A" method is compared to existing methods for the concept of domain specific transfer learning.

5.1. Results on Landsat-8 Corpus with Discussion

In this subsection, the experiment was conducted on the Landsat-8 corpus. The result is shown in Table 2 and Table 3 by comparing between baseline and variations of the proposed techniques. It shows that our network with all strategies "GCN152-TL-A" outperforms other methods. More details will be discussed to show that each of the proposed techniques can really improve an accuracy. Only in this experiment, there are state of the art baseline, including Deep Convolutional Encoder-Decoder (DCED) [30–32].

5.1.1. Effect of enhanced GCN on Landsat-8 corpus

Our first strategy aims to increase an $F1$ and *Mean IoU* score of the network by varying backbones using ResNet 50, ResNet 101, and ResNet 152 rather than the traditional one, DCED method. From Table 2 and Table 3, $F1$ of GCN152 (0.7563) outperforms that of GCN50(0.6847) and GCN101 (0.7290), and baseline method; DCED(0.6495); this yields higher $F1$ at 2.74%, 3.52%, and 4.43% respectively. *Mean IoU* of GCN152 (0.6364) outperforms that of GCN50 (0.5734), GCN101 (0.6154), and baseline method; DCED (0.5384); this yields higher *Mean IoU* at 2.10%, 3.50%, and 4.20% consecutively. The main reason is due to higher precision, but slightly lower recall. This can imply that enhanced GCN is more significantly efficient than DCED method (baseline) for this medium resolution corpus and ResNet with a large number of layers is more robust than the small number of layers.

When comparing the results between the original GCN method and the enhanced GCN methods on Landsat-8 corpus (Table 2), it clearly shows that GCN with the larger layer of the backbone can improve the network performance in term of $F1$ and *mean IoU*

5.1.2. Effect of using "Channel Attention" on Landsat-8 corpus

Our second mechanism focuses on applying "Channel Attention Block" (details in Section 3.4) to change the weights of the features on each stage to enhance the consistency. From Table 2 and 3, the $F1$ of GCN152-A (0.7897) is greater than that of GCN152 (0.7563); this yields higher $F1$ score at 3.34%. and the *Mean IoU* of GCN152-A (0.6726) is superior to that of GCN152 (0.6364); this yields higher *Mean IoU* score at 3.62%. The result (Figure 9e and Figure 12e) shows that can make the network to obtain discriminative features stage-wise to make the prediction intra-class consistent. This is based on the consideration that we re-weighted all feature maps of each layer.

5.1.3. Effect of using "Domain Specific Transfer Learning" on Landsat-8 corpus

Our last strategy aims to use approach of domain specific "Transfer Learning" (details in Section 3.3) by reusing the pre-trained weight from "GCN152-A" model on ISPRS Vaihingen corpus. From Table 2 and Table 3, $F1$ of "GCN152-TL-A" method is the winner; it clearly outperforms not only the baseline, but also all previous generations. Its $F1$ is higher than DCED (baseline) at 17.80%. Its *Mean IoU* is higher than DCED at 17.94%. Also, the result illustrates that concept of domain specific "Transfer Learning" can enhance both precision (0.8293) and recall (0.8476).

Figure 9 and Figure 12 shows twelve sample results from the proposed method. By applying all strategies, the images in the last column (Figure 9f and Figure 12f) are similar to the ground truths

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

(Figure 9b and Figure 12b). Furthermore, *F*1-results and *Mean IoU* scores are improved for each strategy we added to the network as shown in Figure 9(c-f) and Figure 12(c-f).

Table 2. Results on the testing data of Landsat-8 corpus between baseline and five variations of our proposed techniques in terms of *precision*, *recall*, *F*1 and *Mean IoU*.

	Pretrained	Backbone	Model	*Precision*	*Recall*	*F*1	*Mean IoU*
Baseline	-	-	DCED [30–32]	0.6137	0.7209	0.6495	0.5384
Proposed Method	-	Res50	GCN [15]	0.6678	0.7333	0.6847	0.5734
	-	**Res101**	GCN	0.6899	0.8031	0.7290	0.6154
	-	**Res152**	GCN	0.7115	0.8131	0.7563	0.6364
	-	**Res152**	**GCN-A**	0.7997	0.7937	0.7897	0.6726
	TL	**Res152**	**GCN-A**	**0.8293**	**0.8476**	**0.8275**	**0.7178**

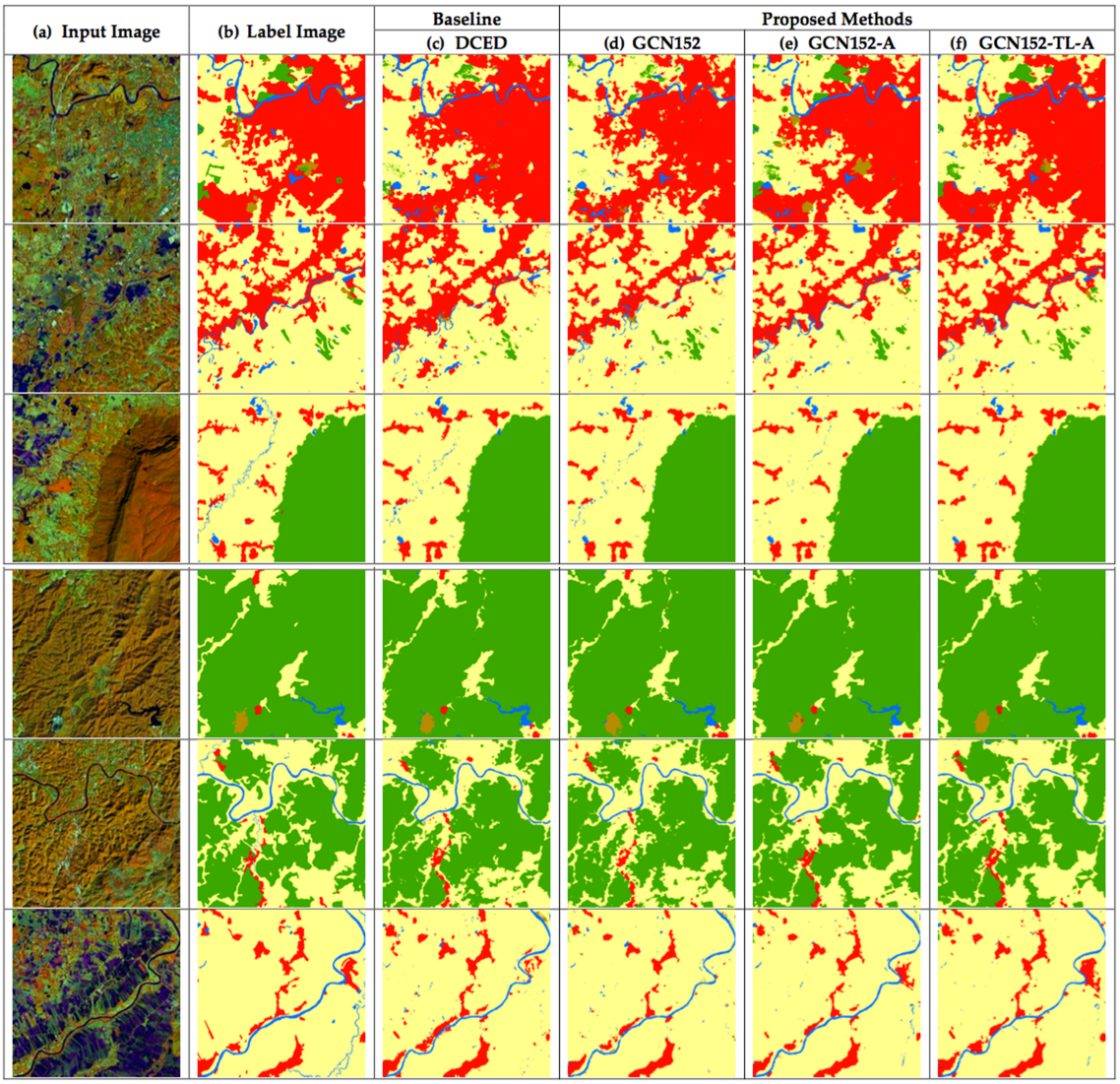


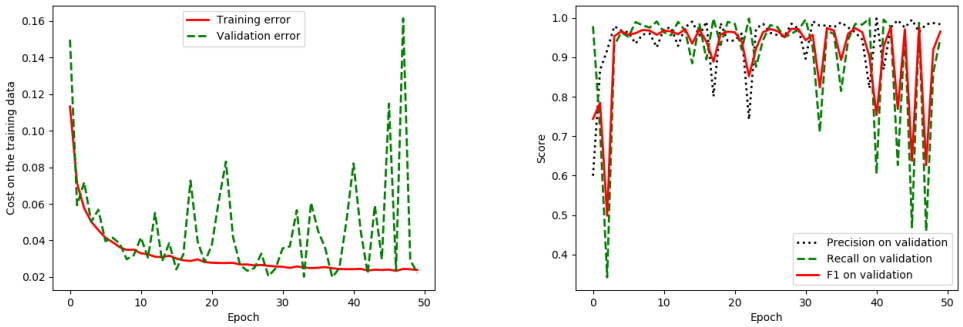
Figure 9. Six testing sample input and output satellite images on Landsat-8 in Nan provinces in Thailand, where rows refer different images. (a) Original input image; (b) Target map (ground truth); (c) Output of Encoder Decoder (Baseline); (d) Output of GCN152; (e) Output of GCN152-A; and (f) Output of GCN152-TL-A.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

Table 3. Results on the testing data of Landsat-8 corpus between each class with our proposed techniques in terms of *AverageAccuracy*

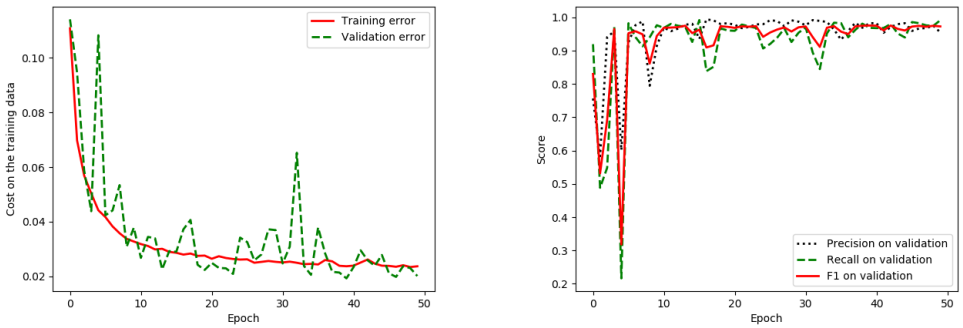
	Model	Agriculture	Forest	Misc	Urban	Water
Baseline	DCED [30–32]	0.9616	0.7472	0.0976	0.7878	0.4742
Proposed Method	GCN50 [15]	0.9407	0.8258	0.1470	**0.8828**	0.5426
	GCN101	0.9677	0.8806	0.2561	0.7971	0.5480
	GCN152	0.9780	0.8444	0.4256	0.7158	0.5937
	GCN152-A	0.9502	**0.9118**	0.6689	0.8675	0.6001
	GCN152-TL-A	**0.9781**	0.8472	**0.8732**	0.7988	**0.6493**

To achieve highest accuracy, the network must be configured and trained many epochs until all parameters in the network are converged. Figure 11(a) illustrates that the proposed network has been properly set and trained until it is really converged and ran more smoothly than baseline in Figure 10(a). Furthermore, Figure 10(b) and Figure 11(b) show that the higher number of epochs tend to show better $F1$ score. Thus, the number of chosen epochs based on the validation data is 49 (The best model for this data set).



(a) **(b)**

Figure 10. Iteration plot on Landsat-8 corpus of the baseline technique, DECE [30–32]; x refers to epochs and y refers to different measures **(a)** Plot of model loss (cross entropy) on training and validation data sets and **(b)** Performance plot on the validation data set.



(a) **(b)**

Figure 11. Iteration plot on Landsat-8 corpus of the proposed technique, "GCN152-TL-A"; x refers to epochs and y refers to different measures **(a)** Plot of model loss (cross entropy) on training and validation data sets, **(b)** Performance plot on the validation data set.

Twelve sample testing results (shown as Figure 9 and Figure 12) from the proposed method on Nan provinces (is one of the northern provinces (changwat) of Thailand and agriculture is the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

province's main industry). The results of the last column look closest to the ground truth in the second column.

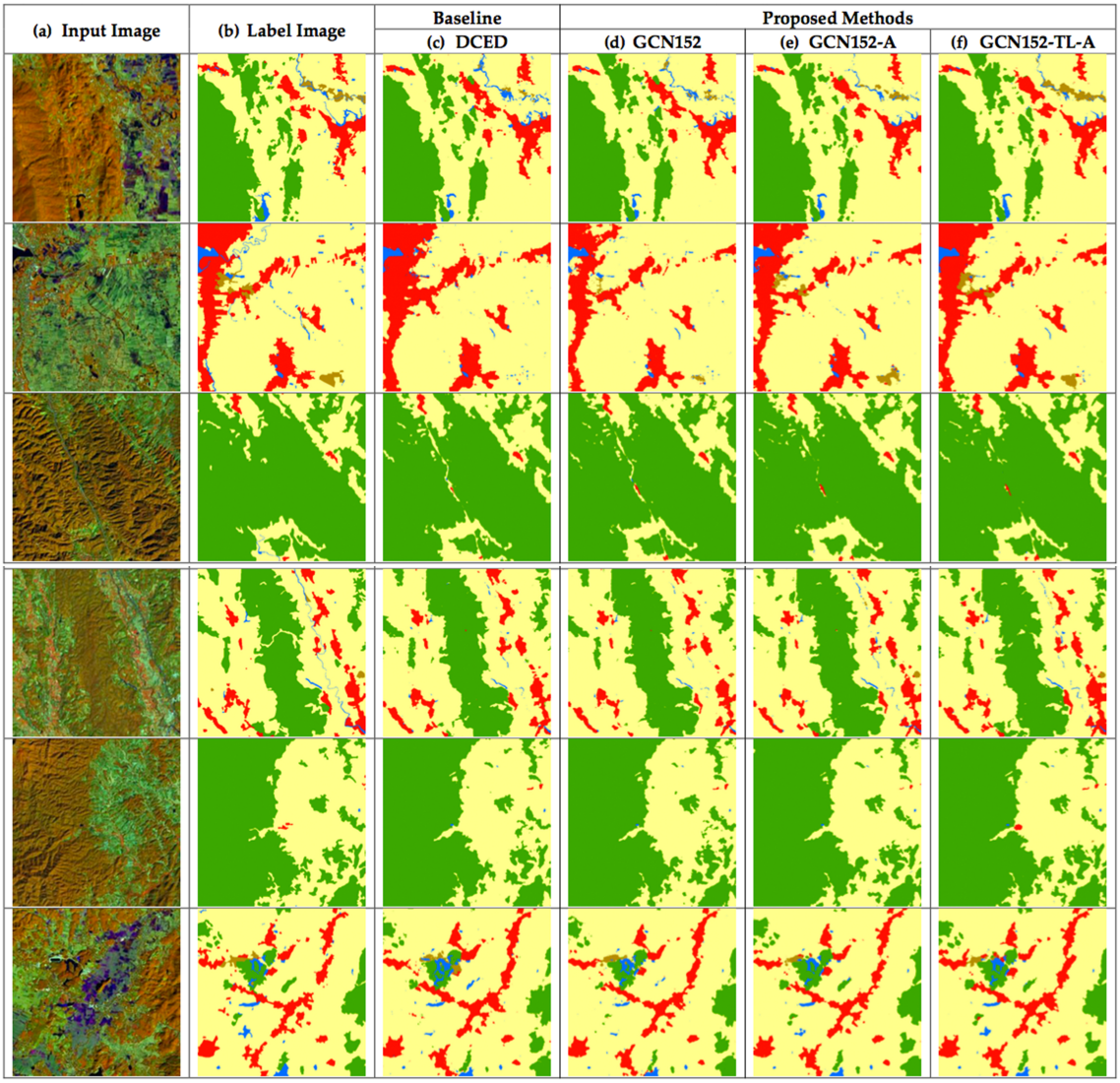


Figure 12. Six testing sample input and output satellite images on Landsat-8 in Nan provinces in Thailand, where rows refer different images. (a) Original input image; (b) Target map (ground truth); (c) Output of Encoder Decoder (Baseline); (d) Output of GCN152; (e) Output of GCN152-A; and (f) Output of GCN152-TL-A.

As can be seen in Figure 9 and Figure 12, the performance of our best model outperforms other advanced models by a considerable margin on each category, especially for the Agriculture, Miscellaneous (Misc), and Water. Furthermore, the loss curves shown in Figure 11(a) exhibit that, our best model performs better on all the given categories.

5.2. Results on ISPRS Vaihingen Challenge Corpus with Discussion

In this subsection, the experiment was conducted on the ISPRS Vaihingen Challenge corpus. The result is shown in Table 4 and Table 5 by comparing between baseline and variations of the proposed techniques. It shows that our network with all strategies (GCN152-TL-A) outperforms other methods. More details will be discussed to show that each of the proposed techniques can really improve an accuracy. Only in this experiment, there are one baseline, including DCED network.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

5.2.1. Effect of enhanced GCN on ISPRS Vaihingen corpus

Our first strategy aims to increase an *F*1 and *Mean IoU* score of the network by varying backbones using ResNet 50, ResNet 101, and ResNet 152 rather than the traditional one, DCED method. From Table 4 and Table 5, *F*1 of GCN152 (0.7864) outperforms that of GCN50 (0.776), GCN101 (0.768), and baseline method; DCED (0.7693); this yields higher *F*1 at 0.02%, 0.68%, and 1.01% respectively. *Mean IoU* of GCN152 (0.8977) outperforms that of GCN50 (0.8776), GCN101 (0.8972), and baseline method; DCED (0.8651); this yields higher *Mean IoU* at 0.02%, 0.68%, and 1.01% consecutively. This can imply that enhanced GCN is also more accurate than DCED approach on very high resolution data set. ResNet with large number of layers is still more robust than small number of layers same as that performed on Landsat-8 corpus (Section 5.1.1).

When comparing the results between the original GCN method and the enhanced GCN methods on Landsat-8 corpus (Table 4), it clearly shows that GCN with the larger layer of the backbone can improve the network performance in term of *F*1 and *mean IoU*

5.2.2. Effect of using "Channel Attention" on ISPRS Vaihingen corpus

Our second mechanism focuses on utilizing "Channel Attention Block" to change the weights of the features on each stage to enhance the consistency. From Table 4 and 5, the *F*1 of GCN152-A (0.7902) is greater than that of GCN152 (0.7864); this yields higher *F*1 score at 0.38%. and the *Mean IoU* of GCN152-A (0.9057) is better than that of GCN152 (0.8977); this yields higher *Mean IoU* score at 0.80%. The results (Figure 13e and Figure 14e) show that can also make the network to obtain discriminative features stage-wise to make the prediction intra-class consistent on very high resolution images.

5.2.3. Effect of using "Domain Specific Transfer Learning" on ISPRS Vaihingen corpus

Our last strategy aims to performing approach of domain specific "Transfer Learning" (details in Section 3.3) by reusing the pre-trained weight from "GCN152-A" model on Landsat-8 corpus. From Table 4 and Table 5, *F*1 of "GCN152-TL-A" method is the winner; it clearly outperforms not only the baseline, but also all previous generations. Its *F*1 is higher than DCED (baseline) at 2.49% and 1.82% consecutively. Its *Mean IoU* is higher than DCED and GCN at 4.76% and 3.51% respectively. Also, the result illustrates that concept of domain specific "Transfer Learning" can enhance both precision (0.7888) and recall (0.8001).

Figure 13 and Figure 14 shows twelve sample results from the proposed method. By applying all strategies, the images in the last column (Figure 13f and Figure 14f) are similar to the ground truths (Figure 13b and Figure 14b). Furthermore, *F*1-results and *Mean IoU* scores are improved for each strategy we added to the network as shown in Figure 13(c-f) and Figure 14(c-f).

To further evaluate the effectiveness of the proposed "GCN152-TL-A" comparisons with baseline' method on the one challenging benchmark and one private benchmark are presented as follows as Table 2 and Table 3 for Landsat-8 data set on Nan province (Thailand) corpus and Table 4 and Table 5 for Vaihengen data set. All extensive experiments on the Landsat-8 and ISPRS dataset demonstrate that the proposed method achieve clear promising gains compared with the baseline approache.

Table 4. Results on the testing data of ISPRS 2D semantic labeling challenge corpus between baseline and five variations of our proposed techniques in terms of *precision*, *recall*, *F*1 and *Mean IoU*.

	Pretrained	**Backbone**	**Model**	*Precision*	*Recall*	*F*1	*Mean IoU*
Baseline	-	-	DCED [30–32]	0.7519	0.7925	0.7693	0.8651
Proposed Method	-	Res50	GCN [15]	0.7636	0.7917	0.776	0.8776
	-	**Res101**	GCN	0.7713	**0.8059**	0.7862	0.8972
	-	**Res152**	GCN	0.7736	0.8021	0.7864	0.8977
	-	**Res152**	**GCN-A**	0.7847	0.7961	0.7902	0.9057
	TL	**Res152**	**GCN-A**	**0.7888**	0.8001	**0.7942**	**0.9123**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

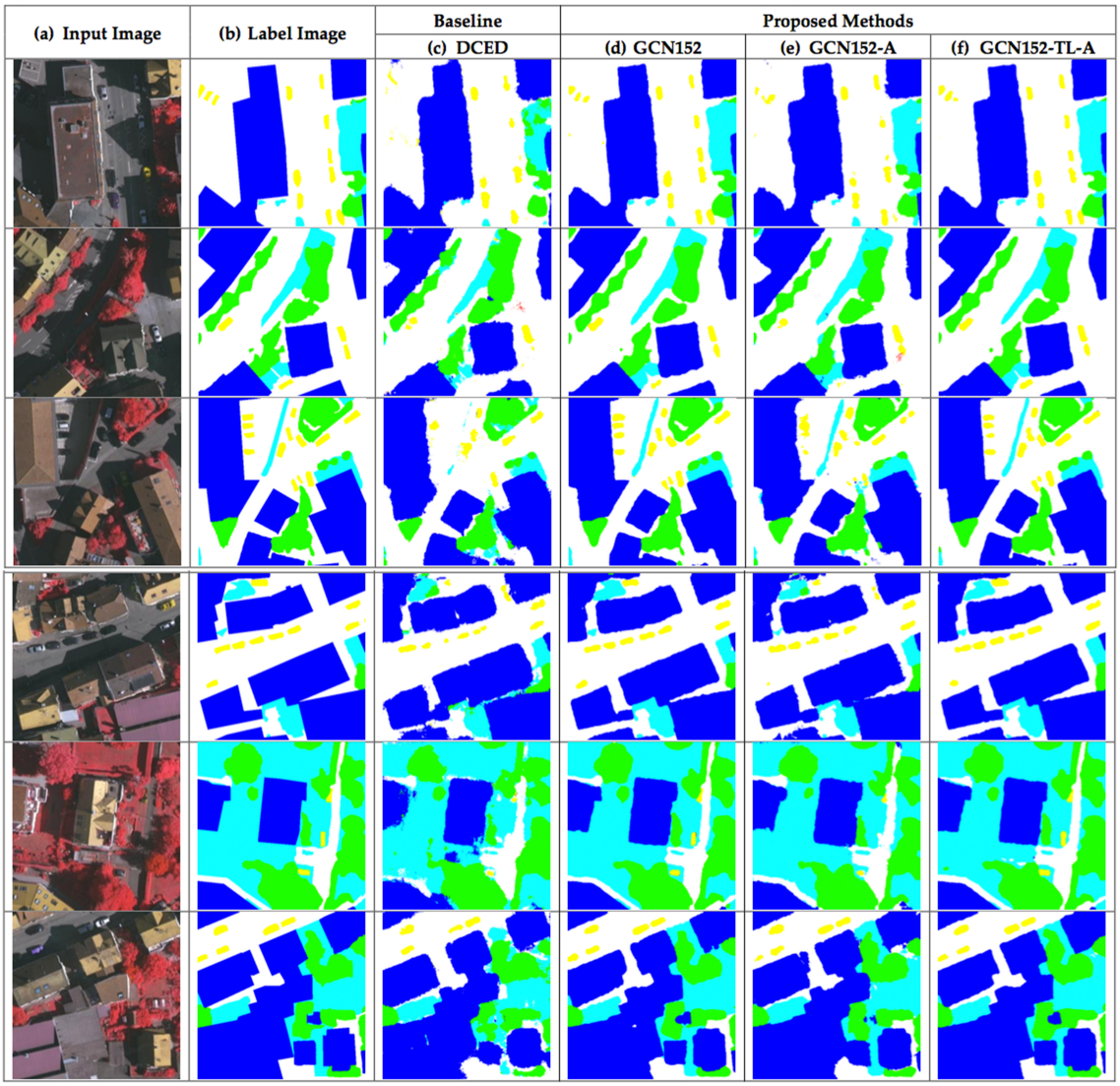


Figure 13. Six testing sample input and output aerial images on ISPRS Vaihingen challenge corpus, where rows refer different images. (a) Original input image; (b) Target map (ground truth); (c) Output of Encoder Decoder (Baseline); (d) Output of GCN152; (e) Output of GCN152-A; and (f) Output of GCN152-TL-A.

Table 5. Results on the testing data of ISPRS Vaihingen challenge corpus between each class with our proposed techniques in terms of *AverageAccuracy*

	Model	IS	Buildings	LV	Tree	Car
Baseline	DCED [30–32]	0.9590	0.9778	0.9108	0.9805	0.6832
Proposed Method	GCN50 [15]	0.9595	0.9628	0.9403	0.9896	0.7292
	GCN101	0.9652	0.9827	**0.9615**	0.9797	0.7387
	GCN152	0.9543	**0.9962**	0.9445	0.9754	0.7710
	GCN152-A	0.9614	0.9865	0.9554	0.9871	0.8181
	GCN152-TL-A	**0.9664**	0.9700	0.9499	**0.9901**	**0.8567**

Figure 13 and Figure 14 show twelve sample testing results from the proposed method on ISPRS Vaihingen corpus. The results of the last column are also similar to the ground truth in the second column same as performed on Landsat-8 corpus. Considering to each class (are shown in Table 3 and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2018 doi:10.20944/preprints201812.0090.v1

Table 5), almost every classes (three out of five) from our proposed methods are the winner in term *Average Accuracy*.

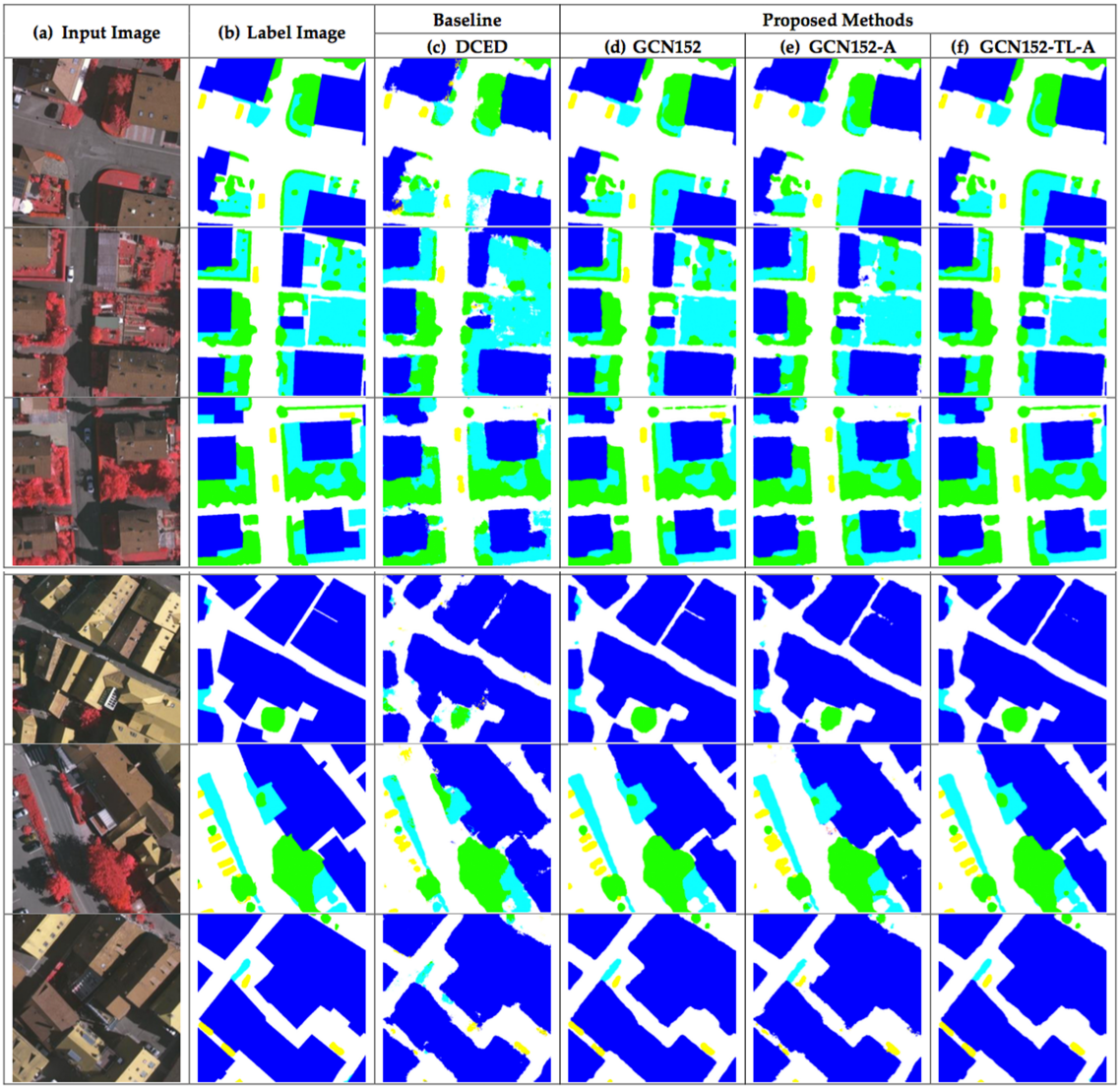


Figure 14. Six testing sample input and output aerial images on ISPRS Vaihingen challenge corpus, where rows refer different images. (a) Original input image; (b) Target map (ground truth); (c) Output of Encoder Decoder (Baseline); (d) Output of GCN152; (e) Output of GCN152-A; and (f) Output of GCN152-TL-A.

As can be seen in Figure 13 and Figure 14, the performance of our best model outperforms other advanced models by a considerable margin on each category, especially for the Impervious Surfaces (IS), Tree, and Car. To show the effectiveness of the proposed methods, we have performed comparisons against a number of state-of-the-art semantic segmentation methods, as listed in Table 4, Table 5 that performs on ISPRS corpus and Table 2, Table 3 that performs on Landsat-8 corpus. Encoder Decoder (DCED) [30–32] and GCN [15] are the versions with ResNet-50 as their backbone. In particular, we re-implement the DCED with Tensorflow-Slim [35], since the released code is built on Caffe [37]. We can see that our proposed methods significantly outperforms other methods on both *F1 score* and *mean IoU*.

In terms of the computational cost, our framework requires slightly additional training time compared to the baseline approach, DCED, by about 6.25% (6-7 hours), and GCN, by about 4.5% (4-5 hours). In our experiment, DCED's training procedure took approximately 16 hours per data set, and

 doi:10.20944/preprints201812.0090.v1

finished after 50 epochs with 1,152 second per epoch. Our framework is modify on GCN-based deep learning architecture. The "Channel Attention" model increases the time 20 minutes from "GCN152" method. There is no additional time required by reusing pre-trained weights.

6. Conclusions and Future Work

In this study, we propose a novel CNN framework to perform semantic labeling on remote-sensed images. Our proposed method achieves excellent performance by presenting three aspects. First, Global Convolutional Network (GCN) is employed and enhanced by adding larger numbers of layers to better capture the complex features. Second, "Channel Attention" is proposed to assign a proper weight for each extracted feature on different stages of the network. Finally, "Domain Specific Transfer Learning" is introduced to to allay the scarcity issue by training the initial weights using other remotely sensed corpora whose resolutions can be different. The experiments were conducted on two data sets: Landsat-8 (medium resolutions) and ISPRS Vaihingen Challenge (very hign resolution) data sets. The results show that our model that combines all proposed strategies outperforms baseline models in terms of $F1$ and $MeanIoU$. The final results show that our enhanced GCN outperforms the baseline (DCED)—17.48% for F1 on Landsat-8 corpus and 2.48% on ISPRS corpus.

In the future, more choices of semantic labeling, modern optimization techniques and/or other novel activation functions will be investigated and compared to obtain the best GCN-based framework for semantic segmentation in remotely-sensed images. Moreover, incorporating other data sources (e.g. digital surface model) might be needed to increase the accuracy of the Deep Learning for both the CNN and modern Deep Learning layer with very low confidence simultaneously. These aforementioned issues will be investigated in future research.

Acknowledgments: T. Panboonyuen thanks the scholarship from The 100^{th} Anniversary Chulalongkorn University Fund granted and The 90^{th} Anniversary Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund). We greatly acknowledge Geo-informatics and Space Technology Development Agency (GISTDA), Thailand, for providing satellite imagery used in this study.

Author Contributions: Teerapong Panboonyuen performed all the experiments and wrote the paper; and Peerapon Vateekul performed the results analysis and edited manuscript. Kulsawasd Jitkajornwanich, SiamLawawirojwong and Panu Srestasathiern reviewed results. Teerapong Panboonyuen revised manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BR	Boundary Refinement
CNN	Convolutional Neural Network
DCED	Deep Convolutional Encoder-Decoder
GCN	Global Convolutional Network
IS	Impervious Surfaces
Misc	Miscellaneous
MR	Medium Resolution
RGB	Red-Green-Blue
LS	Landsat
LV	Low Vegetation
TL	Transfer Learning
VHR	Very High Resolution

References

1. Liu, Y.; Fan, B.; Wang, L.; Bai, J.; Xiang, S.; Pan, C. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS Journal of Photogrammetry and Remote Sensing* **2018**, *145*, 78–95.

2. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sensing* **2017**, *9*, 446.
3. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine* **2017**, *5*, 8–36.
4. Panboonyuen, T.; Vateekul, P.; Jitkajornwanich, K.; Lawawirojwong, S. An Enhanced Deep Convolutional Encoder-Decoder Network for Road Segmentation on Aerial Imagery. *Recent Advances in Information and Communication Technology Series* **2017**, *566*.
5. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915* **2016**.
6. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* **2017**.
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
8. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Road Segmentation of Remotely-Sensed Images Using Deep Convolutional Neural Networks with Landscape Metrics and Conditional Random Fields. *Remote Sensing* **2017**, *9*, 680.
9. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017, pp. 2980–2988.
10. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* **2015**.
11. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
12. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
13. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
14. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp. 234–241.
15. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters—improve semantic segmentation by global convolutional network. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017, pp. 1743–1751.
16. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a Discriminative Feature Network for Semantic Segmentation. *arXiv preprint arXiv:1804.09337* **2018**.
17. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507* **2017**, *7*.
18. Xie, M.; Jean, N.; Burke, M.; Lobell, D.; Ermon, S. Transfer learning from deep features for remote sensing and poverty mapping. *arXiv preprint arXiv:1510.00098* **2015**.
19. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? Advances in neural information processing systems, 2014, pp. 3320–3328.
20. Liu, J.; Wang, Y.; Qiao, Y. Sparse Deep Transfer Learning for Convolutional Neural Network. AAAI, 2017, pp. 2245–2251.
21. Valada, A.; Vertens, J.; Dhall, A.; Burgard, W. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, 2017, pp. 4644–4651.
22. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611* **2018**.
23. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *arXiv preprint arXiv:1808.00897* **2018**.
24. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
25. Bilinski, P.; Prisacariu, V. Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6596–6605.

26. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. DenseASPP for Semantic Segmentation in Street Scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3684–3692.
27. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. *arXiv preprint arXiv:1611.07709* **2016**.
28. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. *arXiv preprint arXiv:1704.08545* **2017**.
29. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881–2890.
30. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293* **2015**.
31. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561* **2015**.
32. Kendall, A.; Badrinarayanan, V.; Cipolla, R. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680* **2015**.
33. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3150–3158.
34. International Society for Photogrammetry and Remote Sensing. 2D Semantic Labeling Challenge. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html. Accessed: 2018-09-09.
35. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; others. Tensorflow: a system for large-scale machine learning. OSDI, 2016, Vol. 16, pp. 265–283.
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, pp. 1097–1105.
37. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 675–678.