Article

Road Segmentation on Remotely-Sensed Images Using Deep Convolutional Neural Networks with Landscape Metrics and Conditional Random Fields

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

[1] Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Phayathai Rd, Pathumwan, Bangkok 10330, Thailand; peerapon.v@chula.ac.th

[2] Department of Computer ScienceFaculty 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 (S.L.); panu@gistda.or.th (P.S.)

* Correspondence: teerapong.pan@student.chula.ac.th;

Abstract: Object segmentation on remotely-sensed images: aerial (or very high resolution, VHS) images and satellite (or high resolution, HR) images, has been applied to many application domains, especially road extraction in which the segmented objects are served as a mandatory layer in geospatial databases. Several attempts in applying deep convolutional neural network (DCNN) to extract roads from remote sensing images have been made; however, the accuracy is still limited. In this paper, we present an enhanced DCNN framework specifically tailored for road extraction on remote sensing images by applying landscape metrics (LMs) and conditional random fields (CRFs). To improve DCNN, a modern activation function, called exponential linear unit (ELU), is employed in our network resulting in a higher number of and yet more accurate extracted roads. To further reduce falsely classified road objects, a solution based on an adoption of LMs is proposed. Finally, to sharpen the extracted roads, a CRF method is added to our framework. The experiments were conducted on Massachusetts road aerial imagery as well as THEOS satellite imagery data sets. The results showed that our proposed framework outperformed Segnet, the state-of-the-art object segmentation technique on any kinds of remote sensing imagery, in most of the cases in terms of *precision*, *recall*, and *F*1.

Keywords: deep convolutional neural networks; road segmentation; conditional random fields; landscape metrics; satellite images; aerial images; THEOS

1. Introduction

Extraction of terrestrial objects: buildings and roads, from remotely-sensed images has been employed in many applications on various areas, e.g., urban planning, map updates, route optimization, and navigation. For road extraction, most primary researches are based on unsupervised learning, such as graph cut and global optimization techniques [1]. Anyway, these unsupervised works have one common limitation, which is color-sensitive since they reply on color features. That is, the segmentation algorithms will not perform well if the road colors presented in the suburban remotely-sensed images contain more than one color (e.g., yellowish brown roads in the countryside regions and cement-grayed roads in the suburban regions). This, in fact, has become a motivation of this work to overcome the color-sensitive issue.

Deep learning, a large convolutional neural network whose performance can be scaled depending on size of training data, model complexity as well as processing power, has shown significant improvements in object segmentation from images as seen in many of the recent works [2–7], [8], [9–12] and [13]. Unlike unsupervised learning, more than one features—other than color—can be extracted: line, shape, and texture, among others. The traditional deep learning methods such as deep convolutional neural networks (DCNN) [14] and [3], deep deconvolutional neural networks (DeCNN) [5], recurrent neural network, namely reSeg [15], and fully convolutional networks [4]. However, are all suffering from the accuracy performance issues.

A deep convolutional encoder-decoder (DCED) architecture, one of the most efficient newly developed neural networks, has been proposed for object segmentation, is designed to be a core segmentation engine for pixel-wise semantic segmentation, and given good performance in the experiments tested on PASCAL VOC 2012 data—a well-known benchmark data set for image segmentation research [6], [8], and [16]. In this architecture, Rectified Linear Unit (ReLU) is employed as an activation function.

In the road extraction task, there are many issues that can cause a limited detection performance. First, based on [6] and [8], although the most recent DCED approach for object segmentation, namely SegNet, showed promising detection performance on overall classes, the result on road objects is still limited since it misses to detect many road objects. This should be caused by ReLU is sensitive to the gradient vanishing issue. Second, it is common to apply Gaussian smoothing at the last step to connect detected roads together. This yields an issue of excessive detected road objects (false road objects).

In this paper, we present an improved deep convolutional encoder-decoder network (DCED) for segmenting road objects from aerial and satellite images. Several aspects of the proposed method are enhanced, incl. incorporation of ELU (exponential linear unit)—as opposed to ReLU that typically outperforms ELU in most object classification cases; and next, adoption of landscape metrics to further improve the overall quality of results by removing false road objects. Finally, we even get the better result when combined with the traditional fully connected conditional random fields (CRFs) algorithms used in semantic segmentation problem. The nature of ELU-SegNet network restricts the performance due to the loss of spatial accuracy. However, the drawbacks of convolutional networks can be alleviated by conditional random fields algorithm, which takes in the low-level information captured by the local interactions of pixels and edges [17], [18], and [19]. Accordingly, it is worthy to discover the details of the ELU-SegNet-LMs-CRFs model as discussed in this paper. SegNet is used as one of the benchmarks in evaluating our method. The experiments were conducted on a well-known aerial imagery, Massachusetts roads data set (Mass. Roads), which is publicly available and satellite imagery (THEOS satellite) which is provided by GISTDA. The results showed that our method outperforms all of the baselines in terms of precision, recall, and F1 scores.

The paper is organized as follows. Related work is discussed in section 2. Section 3 describes our proposed method. Experimental data sets and evaluation were described in section 4. Next, 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 [20] and has increasingly become a promising tool for accelerating image recognition process with high accuracy results [4][6][21] 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 road object segmentation techniques using deep learning, and finally activation functions and post processing technique of deep learning are discussed.

2.1. Deep Learning for Semantic Segmentation

Semantic segmentation algorithms are often formulated to solve structured pixel-wise labeling problems based on deep convolutional neural network (DCNN), the state-of-the-art supervised learning algorithms in modeling and extracting latent features hierarchies. Noh et al. [5] proposed a novel semantic segmentation technique utilizing a deconvolutional neural network (DeCNN) and the top layer from DCNN adopted from VGG16 [22]. 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 data set [16], 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. [4] proposed an adapted contemporary classification networks incorporating Alex, VGG and GoogLe networks into fully DCNN. 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% to 62.2% in the experiments tested on PASCAL VOC 2012 data. Ronneberger et al. [12] 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. In this work, VGG16 is selected as our baseline architecture since it is the most popular architecture used in various networks for object recognition. Furthermore, we will investigate the effect of the skipped layer technique, especially FCN-8s, since it is the top-ranking architecture as shown in Long et al. [4].

2.2. Deep Learning for Road Segmentation

There have been many approaches in road network extraction from very-high-resolution (VHR) aerial and satellite imagery literature. Wand et al. [14] proposed a DCNN and FSM (finite state machine)-based framework to extract road networks from aerial and satellite images. DCNN recognizes patterns from a sophisticated and arbitrary environment while FSM translates the recognized patterns to states such that their tracking behaviors can be captured. The results showed that their approach is more accurate compared to the traditional methods. The extension of the method for automatic road point initialization was left for future work. DCNN for multiple object extraction from aerial imagery was proposed in [3] by Saito et al. Both features (extractors and classifiers) of DCNN were automated in that a new technique to train a single DCNN for extracting multiple kinds of objects simultaneously was developed. Two objects were extracted: buildings and roads, thus a label image consists of three channels: buildings, roads, and background. Finally, the results showed that the proposed technique not only improved the prediction performance but also outperformed the cutting-edge method tested on a publicly available aerial imagery data set. Muruganandham et al. [2] designed an automated framework to extract semantic maps of roads and highways, so the urban growth of cities from satellite images can be tracked. They used VGG16 model—a simplistic architecture with homogeneous 3x3 convolution kernels and 2x2 max pooling throughout the pipeline—as a baseline for fixed feature extractor. The experimental results showed that their proposed technique for the prediction performance was improved with F1 scores of 0.76 on the Mass. Roads data set.

2.3. Recent Techniques in Deep Learning

Activation function is an important factor for an accuracy of DCNN. While the most popular activation function for neural networks is the rectified linear unit (ReLU), Clevert et al. [21] have just proposed exponential linear unit (ELU), which can speed up the learning process in DCNN and therefore, lead to higher classification accuracies as well as overcome the previously unsolvable problem, i.e., vanishing gradient problem. Comparing to other methods with different activation functions, ELU has greatly improved many of the learning characteristics. In the experiments, ELUs enable fast learning as well as more effective generalization performance than ones of ReLUs and LReLUs (leaky rectified linear unit) on the networks with five layers or more. In ImageNet, ELU networks substantially increase the learning time compared to ReLU networks with the identical architecture; less than 10% classification error was presented for a single crop, model network.

Recently, there are some efforts to enhance a performance of DCNN by combining it with other classifier as a post-processing step. Conditional Random Fields (CRFs) has been reported its success to increase an accuracy of DCNN especially in the image segmentation domain. CRFs have been employed to smooth maps [7], [17], [18], and [19] Typically these models contain energy terms that couple neighboring nodes, favoring same-label assignments to spatially proximal pixels. Qualitatively, the primary function of these short-range CRFs has been to clean up the spurious predictions of weak classifiers built on top of local hand-engineered features.

3. Proposed Method

We proposed an enhanced, improved DCED network (or SegNet) to efficiently segment road objects from aerial and satellite images. Three aspects of the proposed method are enhanced: ***(i)*** modification of DCED architecture, ***(ii)*** using landscape metrics (LMs), and ***(iii)*** adoption of conditional random fields (CRFs). Overview of the all proposed was shown in Figure 1.

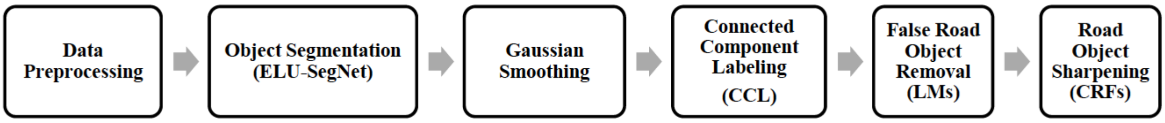

Data Preprocessing
Object Segmentation (ELU-SegNet)
Gaussian Smoothing
Connected Component Labeling (CCL)
False Road Object Removal (LMs)
Road Object Sharpening (CRFs)

Figure 1. A process of our proposed framework

3.1. Data Preprocessing

Data preparation is required when working with neural network and deep learning models. Increasingly data augmentation is also required on more complex object recognition tasks.

So, we increase the size of our data sets to improve efficiency of the method by rotating them incrementally with eight different angles. All images on Massachusetts road data sets are standardized and cropped into 1,500x1,500 pixels with a resolution of 1 m^2/pixel. The data sets consist of 1,108 training images, 49 test images, and 14 validation images. The original training images were further extended to 8,864 training images.

On TEHOS dat sets, we also increase the size of our data sets by rotating them incrementally with eight different angles. All images are standardized and cropped into 1,500x1,500 pixels with a resolution of 2 m^2/pixel.

3.2. Object Segmentation (ELU-SegNet)

SegNet, one of the deep convolutional encoder-decoder architectures, consists of two main networks encoder and decoder, and some outer layers (see Figure 2). 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. The encoder network consists of convolution layer and pooling layer. A technique, called batch normalization (proposed by Ioffe and Szegedy [23]), is used to speed up the learning process of the DCNN by reducing internal covariate shift. In the encoder network, the number of layers are reduced to 13 layers (VGG16) by removing the last three layers (fully connected layers) [6], [8], [24], and [25] due to the following two reasons: to maintain the high-resolution feature maps in the encoder network, and to minimize the countless number of parameters from 134 million features to 14.7 million features compared to the traditional deep learning networks such as DCNN [4] and DeCNN [5], where the fully connected layer remains intact. In the activation function of feature extraction, ReLU, max-pooling, and 7x7 kernel are used in both encoder and decoder networks. For training images, three-channel images (r/g/b) are used. Exponential Linear Unit (ELU) was introduced in [21], which can speed up learning in deep neural networks, offer higher classification accuracies, and give better generalization performance than ReLUs and LReLUs on networks. In SegNet architecture, to do optimization for training networks, stochastic gradient descent (SGD) [26] with a fixed learning rate of 0.1 and momentum of 0.9 are used. In each training round (epoch), a mini-batch (a set of 12 images) is chosen such that each image is used once. The model with the best performance on the validation data set in each epoch will be selected. Our architecture (see Figure 2) is enhanced from SegNet, consisting of two main networks responsible for feature extraction. In each network, there are 13 layers with the last layer being the classification based on softmax supporting pixel-wise classification. In our work, an activation function called ELU is used—as opposed to ReLU—based on its performances. For the network training optimization, stochastic gradient descent (SGD) is used and configured with a fixed learning rate of 0.001 and momentum of 0.9 to delay the convergence time and so, can avoid local optimization trap.

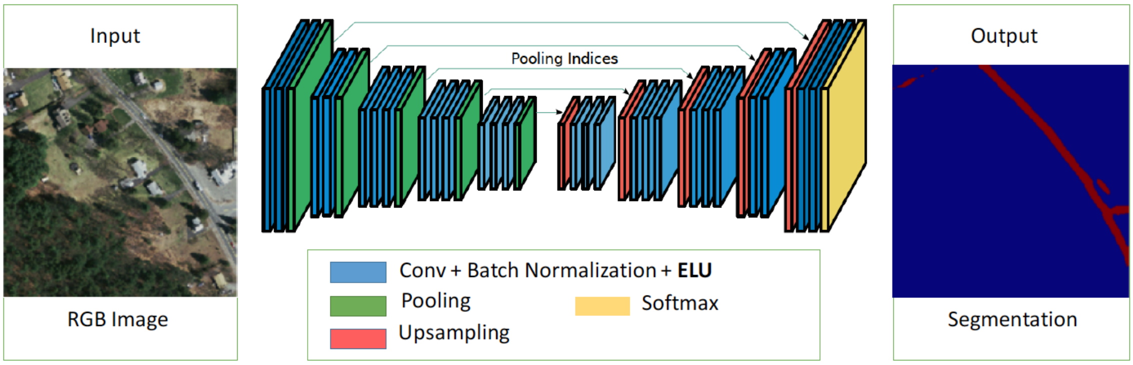


Figure 2. A proposed network architecture for object segmentation (ELU-SegNet)

3.3. Gaussian Smoothing

The Gaussian smoothing technique [27] is a 2d convolution operator that is used to 'blur' images and remove detail and noise. It is a type of image-blurring filter that uses a Gaussian function (which also expresses the normal distribution in statistics) for calculating the transformation to apply to each pixel in the image.

In this paper, Gaussian smoothing technique is the first step of post processing technique. It aims to expand objects close to each other for can put together to be one object by using the next algorithm (in Section 3.4.). The equation of a Gaussian function in one dimension is

$$G(x) = \frac{1}{2\pi\sigma^2}e^{-\frac{x^2}{2\sigma^2}} \tag{1}$$

In two dimensions, it is the product of two such Gaussian, one in each dimension:

$$G(x) = \frac{1}{2\pi\sigma^2}e^{\frac{-x^2-y^2}{2\sigma^2}} \tag{2}$$

where x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis, and σ is the standard deviation of the Gaussian distribution.

doi:10.20944/preprints201706.0012.v1

3.4. Connected Component Labeling (CCL)

Connected components labeling (CCL) [27] scans an image and groups its pixels into components based on pixel connectivity, i.e., all pixels in a connected component share similar pixel intensity values and are in some way connected with each other.

In this paper, CCL is the step after gaussian smoothing technique finished. It aims to group the pixels into connected components based on pixel connectivity (eight neighbours) and calculates some geometrical attributes for each component, such as area and perimeter for using in calculated on Landscape Metrics (LMs) approach (in Section 3.5.).

3.5. False Road Object Removal (LMs)

After connected components labeling and gaussian smoothing technique finished. In this paper, we compute all object that get from the past processing step by shape metrics (one of the landscape metrics for measuring spatial object complexity) is used [28]. Sample result that get from computing shape score was shown in Figure 3. Geometrical characteristics of the roads are captured and differentiated from other spatial objects in the given image. Other geometry metrics can also be used such as rectangular degree, aspect ratio, etc. More information on other landscape metrics can be found in [28], [29].

$$shape\ index = \frac{e(i)}{4x\sqrt{A(i)}} \tag{3}$$

where e(i) and A(i) denote the perimeter and area for object i, respectively.

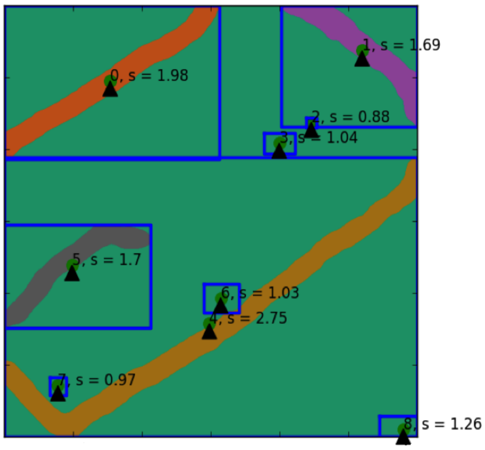


Figure 3. Illustration of shape index scores on each extracted road object. Any objects with shape index score lower than 1.25 are considered as noises and will be removed

3.6. Road Object Sharpening (CRFs)

We explore extending our ELU-SegNet-LMs model to ELU-SegNet-LMs-CRFs model by adding explicit dependencies between outputs of a neural network. Especially, we add a smoothness term between neighboring pixels to our model and removing the need to learn smoothness from the satellite image. Using the resulting models for post-processing leads to improvements over unstructured and post-processing deep neural networks.

Traditionally, conditional random fields (CRFs) have been employed to smooth noisy segmentationmaps [18]. Typically, these models contain energy terms that couple neighboring nodes, favoring same-label assignments to spatially proximal pixels. To overcome these limitations of short-range CRFs, we integrate into our system the fully connected CRF model of [19]. The energy function of the dense CRFs is

Table 1. Number of training, validation, and testing sets

	Training set	Validation set	Testing set
Massachusetts	1,108	14	49
Nakhonpathom	200	14	49
Chonburi	100	14	49
Songkhla	100	14	49
Surin	70	14	49
Ubonratchathani	70	14	49

$$E(x) = \sum_{i} \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \tag{4}$$

where x is the label assignment for pixels. We use as unary potential $\theta_i(x_i)) = -logP(x_i)$, while $P(x_i)$ is the label assignment probability at pixel i as computed by a DCNN. The pairwise potential has a form that allows for efficient inference while using a fully-connected graph.

Particularly, the unary potentials can be treated as local classifiers and well defined by the output of the ELU-SegNet-LMs model (The output of the ELU-SegNet-LMs on satellite image is a probability map for each classes and each pixel). The pairwise potentials usually model the relationship among neighboring pixels and weighted by color similarity. In the DeepLab CRFs model [19], they use dense CRFs, which means the model considers long range interactions among pixels instead of neighboring information. Furthermore, they define the following pairwise potential

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j)[w_1 \exp(-\frac{\| p_i - p_j \|^2}{2\sigma_\alpha^2} - \frac{\| I_i - I_j \|}{2\sigma_\beta^2{}^2}) + w_2 \exp(-\frac{\| p_i - p_j \|^2}{2\sigma_\gamma^2{}^2})] \tag{5}$$

where $\mu(x_i, x_j) = 1 \ if \ x_i \neq x_j$ and zero otherwise, which, as in the Potts model, means that only nodes with distinct labels are penalized. The remaining expression uses two Gaussian kernels in different feature spaces; the first, 'bilateral' kernel depends on both pixel positions (denoted p) and RGB color (denoted as I), and the second kernel only depends on pixel positions. The hyperparameters $\sigma_\alpha, \ \sigma_\beta \ and \ \sigma_\gamma$ control the scale of Gaussian kernels. The first kernel forces pixels to similar color and position to have similar labels, while the second kernel only considers spatial proximity when enforcing smoothness.

Basically, The first term depends on both pixel positions and pixel color intensities and the second term only depends on pixel positions [18], [19]. However, the dense CRF has billion edges, which it is computationally hard to inference. Nevertheless, the mean-field algorithm allows us to approximate the maximum posterior efficiently.

4. Experimental Data Sets and Evaluation

In our experiments, two types of data sets are used: aerial images and satellite images. Table 1 shows one aerial data set (Massachusetts) and five satellite data sets (Nakhonpathom, Chonburi, Songkhla, Surin, and Ubonratchathani). All experiments are evaluated based on $precision$, $recall$, and $F1$.

4.1. Massachusetts Road Data Set (Aerial Imagery)

This data set (made publicly available by [7]) consists of 1,171 aerial images of the state of Massachusetts. Each image is 1,500x1,500 pixels in size, covering an area of 2.25 square kilometers. We randomly split the data into a training set of 1,108 images, a validation set of 14 images and a testing set of 49 images. The samples of this data set are shown in Figure 4. The data set covers a wide variety of urban, suburban, and rural regions with a total area of over 2,600 square kilometers. With our test set alone, it covers more than 110 square kilometers which is by far the largest and most challenging aerial image labeling data set.

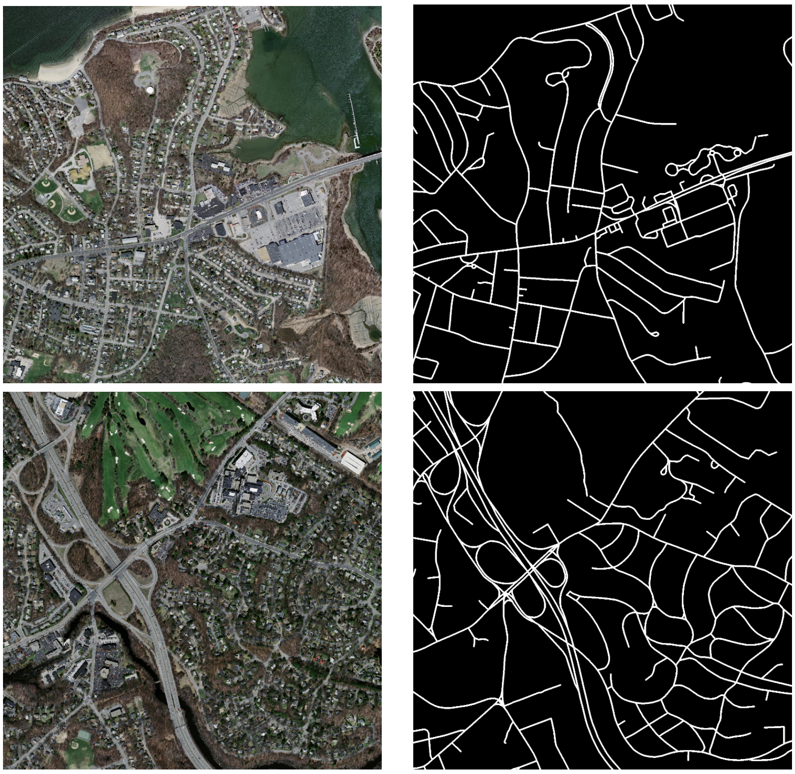
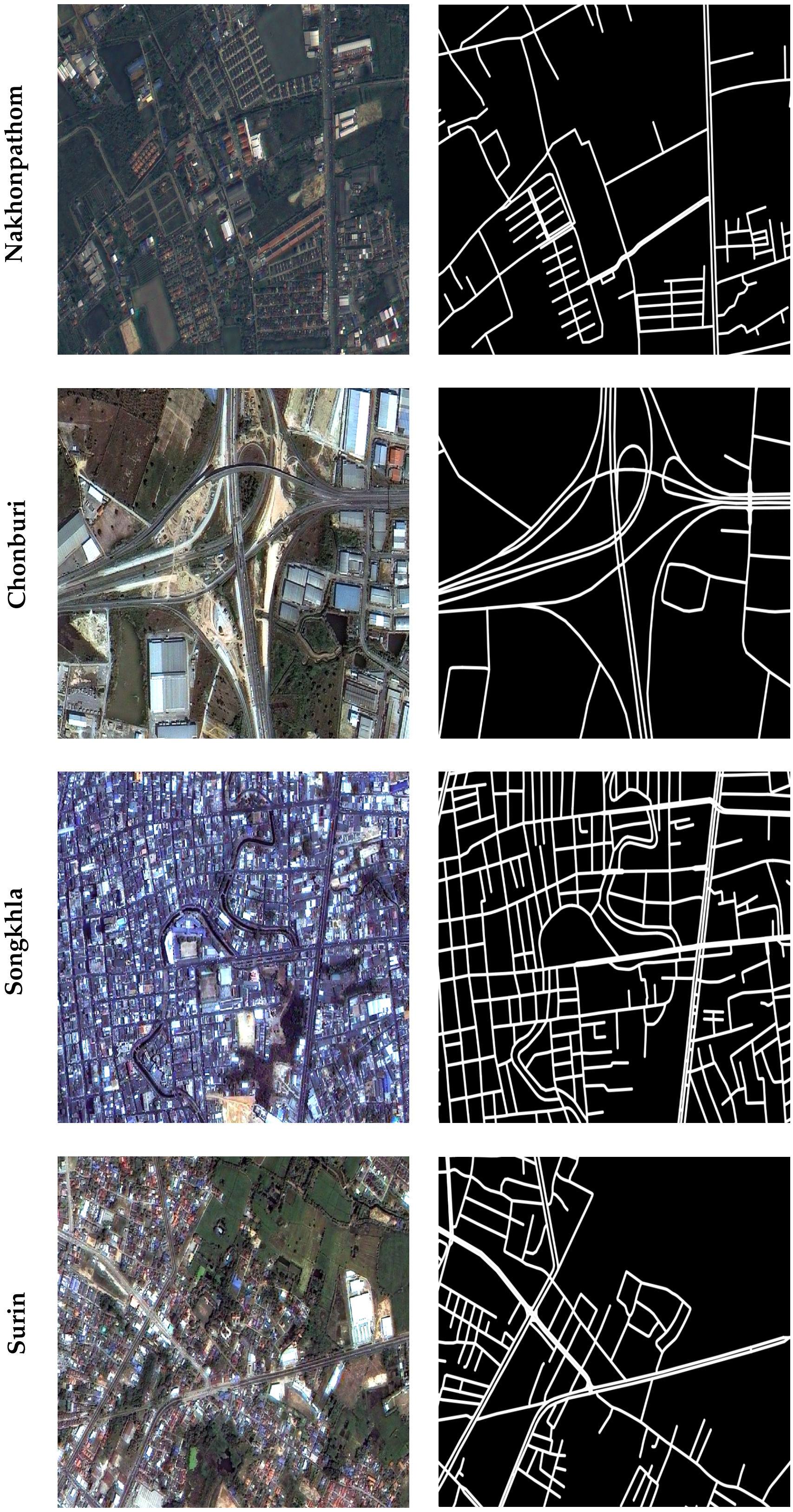


Figure 4. Two sample aerial images from Massachusetts road corpus, where a row refers to each image **(a)** aerial image and **(b)** binary maps which is a ground truth image; denoting the location of roads

4.2. THEOS Data Sets (Satellite Imagery)

In this data set was separated into 5 data sets which contained 5 provinces (1 data set/province). THEOS, also known as Thailand Earth Observation System or Thaichote, is an earth observation mission of Thailand, developed at EADS Astrium SAS, Toulouse, France. In July 2004, EADS Astrium SAS signed a contract for delivery of THEOS with GISTDA (Geo-Informatics and Space Technology Development Agency) of Bangkok, Thailand. GISTDA is Thailand's leading national organization (For example, space agency) in the field of space activities and applications. The Thai Ministry of Science and Technology funds the program.

This data set consists of 855 satellite images, the sample of this data set was shown in Figure 5. The five data sets was seperated to 263 satellite images (200 training images, 49 testing images, and 14 validation images) of the state of Nakhonpathom, 163 satellite images (100 training images, 49 testing images, and 14 validation images) of the state of Chonburi, 163 satellite images (100 training images, 49 testing images, and 14 validation images) of the state of Songkhla, 133 satellite images of the state (70 training images, 49 testing images, and 14 validation images) of Surin, and 133 satellite images of the state (70 training images, 49 testing images, and 14 validation images) of Ubonratchathani.


Nakhonpathom
Chonburi
Songkhla
Surin

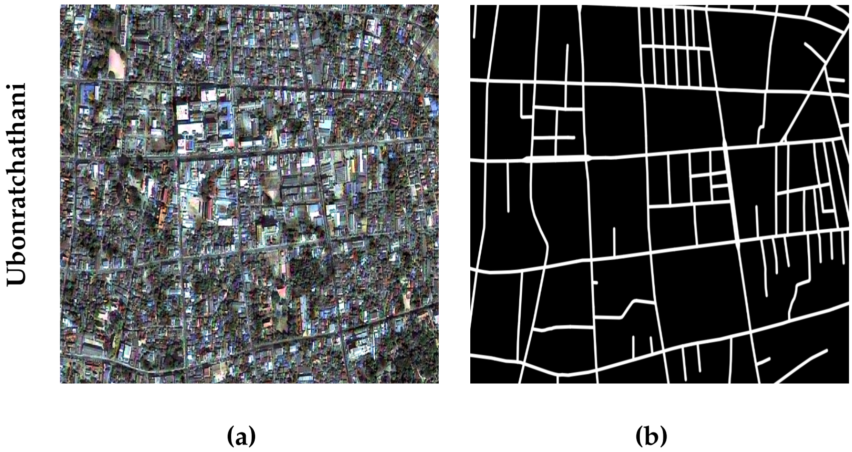


Figure 5. Sample satellite images from five provinces of our data sets, each row refers to a single sample image from one province (Nakhonpathom, Chonburi, Songkhla, Surin, and Ubonratchathani) in a satellite image format **(a)** and in a binary map **(b)**, which is served as a ground truth image; denoting the location of roads

4.3. Evaluation

The road extraction task can be considered as binary classification, where road pixels are positives and the remaining non-road pixels are negatives. Let TP denote the number of true positives (the number of correctly classified road pixels), TN denote the number of true negatives (the number of correctly classified non-road pixels), FP denote the number of false positives (the number of mistakenly classified road pixels), and FN denote the number of false negatives (the number of mistakenly classified non-road pixels).

The performance measures used are precision, recall, and F1 as shown in equations (Eq. 6-8). Precision is the percentage of correctly classified road pixels among all predicted pixels by the classifier. Recall is the percentage of correctly classified road pixels among all actual road pixels. F1 is a combination of precision and recall.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precison + Recall} \tag{8}$$

5. Experimental Results and Discussions

This section illustrates details of the experiment on two kinds of remotely-sensed data: aerial and satellite images. The proposed deep learning network is based on SegNet with three improvements: ***(i)***, employ ELU activation function, ***(ii)*** use LMs to filter incorrect detected roads, and ***(iii)*** apply CRFs to sharpen broad roads. Thus, there are three variations of the proposed methods as shown in 3.

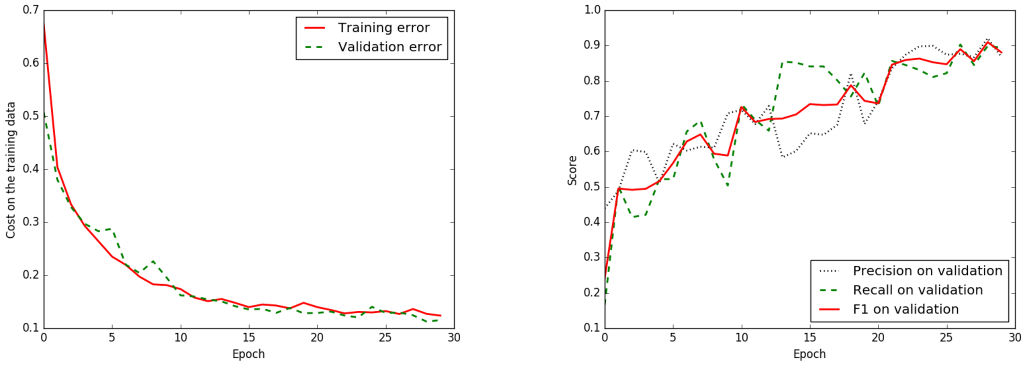
Table 3. Variations of our proposed deep learning methods

Abbreviation	Description
ELU-SegNet	SegNet + **ELU activation**
ELU-SegNet-**LMs**	SegNet + ELU activation + **Landscape Metrics**
ELU-SegNet-LMs-**CRFs**	SegNet + ELU activation + Landscape Metrics + **CRFs**

The implementation is based on a deep learning framework, called "Lasagne", which is extended from Theano. All experiments were conducted on a server with Intel Core i5-4590S Processor (6M Cache, up to 3.70 GHz), 8 GB of memory, and Nvidia GeForce GTX 960 (4 GB) and Nvidia GeForce GTX 1080 (8 GB). In stead of using the whole image (1,500×1,500 pixels) to train the network, we randomly crop all images to be 224×224 as inputs of each epoch.

5.1. Results on Aerial Imagery (Massachusetts Data Set)

In this sub-section, the experiment was conducted on Massachusetts aerial corpus. To achieve highest accuracy, the network must be configured and trained many epochs until all parameters in the network are converged. Figure 6(a) illustrates that the proposed network has been properly set and trained until it is really converged. Furthermore, Figure 6(b) shows that the higher number of epochs tend to show better $F1$-score. Thus, the number of chosen epochs based on the validation data is 27 epochs.



(a)

(b)

Figure 6. Iteration plot on Massachusetts aerial corpus of the proposed technique, ELU-SegNet-LMs-CRFs; 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

The result is shown in Table 4 by comparing between baselines and variations of the proposed techniques. It shows that our network with all strategies (ELU-SegNet-LMs-CRFs) 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 four baselines including Basic-model, FCB-no-skip, FCN-8s, and SegNet. Note that SegNet has been implemented and tested on the experimental data set, while the results of other three baselines are carried from the original paper [2].

Table 4. Results on the testing data of Massachusetts aerial corpus between four baselines and three variations of our proposed techniques in terms of *precision*, *recall*, and $F1$

	Model	**Precsion**	**Recall**	**F1**
Baselines	Basic-model [2]	0.657	0.657	0.657
	FCN-no-skip [2]	0.742	0.742	0.742
	FCN-8s [2]	0.762	0.762	0.762
	SegNet	0.773	0.765	0.768
Proposed Method	**ELU**-SegNet	0.852	0.733	0.788
	ELU-SegNet-**LMs**	0.854	0.861	0.857
	ELU-SegNet-LMs-**CRFs**	**0.858**	**0.894**	**0.876**

5.1.1. Results of Enhanced SegNet (ELU-SegNet)

Our first strategy aims to increase an accuracy of the network by using ELU as an activation function (ELU-SegNet) rather than the traditional one, ReLU (SegNet). Details are shown in Section 3.2. From Table 4, $F1$ of ELU-SegNet (0.788) outperforms that of SegNet (0.768); this yields higher $F1$ for 2.6%. The main reason is due to higher *precision*, but slightly lower *recall*. This can imply that ELU is more robust than ReLU to detect road pixels.

5.1.2. Results of Enhanced SegNet with Landscape Metrics (ELU-SegNet-LMs)

Our second mechanism focuses on applying LMs (details in Section 3.5) on top of ELU-SegNet to filter false road objects. From Table 4, $F1$ of ELU-SegNet-LMs (0.857) is superior to that of ELU-SegNet (0.788) and SegNet (0.768); this yields higher $F1$ for 6.9% and 8.9%, consecutively. Although LMs is specifically designed to increase *precision*, the result shows that it can increase both *precision* (0.854) and *recall* (0.861). It is interesting that *recall* is also improved since all noises in the training images have been removed by the LMs filtering technique resulting in a better quality of the training data set.

5.1.3. Results of All Modules (ELU-SegNet-LMs-CRFs)

Our last strategy aims to sharpen road objects (details in Section 3.6) by integrating CRFs into our deep learning network. From Table 4, $F1$ of ELU-SegNet-LMs-CRFs (0.876) is the winner; it clearly outperforms not only baselines, but also all previous generations. Its $F1$ is higher than SegNet (0.768) for 10.8%. Also, the result illustrates that CRFs can enhance both *precision* (0.858) and *recall* (0.894).

Figure 7 shows two sample results from the proposed method. By applying all strategies, the images in the last column (Figure 7(e)) look really close to the ground truths (Figure 7(b)). Furthermore, $F1$-results are improved for each strategy we added to the network as shown in Figure 7(c) to (e).

5.2. Results on Satellite Imagery (THEOS Data Sets)

In this sub-section, the experiment was conducted on THEOS satellite images. There are five data sets referring to different provinces: Nakhonpathom, Chonburin, Songkla, Surin, and Ubonratchathani; therefore, there are five learning models. Figure 8 shows that each model is properly setup and trained until it is converged and obtained the best $F1$. The best epochs for each province are 25, 15, 30, 21, and 20, consecutively.

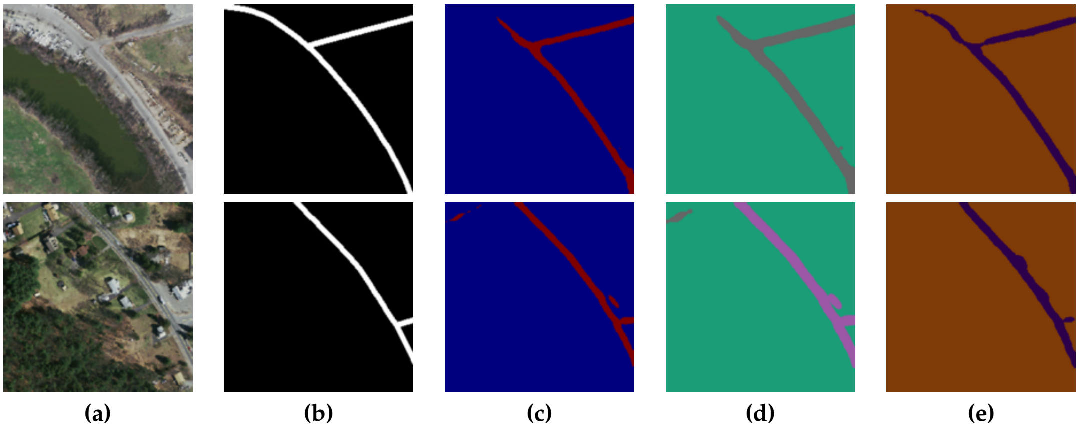
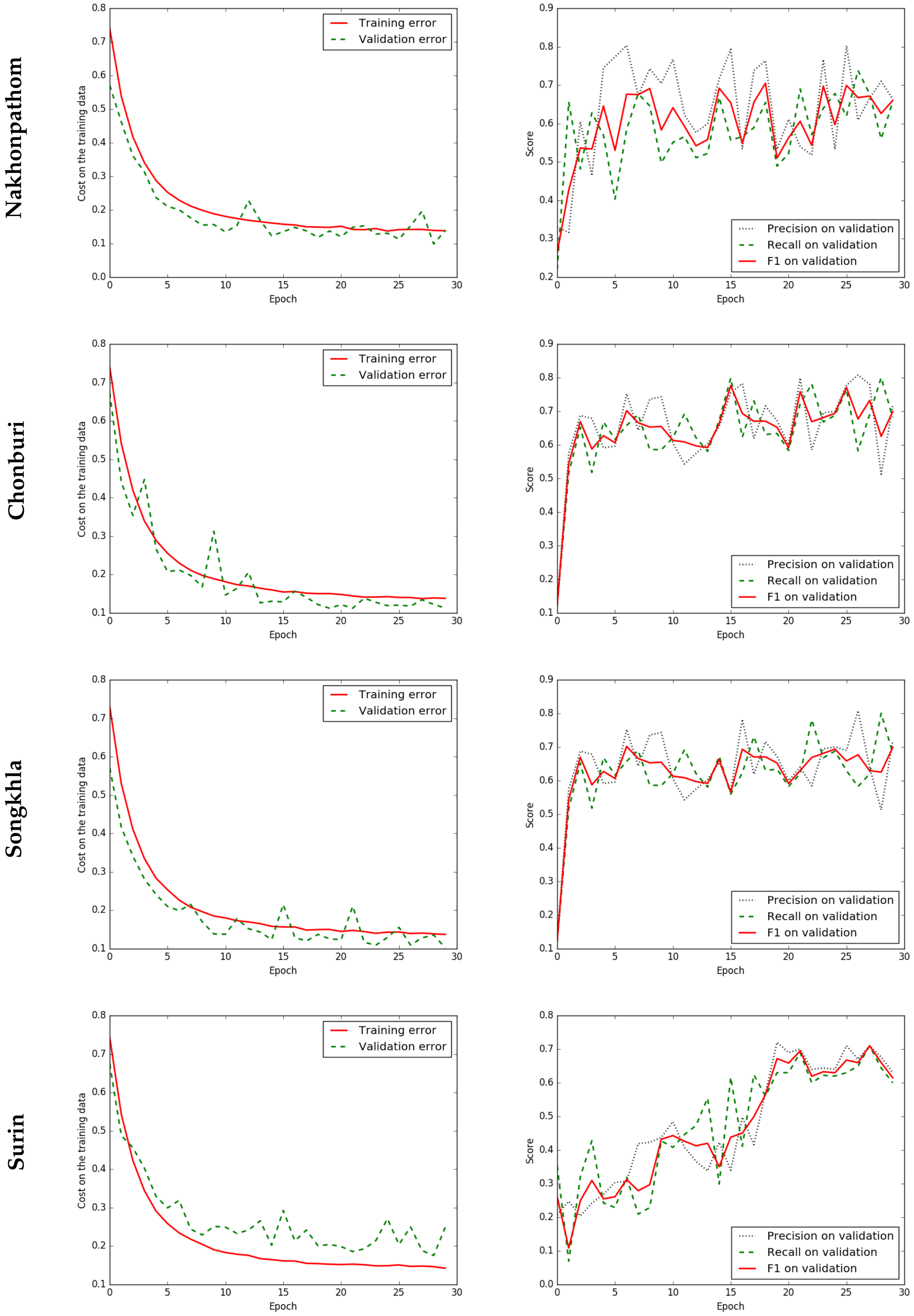


Figure 7. Two sample input and output aerial images on Massachusetts corpus, where rows refer different images **(a)** original input image; **(b)** target road map (ground truth); **(c)** ELU-SegNet's output; **(d)** ELU-SegNet-LMs's output; and **(e)** ELU-SegNet-LMs-CRFs's output


Nakhonpathom
Chonburi
Songkhla
Surin
Training error
Validation error
Cost on the training data
Epoch
Precision on validation
Recall on validation
F1 on validation
Score
0
5
10
15
20
25
30
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

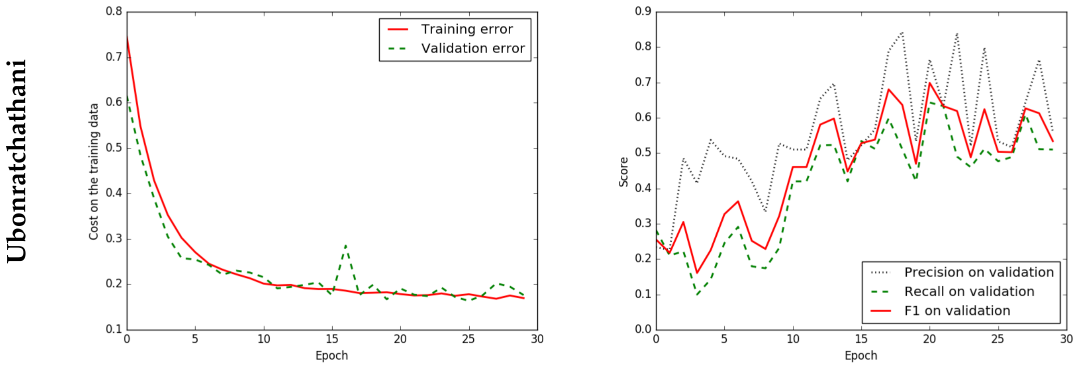


Figure 8. Iteration plot on THEOS satellite data sets of the proposed technique, ELU-SegNet-LMs-CRFs. x refers to epochs and y refers to different measures. Each row refers to different data set (province) **(a)** Plot of model loss (cross entropy) on training and validation data sets, and **(b)** Performance plot on the validation data set

The results are shown in Tables 6, 7, and 8 for measures in terms of $F1$, *precision*, and *recall*, respectively. It is interesting that the proposed network with all strategies (ELU-SegNet-LMs-CRFs) is the winner showing the best performance on any measures and provinces. Also, an improvement in the satellite images is higher than that in the aerial images. More details on each proposed strategy will be discussed.

Table 6. $F1$ on the testing data of THEOS satellite data sets between baseline (SegNet) and three variations of our proposed techniques; columns refer to five different provinces (data sets)

	Model	**Nakhon.**	**Chonburi**	**Songkhla**	**Surin**	**Ubon.**
Baseline	SegNet	0.422	0.572	0.424	0.501	0.406
Proposed Method	**ELU**-SegNet	0.463	0.690	0.497	0.591	0.534
	ELU-SegNet-**LMs**	0.488	0.732	0.526	0.625	0.562
	ELU-SegNet-LMs-**CRFs**	**0.550**	**0.775**	**0.607**	**0.707**	**0.608**

Table 7. *precision* on the testing data of THEOS satellite data sets between baseline (SegNet) and three variations of our proposed techniques; columns refer to five different provinces (data sets)

	Model	**Nakhon.**	**Chonburi**	**Songkhla**	**Surin**	**Ubon.**
Baseline	SegNet	0.435	0.668	0.456	0.598	0.601
Proposed Method	**ELU**-SegNet	0.410	0.702	0.478	**0.840**	0.852
	ELU-SegNet-**LMs**	0.494	0.852	0.557	0.770	0.867
	ELU-SegNet-LMs-**CRFs**	**0.535**	**0.909**	**0.650**	0.786	**0.871**

Table 8. *recall* on the testing data of THEOS satellite data sets between baseline (SegNet) and three variations of our proposed techniques; columns refer to five different provinces (data sets)

	Model	**Nakhon.**	**Chonburi**	**Songkhla**	**Surin**	**Ubon.**
Baseline	SegNet	0.410	0.499	0.395	0.431	0.306
Proposed Method	**ELU**-SegNet	0.532	**0.678**	0.517	0.456	0.389
	ELU-SegNet-**LMs**	0.483	0.642	0.498	0.526	0.416
	ELU-SegNet-LMs-**CRFs**	**0.566**	**0.676**	**0.570**	**0.643**	**0.467**

doi:10.20944/preprints201706.0012.v1

5.2.1. Results of Enhanced SegNet (ELU-SegNet)

ELU activation function can increase the performance of the network. In terms of $F1$, Table 6 shows that ELU-SegNet outperforms the traditional network (SegNet) on all provinces. It is higher than SegNet for 9.08% on average of all provinces, where Ubonratchathani and Chonburi show the highest $F1$-improvement greater than 10%. For *precision* and *recall*, Tables 7 and 8 illustrate that almost all data sets can be improved by employing the ELU function with 10.48% and 10.62%, consecutively, on average of all provinces.

5.2.2. Results of Enhanced SegNet with Landscape Metrics (ELU-SegNet-LMs)

The LMs filtering strategy aims to remove all inaccurately extracted roads (false positives: FP) resulting in higher *precision* and $F1$, but might loose a slight *recall*. Comparing to the previous generation (ELU-SegNet), there are improvements by LMs on average of all provinces for 3.16% and 5.16% in terms of *precision* (Table 7) and $F1$ (Table 6) with a slight lost for -1.22% in terms of *recall* (Table 8). Comparing to the baseline, LMs outperforms SegNet on any performance measures.

5.2.3. Results of All Modules (ELU-SegNet-LMs-CRFs)

To further improve the performance, CRFs is integrated into the network from the previous section. This is considered as using all proposed modules: ELU, LMs, and CRFs. From Table 7, 8, 6, the results show that ELU-SegNet-LMs-CRFs is the winner comparing the previous generations and baseline (SegNet) on any measures (*precision*, *recall*, and $F1$). As of $F1$ average of all provinces, it outperforms ELU-SegNet-LMs, ELU-SegNet, and SegNet for 6.28%, 9.44% and 18.48%, respectively.

Figures 9, 10, 11, 12, and 13 show sample results from the proposed method on five provinces. The results of the last column look closest to the ground truth in the second column.

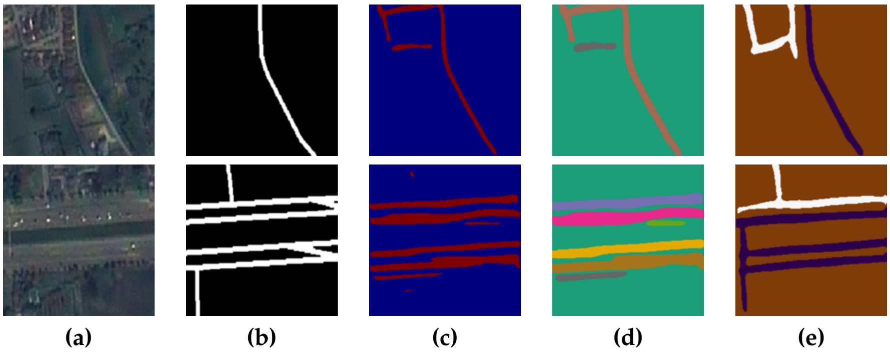


Figure 9. Two sample input and output THEOS' satellite images on Nakhonpathom data set, where rows refer different images **(a)** original input image; **(b)** target road map (ground truth); **(c)** ELU-SegNet's output; **(d)** ELU-SegNet-LMs's output; and **(e)** ELU-SegNet-LMs-CRFs's output

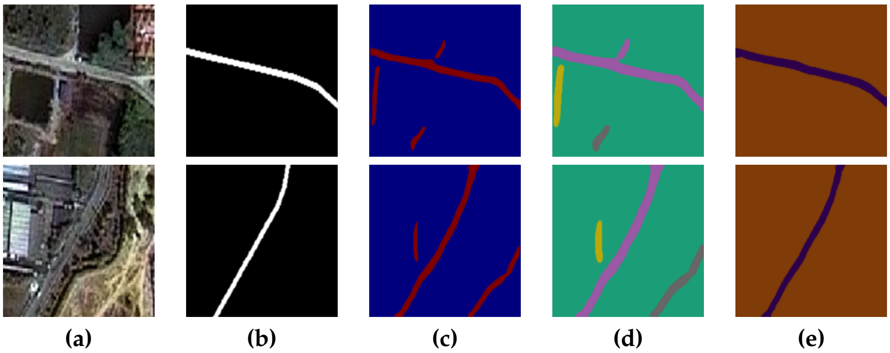


Figure 10. Two sample input and output THEOS' satellite images on Chonburi data set, where rows refer different images **(a)** original input image; **(b)** target road map (ground truth); **(c)** ELU-SegNet's output; **(d)** ELU-SegNet-LMs's output; and **(e)** ELU-SegNet-LMs-CRFs's output

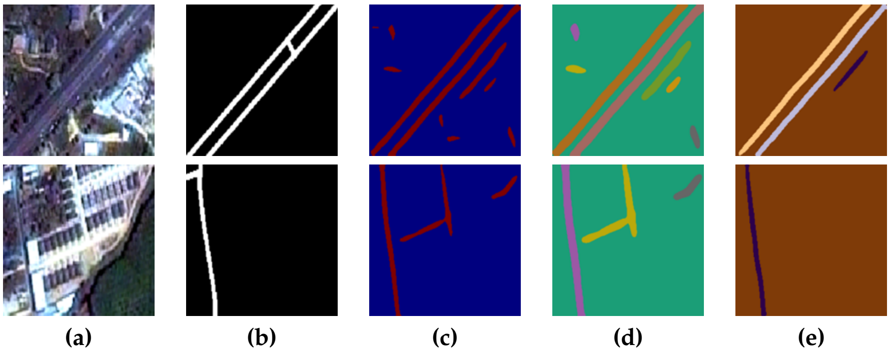


Figure 11. Two sample input and output THEOS' satellite images on Songkhla data set, where rows refer different images **(a)** original input image; **(b)** target road map (ground truth); **(c)** ELU-SegNet's output; **(d)** ELU-SegNet-LMs's output; and **(e)** ELU-SegNet-LMs-CRFs's output

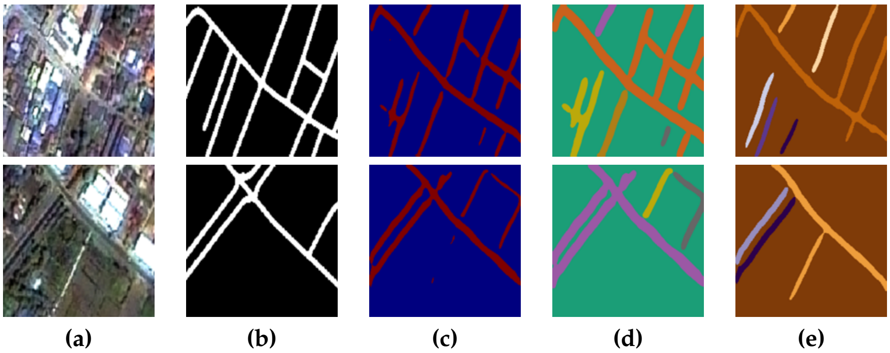


Figure 12. Two sample input and output THEOS' satellite images on Surin data set, where rows refer different images **(a)** original input image; **(b)** target road map (ground truth); **(c)** ELU-SegNet's output; **(d)** ELU-SegNet-LMs's output; and **(e)** ELU-SegNet-LMs-CRFs's output

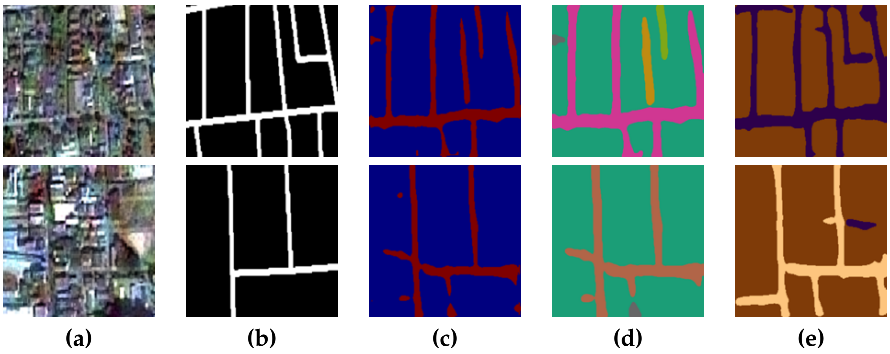


Figure 13. Two sample input and output THEOS' satellite images on Ubonratchathani data set, where rows refer different images **(a)** original input image; **(b)** target road map (ground truth); **(c)** ELU-SegNet's output; **(d)** ELU-SegNet-LMs's output; and **(e)** ELU-SegNet-LMs-CRFs's output

6. Conclusions and Future Work

In this study, we present a novel deep learning network framework to extract road objects from aerial and satellite images. The network is based on Deep Convolutional Encoder-Decoder Network (DCED), called "SegNet." To improve the network's precision, we incorporate the recent activation function, called Exponential Linear Unit (ELU), into our proposed method. The method is also further improved to detect more road patterns by utilizing landscape metrics and conditional random fields. Excessive detected roads are then eliminated by applying landscape metrics thresholding. Finally, we extend the SegNet network to ELU-SegNet-LMs-CRFs. The experiments were conducted on Massachusetts roads data set as well as THEOS (Thailand) roads data sets and compared to the existing techniques. The results show that our proposed (ELU-SegNet-LMs-CRFs) outperforms the original method on both aerial and satellite imagery for F1—as well as for all other baselines.

In future work, more choices of image segmentation, optimization techniques and/or other activation functions will be investigated and compared to obtain the best DCED-based framework for semantic road segmentation.

Acknowledgments: T. Panboonyuen thanks the scholarship from Chulalongkorn University to commemorate the 72^{nd} Anniversary of H.M. King Bhumibala Aduladeja.

Author Contributions: The experiment design was carried out by all of the authors. Teerapong Panboonyuen and Peerapon Vateekul performed the experiments and results analysis. Kulsawasd Jitkajornwanich, SiamLawawirojwong and Panu Srestasathiern reviewed results. The article was co-written by the five authors. All authors read and approved the submitted manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

CCL	Connected Component Labeling
CNN	Convolutional Neural Network
CRFs	Conditional Random Fields
DCED	Deep Convolutional Encoder-Decoder
DCNN	Deep Convolutional Neural Network
DL	Deep Learning
ELU	Exponential Linear Unit
FCN	Fully Convolutional Network
FN	False Negative
FP	False Positive
HR	High Resolution
LMs	Landscape Metrics
ReLU	Rectified Linear Unit
RGB	Red-Green-Blue
SGD	Stochastic Gradient Descent
TN	True Negative
TP	True Positive
VGG	Visual Geometry Group
VHR	Very High Resolution
VOC	Visual Object Classes

References

1. Poullis, C. Tensor-Cuts: A simultaneous multi-type feature extractor and classifier and its application to road extraction from satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing* **2014**, *95*, 93–108.
2. Muruganandham, S. Semantic Segmentation of Satellite Images using Deep Learning, 2016.
3. Saito, S.; Yamashita, T.; Aoki, Y. Multiple object extraction from aerial imagery with convolutional neural networks. *Electronic Imaging* **2016**, *2016*, 1–9.
4. 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.
5. 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.
6. 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**.
7. Mnih, V. Machine learning for aerial image labeling. PhD thesis, University of Toronto, 2013.
8. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561* **2015**.
9. Volpi, M.; Ferrari, V. Semantic segmentation of urban scenes by learning local class interactions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 1–9.
10. Liu, J.; Liu, B.; Lu, H. Detection guided deconvolutional network for hierarchical feature learning. *Pattern Recognition* **2015**, *48*, 2645–2655.
11. Hong, S.; Noh, H.; Han, B. Decoupled deep neural network for semi-supervised semantic segmentation. Advances in Neural Information Processing Systems, 2015, pp. 1495–1503.
12. 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.
13. Andrearczyk, V.; Whelan, P.F. Using filter banks in convolutional neural networks for texture classification. *Pattern Recognition Letters* **2016**, *84*, 63–69.
14. Wang, J.; Song, J.; Chen, M.; Yang, Z. Road network extraction: A neural-dynamic framework based on deep learning and a finite state machine. *International Journal of Remote Sensing* **2015**, *36*, 3144–3169.
15. Visin, F.; Ciccone, M.; Romero, A.; Kastner, K.; Cho, K.; Bengio, Y.; Matteucci, M.; Courville, A. Reseg: A recurrent neural network-based model for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 41–48.

doi:10.20944/preprints201706.0012.v1

16. Liu, Z.; Li, X.; Luo, P.; Loy, C.C.; Tang, X. Deep Learning Markov Random Field for Semantic Segmentation. *arXiv preprint arXiv:1606.07230* **2016**.
17. Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst* **2011**, *2*, 4.
18. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* **2014**.
19. 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**.
20. Audebert, N.; Saux, B.L.; Lefèvre, S. Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sensing* **2017**, *9*, 368.
21. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289* **2015**.
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
23. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* **2015**.
24. 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*.
25. 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**.
26. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
27. Gonzalez, R.; Wintz, P. Digital image processing **2008**.
28. McGarigal, K. Landscape metrics for categorical map patterns, 2008.
29. Huang, X.; Zhang, L. Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines. *International Journal of Remote Sensing* **2009**, *30*, 1977–1987.