Research Article

Land Cover Classification Using SegNet with Slope, Aspect, and Multidirectional Shaded Relief Images Derived from Digital Surface Model

Dae Geon Lee,¹ Young Ha Shin,² and Dong-Cheon Lee ²

¹Technology Department, Geo & Co., Ltd., Goyang-si 10495, Republic of Korea

²Department of Environment, Energy & Geoinformatics, Sejong University, Seoul 05006, Republic of Korea

Correspondence should be addressed to Dong-Cheon Lee; dclee@sejong.ac.kr

Received 28 July 2020; Revised 20 August 2020; Accepted 28 August 2020; Published 12 September 2020

Academic Editor: Sang-Hoon Hong

Copyright © 2020 Dae Geon Lee et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Most object detection, recognition, and classification are performed using optical imagery. Images are unable to fully represent the real-world due to the limited range of the visible light spectrum reflected light from the surfaces of the objects. In this regard, physical and geometrical information from other data sources would compensate for the limitation of the optical imagery and bring a synergistic effect for training deep learning (DL) models. In this paper, we propose to classify terrain features using convolutional neural network (CNN) based SegNet model by utilizing 3D geospatial data including infrared (IR) orthoimages, digital surface model (DSM), and derived information. The slope, aspect, and shaded relief images (SRIs) were derived from the DSM and were used as training data for the DL model. The experiments were carried out using the Vaihingen and Potsdam dataset provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) through the International Society for Photogrammetry and Remote Sensing (ISPRS). The dataset includes IR orthoimages, DSM, airborne LiDAR data, and label data. The motivation of utilizing 3D data and derived information for training the DL model is that real-world objects are 3D features. The experimental results demonstrate that the proposed approach of utilizing and integrating various informative feature data could improve the performance of the DL for semantic segmentation. In particular, the accuracy of building classification is higher compared with other natural objects because derived information could provide geometric characteristics. Intersection-of-union (IoU) of the buildings for the test data and the new unseen data with combining all derived data were 84.90% and 52.45%, respectively.

1. Introduction

The field of DL has grown significantly over the past decade coupled with rapid improvements in computer performance. Since McCulloch and Pitts introduced artificial neuron that is a computational model of the neural networks to mimic the human brain in 1943 [1], DL as a branch of machine learning has evolved steadily to this day. In recent years, advances in image processing computer vision, information and communication technology, and geoinformatics have accelerated the development and use of the DL. Many DL tasks involve visual information processing such as object recognition and identification from imagery [2–4]. It is true that the optical images provide rich, diverse, and explicit information.

“Seeing is believing,” i.e., human beings have been relying on the visual information to understand the real world than through any other media. Images obtained from optical sensors are formed by recording reflected light, in the visible spectral range, from the surfaces of the terrain objects. In this aspect, it is not sufficient to reveal real-world features by utilizing image alone.

Performance of the DL depends on the training data that sometimes leads to the issue of overfitting. Various schemes have been suggested to avoid overfitting such as drop-out, early-stopping, regularization, cross-validation, and hyperparameter tuning. However, such efforts might not be the fundamental solutions to prevent overfitting. Extracting intrinsic and characteristic information from the original

data to utilize training data would make DL more robust. Recent deep learning researches with a similar concept are found. Maltezos et al. proposed building detection method by training the CNN model with multidimensional features that include entropy, height variation, intensity, normalized height, planarity, and standard deviation extracted from light detection and ranging (LiDAR) data. Each feature reflects the unique physical property of the objects [5]. Audebert et al. introduced the SegNet DL model with multimodal and multiscale data for semantic labeling. Optical IR images, DSM created from LiDAR data, normalized DSM (nDSM), and normalized difference vegetation index (NDVI) derived from multispectral data are used to train the DL model [6]. Zhou et al. proposed a CNN-based AlexNet DL model with fusion of point cloud data and high-resolution images for land cover classification. Specifically, a stratified segmentation scheme with grey-level cooccurrence matrix (GLCM) [7, 8] was introduced to improve segmentation efficiency for increasing classification accuracy. Mean, variance, homogeneity, dissimilarity, and entropy from GLCM features were DL model training process [9]. Alidoost and Arefi applied LeNet-5 [10] for automatic recognition of various roof types by training features extracted from both LiDAR data and orthoimages [11]. Pibre et al. presented multiple source data fusion strategies to detection trees [12]. Two different modes were applied to integrate heterogeneous data for training DL models that is a similar idea to [6]. In the study, early fusion and late fusion of the IR aerial images and NDVI and DSM created from LiDAR data were performed. Better results were obtained with early fusion when both NDVI and DSM were used [12, 13].

The recent researches of DL have focused on efforts to improve training performance by utilizing multisource data and/or creating information from the raw data. In the field of geoinformatics and remote sensing, the major data sources are optical and laser sensors. Point cloud data that consists of 3D coordinates can be obtained directly from laser sensors (i.e., LiDAR data) or indirectly by stereo image matching. In addition, derived information from the original raw data are utilized. Examples of the derived information are NDVI and DSM created from multispectral imagery and LiDAR data, respectively. Combining multisource data (e.g., optical and multispectral imagery, point cloud data, and DSM) with derived information from original raw data (e.g., NDVI, cooccurrence features, and surface orientation) provide more reliable results.

As for the DL models, CNN is one of the most extensively used models and has been successfully applied to high-level image processing (or late-vision in computer vision) tasks such as image recognition, scene analysis and description, classification, and medical image computing. Well-known CNN models are LeNet, AlexNet, GoogLeNet, VGGNet, ResNet, Mask R-CNN, SegNet, etc. Most of them are winners of the Image Net Large Scale Visual Recognition Challenge (ILSVCR) [14]. One of the major applications of DL is object classification from images [15, 16]. The goal of this paper is to perform semantic segmentation for land cover classification using CNN based SegNet model. The training datasets are composed of IR images and DSM with derived informa-

tion from the DSM. The derived information includes surface orientation (i.e., slope and aspect of each DSM grid cell) and multidirectional SRIs. DSM is a 3D representation of the terrain surface features, including natural and man-made objects, that are formed with dense point clouds of the 3D coordinates (X , Y , Z). In the field of geoinformatics, the point clouds with 3D coordinates can be obtained directly from laser sensors (i.e., LiDAR data) or indirectly by stereo image matching. Since images are 2D data and provide reflected light from the terrain surfaces, images lack 3D information about the real-world objects. On the other hand, DSM could provide richer 3D geometric information of the objects than the imagery [17].

Feature is a primitive characteristic or attribute of the data. Thus, it is important to extract or derive unique features from the various data, then use those features for DL model training for semantic segmentation. Segmentation entails division or separation of the data into regions of similar characteristics or attributes. Therefore, the purpose of the segmentation is to form meaningful regions by grouping features that have common properties distinct from their neighboring regions [18]. In general, segmentation does not involve classifying each segment but rather subdivides regions without attempting to recognize the individual regions. On the other hand, semantic segmentation or classification involves identification and labeling of the regions or individual objects according to the category [15]. There is no theory or standard method available for the segmentation yet. Most of the existing methods are based on the hoc approach; therefore, DL is expected to be a promising innovative method to solve such a challenging task that requires human intelligence.

The experiments for this paper were carried out using the ISPRS benchmark dataset for terrain feature classification. The land cover of the study site is categorized into six classes; building, tree, low vegetation (grass), impervious surface (road), car, water, and clutter/background in the label data. The main intent of this paper is to classify terrain objects by training the DL model using multisource data; optical IR images, DSM, and DSM-derived data including slope, aspect, and multidirectional SRIs. The experiments were carried out as follows: (1) training with each type of data independently, and (2) training with combining all data. Each data collected from a specific sensor could not convey sufficient information about the real-world. In this aspect, multisource data would be complementary in the training process. The results were analyzed based on evaluation metrics and visual inspection. In conclusion, training by combining multisource data could provide a synergistic effect and multidirectional SRI plays an important role.

2. Materials and Proposed Methods

2.1. Description of Datasets. The ISPRS benchmark dataset [19, 20] was used for training, evaluation, and test of the SegNet model. Table 1 shows the description of the datasets, and Figure 1 presents the configuration of Vaihingen datasets. The datasets consist of IR true orthoimages, DSM, airborne LiDAR data, and label data. True orthoimages are

TABLE 1: Description of ISPRS benchmark dataset for the Vaihingen area.

Data type	GSD (m)	Data collection	Comment
Color IR (IR-R-G) true orthoimage	0.09	Intergraph DMC (altitude: 900 m)	33 patches with georeferenced TIFF format
DSM	0.09	Point clouds created by image matching	16 patches with georeferenced TIFF format
LiDAR data	0.50 (4 pts/m ²)	Leica ALS50 (altitude: 500 m)	10 strips with LAS format
Label data	0.09	Manually digitized ground truth	Land cover classification of 6 classes with TIFF format

FIGURE 1: Vaihingen dataset is composed with IR images, DSM, and label data for training, validation, and test.

geometrically corrected, occlusion-free, and georeferenced images. 60%, 20%, and 20% of the datasets were used as training, evaluation, and test data, respectively. Figure 2 is another dataset from the Potsdam area that was used as “new unseen” data to apply to the trained DL model that was trained with the Vaihingen datasets. Four datasets that are 3_12, 4_10, 4_11, and 7_08 area were selected as new unseen data for the experiments. It is notable to mention that LiDAR data was not used in the experiment; instead, DSM was used to derive slope and aspect and to create multi-

directional SRIs. DSM could be generated using LiDAR data. However, some significant preprocessing is required such as noise removing and interpolation because of the irregularly distributed point clouds with low point density of the LiDAR data. The high-resolution DSM in the ISPRS dataset was created using INPHO MATCH-DSM software with sequential multiimage matching and finite element interpolation [21]. In this regard, we decided that utilizing DSM is more appropriate than LiDAR data to perform the proposed approach.

FIGURE 2: New unseen dataset from the Potsdam area.

2.2. Slope and Aspect. The slope and aspect (i.e., surface orientation) of the surface elements (i.e., DSM grid cells) can be computed from the coordinates of the 3×3 grid window of the DSM (see Figure 3).

Slope and aspect of each surface element is computed by Equations (1)–(10).

$$dX_1 = \frac{Z(X+1, Y+1) - Z(X-1, Y+1)}{2 \cdot \Delta X}, \quad (1)$$

$$dX_2 = \frac{Z(X+1, Y) - Z(X-1, Y)}{2 \cdot \Delta X}, \quad (2)$$

$$dX_3 = \frac{Z(X+1, Y-1) - Z(X-1, Y-1)}{2 \cdot \Delta X}. \quad (3)$$

The average slope in X-direction is

$$S_X = \frac{dX_1 + dX_2 + dX_3}{3}. \quad (4)$$

Slopes in Y-direction are computed by

$$dY_1 = \frac{Z(X-1, Y+1) - Z(X-1, Y-1)}{2 \cdot \Delta Y}, \quad (5)$$

$$dY_2 = \frac{Z(X, Y+1) - Z(X, Y-1)}{2 \cdot \Delta Y}, \quad (6)$$

$$dY_3 = \frac{Z(X+1, Y+1) - Z(X+1, Y-1)}{2 \cdot \Delta Y}. \quad (7)$$

The average slope in Y-direction is

$$S_Y = \frac{dY_1 + dY_2 + dY_3}{3}, \quad (8)$$

where ΔX and ΔY are ground sampling distance (GSD) of DSM in X- and Y-direction, respectively. Finally, slope (S) and aspect (A) at the center of the 3×3 DSM grid window are computed as follows [22]:

$$S = \tan^{-1} \left(\sqrt{S_X^2 + S_Y^2} \right), \quad (9)$$

$$A = \tan^{-1} \left(\frac{S_Y}{S_X} \right). \quad (10)$$

Figure 4 illustrates examples of DSM, slope, and aspect using Vaihingen test data.

2.3. Multidirectional Shaded Relief Images. Shading is an important visual cue (i.e., shape-from-shading) to recognize

FIGURE 3: 3 × 3 grid window of DSM for computing slope and aspect.

(a)

(b)

(c)

FIGURE 4: DSM and derived data. (a) DSM. (b) Slope. (c) Aspect.

the shape of the objects because each surface patch of the objects appears in different brightness depending on the orientation of the surfaces [23–25]. The SRIs can be created using surface orientation (i.e., slope and aspect) of the DSM and virtual light source. The amount of the reflected light from each surface element is recorded in the corresponding pixel of the SRIs and computed by Equation (11).

$$R = \cos(90^\circ - \beta) \cos(S) + \sin(90^\circ - \beta) \sin(S) \cos(\alpha - A), \quad (11)$$

where R is the magnitude of the reflected light, α and β are azimuth and elevation angle of the light source, respectively, and S and A are slope and aspect of the surface element, respectively. Each value of R is converted to an 8-bit image

with range from 0 to 255. The multidirectional SRIs are generated by changing the location of the light source [26]. Four SRIs were generated with light sources of NW, NE, SE, and SW direction (see Figure 5).

2.4. Preparation of Dataset for Training, Evaluation, and Label Data. Each region of the dataset, including training, evaluation, and label data, is partitioned with a 50% overlap along both X and Y directions, and the partitioned tile size is 150×250 as shown in Figure 6. Data partitioning, especially with overlapping, could provide significant benefits: (1) increasing amount of the training data because DL requires a large number of data, and (2) improving DL performance by training the same object (or area) in various situations.

FIGURE 5: Multidirectional shaded relief images created from different locations of the light source.

2.5. SegNet Model. DL using the CNN has been proved to be successful and effective on semantic segmentation (or called semantic labeling) [15]. The SegNet model is a deep fully CNN for multiclass pixel-wise semantic segmentation. SegNet, by Computer Vision and Robotics Group at the University of Cambridge, was designed to be efficient in memory and computational time. It is also significantly smaller in the number of the trainable parameters than other network models. The SegNet is composed of encoder and decoder with symmetrical architecture, and the encoder is based on the convolutional layers from VGG16 that has 13 convolutional and 3 fully-connected layers. VGG16 has been developed by Visual Geometry Group of the University of Oxford and 1st runner-up of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with an error rate of under 10% in 2014 [27].

On the other hand, the fully-connected layers in VGG16 are replaced by convolutional layers in SegNet. The SegNet uses a bottleneck architecture in which the feature maps are upsampled to match the original input resolution, and therefore it is possible to perform pixel-wise segmentation with one-to-one resolution. This is beneficial to label data with the same resolution of the original input data [6]. As with most of the CNN models, the encoder network of the SegNet

performs convolution with a filter bank to produce feature maps, batch normalization to increase the stability of the network, and accelerate training [28]. Then, rectified linear unit (ReLU) is applied as an activation function, and max-pooling with a 2×2 window and stride 2 is performed.

The number of convolutional kernels (i.e., filters) is from 64 to 512 with a size of 3×3 . The decoder network upsamples the feature maps using max-pooling indices from the corresponding encoder feature maps. The feature maps are convolved with a trainable decoder filter bank to produce dense feature maps. Then, batch normalization is applied to each feature map and then fed to the softmax classifier at the output layer of the decoder (see Figure 7). Softmax is used to map the nonnormalized output of the network to a probability distribution over predicted classes [29].

2.6. SegNet Model Training. Partitioned patches (or tiles) of the multisource dataset (i.e., IR orthoimage, DSM, slope, aspect, and multidirectional SRIs) generated by partitioning are fed into the SegNet model for training and evaluation. The numbers of the input patches of each data are 4870, 1730, and 1390 for training, evaluation, and test, respectively. Most of the current DL models require a trial-and-error method to achieve optimal solutions. This approach is not

FIGURE 6: Preparation of training, evaluation, and label dataset by partitioning with 50% overlap.

FIGURE 7: SegNet model architecture. Series of dots represent convolutional layer groups.

trivial and takes ample training time because diverse parameters are involved in the DL model. The combination of different parameter setting leads to tremendous numbers of cases. Therefore, transfer learning was adopted to determine the hyper-parameters including learning rate, mini-batch size, and number of the epochs. Another consideration is normalized DSM (nDSM). The nDSM is obtained by subtracting bare-ground height (i.e., digital terrain model (DTM)) from the DSM as Equation (12):

$$\text{nDSM} = \text{DSM} - \text{DTM}. \quad (12)$$

Many DL models utilize nDSM instead of DSM [6, 9, 30]. The main reason to use nDSM is that the nDSM only reflects relative heights of the objects regardless of the terrain elevation. In other words, differences among objects in elevation are taken into account in nDSM [31]. DTM has to be available to create nDSM (e.g., contours from topographic maps) or complicated filtering process to separate ground and

FIGURE 8: nDSM changes the original shape of object. (a) DSM. (b) nDSM. Z_T and Z_S denote terrain and object height from datum, respectively.

FIGURE 9: Workflow of proposed method for land cover classification.

nonground features (e.g., morphological filtering, progressive densification, surface-based method, and segment-based method) is required [32]. However, the robust DL has to require less preprocessing such as generating nDSM. There might be some controversial issues about using nDSM for object recognition or classification. The ultimate goal of DL is to resemble human intelligence. Human beings are able

to recognize objects regardless of their location in 3D space. Namely, it would be invariant with respect to geometric transformation including shift, rotation, and scale.

nDSM might be a controversial issue in DL. The concept of nDSM is that all objects have to be vertically relocated onto the reference datum before training DL models that utilize DSM. However, nDSM might distort the shape of the objects

TABLE 2: Hyper-parameter values for experiments.

Parameter	Value
Learning rate	0.001
Training cycle (iteration)	334,675
Maximum iteration	608,500
Iteration/epoch	1217/275
Validation frequency	1217
Mini-batch size	4
Shuffle	Every epoch

in some cases. For example, a building with flat roof on a sloped surface is to be a building with a sloped roof if nDSM is applied as shown in Figure 8.

Another issue of the nDSM is for DSM obtained from terrestrial LiDAR data or images of the street scene. If nDSM should be used, then objects have to be moved horizontally to the reference plane. In this case, the horizontal reference plane must be defined. Since DL attempts to resemble human recognition ability, it is much better and robust models that nDSM is not required because human beings recognize objects regardless of objects' location. nDSM should not affect the computation of slope and aspect and creating SRI. In this matter, we did not use nDSM as training but the original DSM. DSM and DSM-derived data (i.e., slope, aspect, and SRI) could provide 3D spatial geometric characteristics that are important information to identify and distinguish different types of objects while images provide only 2D visual information. In this regard, we propose utilizing various data for training the DL model to obtain reliable results. Figure 9 shows the workflow of our proposed method.

2.7. Evaluation Metrics for Performance Measures. Evaluating performance is one of the fundamental tasks in DL. Classification accuracy, in general, can be described as the number of the correct predictions from the total predictions performed. Classification accuracy alone is not sufficient to evaluate the performance of the DL model. Since the ultimate goal of DL is to expand the trained model to other datasets (i.e., new unseen data) that is not involved with training, it is an important issue to evaluate the robustness of a DL model for the new unseen data. The test data has similar characteristics to the training data since the test data, in most cases, belongs to the same region of the training data (i.e., Vaihingen dataset), while the new unseen data has somewhat different characteristics because the new unseen data is selected from a different place (i.e., Potsdam dataset).

Different evaluation criteria have been proposed to assess the quality of the semantic segmentation. Commonly used evaluation metrics for classification are accuracy, precision, recall, F1 score, and intersection-of-union (IoU) [33]. Usually, variations on pixel accuracy and IoU have been used frequently [34]. We applied overall accuracy (Equation (13)) and IoU (Equation (14)) to evaluate semantic segmentation results:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (13)$$

$$\begin{aligned} \text{IoU} &= \frac{\text{Area of overlap}}{\text{Area of union}} \\ &= \frac{(\text{Predicted area}) \cap (\text{Actual area})}{(\text{Predicted area}) \cup (\text{Actual area})} \\ &= \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \end{aligned} \quad (14)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Accuracy might lead to misinterpretation when the class representation is small within the image, as the measure is biased in mainly reporting on how well in identifying negative cases. On the other hand, IoU is calculated for each class separately, then averaged over all classes to provide mean IoU score of semantic segmentation prediction. The criterion to be correct or not to be is 0.5. If IoU is larger than 0.5 (i.e., 50%), it is normally considered a good prediction. IoU has been used in numerous papers and popular object detection challenges such as ImageNet ILSVRC, Microsoft COCO, and PASCAL VOC.

3. Experimental Results and Discussion

Semantic segmentation results of three test datasets from Vaihingen data (i.e., areas I, K, and O in Figure 1) and four new unseen datasets from Potsdam data (i.e., areas 3_12, 4_10, 4_11, and 7_08 in Figure 2) are presented. The major hyper-parameters for experiments were set as shown in Table 2.

Figures 10 and 11 are IR image and label data of Vaihingen and Potsdam datasets, respectively. Figures 12 and 13 show the classification results of Vaihingen and Potsdam datasets, respectively. Results of the Potsdam data were obtained by using trained model with the Vaihingen dataset. It is obvious that the accuracy of the test data is higher than that of the new unseen data.

Classification accuracies of the training, evaluation, test, and new unseen data are listed in Tables 3 and 4 (Note: evaluation metrics are expressed in percent (%)). The result from combining all data (IoU = 84.90%) for the test data was improved compared with the results from individually trained data. However, multidirectional SRI provided better results (IoU = 66.27%) than combining all data (IoU = 52.45%) for new unseen data. We expect that multidirectional SRI has a relatively larger contribution to the model. Even though IoU from combining all data is lower than that from multidirectional SRI, in general, IoU with higher than 0.5 (i.e., 50%) is considered a successful classification. Since man-made features such as buildings have distinctive geometric characteristics compared to natural features, training DL model using DSM-derived data is particularly effective to buildings. Therefore, the proposed approach might be feasible to identify and extract buildings for the further application of 3D building modeling.

The evaluation metrics could not provide information on how well ground truth and predicted classes for individual

FIGURE 10: Vaihingen test area. (a) IR images. (b) Label data (ground truth).

FIGURE 11: Potsdam new unseen data. (a) IR images. (b) Label data (ground truth).

FIGURE 12: Continued.

FIGURE 12: Classification results of Vaihingen test data. (a) IR images. (b) DSM. (c) Slope. (d) Aspect. (e) Multidirectional SRI. (f) Combining all data.

objects matched. In this matter, differences between each result and label data were depicted for visual evaluation. White and black indicate correctly classified and misclassified pixels, respectively (see Figures 14 and 15). Interpolation of the DSM might cause delocalization and zagging effect, especially on the objects' boundaries. Thus, most of the misclassified pixels are found along the boundaries of the objects.

For the Vaihingen test data, some of the low vegetation areas were classified as impervious surfaces (i.e., roads) with DSM and DSM-derived data (i.e., slope, aspect, and SRI). However, training the DL model using all data could improve classification results (see Figure 14(f)). As for the Potsdam new unseen data, more misclassified regions were found except for the case of buildings. Similar to the Vaihingen test

(a)

(b)

(c)

(d)

(e)

FIGURE 13: Continued.

FIGURE 13: Classification results of Potsdam new unseen data. (a) IR images. (b) DSM. (c) Slope. (d) Aspect. (e) Multidirectional SRI. (f) Combining all data.

TABLE 3: Classification evaluation of Vaihingen dataset for training, evaluation, and test.

Dataset	Training accuracy (%)	Evaluation accuracy (%)	Test data (%)			
			Overall Accuracy	IoU	Building Accuracy	IoU
IR image	97.74	91.49	83.71	57.98	89.34	84.08
DSM	91.97	85.25	69.90	41.32	83.88	77.66
Slope	94.42	88.41	67.93	40.31	78.59	71.28
Aspect	95.87	89.07	64.55	35.90	79.01	70.58
SRI	93.50	89.37	70.00	42.59	83.74	75.90
All data	91.36	90.61	83.56	55.75	89.45	84.90

TABLE 4: Classification evaluation of Potsdam dataset for new unseen data.

Dataset	New unseen data (%)			
	Overall Accuracy	IoU	Building Accuracy	IoU
IR image	45.56	23.34	94.24	35.95
DSM	43.86	17.66	82.71	44.89
Slope	41.09	18.11	77.68	37.68
Aspect	32.97	11.68	61.80	34.24
SRI	52.53	28.28	77.92	66.27
All data	58.86	30.75	89.37	52.45

data, a training model using a combination of all data provided better results (see Figure 15(f)). Test data might be insufficient to evaluate the performance of the trained DL models particularly to investigate the possibility of universal use. For this reason, we applied not only test data but also new unseen data to the trained model; as a result, some meaningful results were confirmed. Utilizing multisource data and derived information that are appropriate to represent characteristics of the objects could improve semantic segmentation. DSM and DSM-derived information could play an important role in recognizing buildings.

Classification results from training all data (i.e., IR image, DSM, slope, aspect, and SRI) are better than results from training individual data. The experiments show that DSM and its derived information could be appropriate to recognize man-made objects including buildings because DSM

and DSM-derived information have explicit geometric characteristics. In consequence, the classification accuracy of the building is much higher than that of other objects. Figures 16–18 show buildings of the label data and corresponding classification results from using all data of Vaihingen and Potsdam areas, respectively. It is noticeable that buildings in the test data were well identified, while buildings in the new unseen data were less accurately identified. A considerable number of cars of the Potsdam data were identified as buildings; however, most of the major buildings are correctly classified.

Some conventional image processing technique (e.g., morphological filtering with erosion and dilation [35]) might improve the results. Figure 18 illustrates an example of applying morphological filtering to the building class from Potsdam data. Therefore, it is recommendable to consider postprocessing if necessary.

4. Conclusions

Human beings have the ability to integrate various visual cues (i.e., shape, size, color, tone, texture, depth, stereopsis, and motion) based on knowledge, experience, and innate cognition for visual perception. Semantic segmentation or classification involves object detection, recognition, and identification by utilizing various information. In consequence, it is not sufficient to successfully carry out the semantic segmentation of real-world objects by one type of sensory data. Every object has its own unique physical and geometrical characteristics. Such intrinsic characteristics could be obtained partially from imagery, 3D data, and derived

FIGURE 14: Continued.

FIGURE 14: Differences between classification results and label data of Vaihingen test data. (a) IR image. (b) DSM. (c) Slope. (d) Aspect. (e) Multidirectional SRI. (f) Combining all data.

information from various sources. Therefore, one of the key tasks is to utilize multisource data for complementary effects to extract characteristics of the objects from the data.

In this paper, we proposed a DL model that utilizes multisource data including optical IR image, DSM, and DSM-derived slope, aspect, and multidirectional SRIs for semantic

FIGURE 15: Continued.

FIGURE 15: Differences between classification results and label data of Potsdam new unseen data. (a) IR image. (b) DSM. (c) Slope. (d) Aspect. (e) Multidirectional SRI. (f) Combining all data.

FIGURE 16: Building class of Vaihingen test data. (a) Label data. (b) Results from combining all data.

segmentation to classify land cover features. The dataset used for training the DL model are based on the DSM that has distinct 3D geometric characteristics. Therefore, specific objects (i.e., buildings) are identified well dominantly. Nevertheless, overall results from the new unseen data were not quite satisfied; buildings were adequately identified compared to the other objects due to the DSM

with derived information. Each object has its own unique characteristics with various aspects. The key task of DL is to reveal representative features from the objects during training. Therefore, training by utilizing various types of data that are suitable to specific objects could improve the performance of DL. In particular, DSM-derived data would be helpful to identify buildings.

FIGURE 17: Building class of Potsdam new unseen data. (a) Label data. (b) Results from combining all data.

FIGURE 18: Comparison of (a) classification result, (b) morphological filtering, and (c) label data of Potsdam new unseen data.

In general, the scalability of the trained DL models is evaluated using test data. The problem of using the test data is that, in most cases, the test data belongs to the same area with similar properties as the training data. Therefore, we evaluated the model using a dataset from different areas, i.e., test data from the Vaihingen area, and new unseen data from the Potsdam area. We found some meaningful results from the test data and new unseen data. Combining all data yielded the highest IoU of the building for the test data (i.e., 84.90%). On the other hand, multidirectional SRI provided the highest IoU of the building (i.e., 66.27%) while combining all data yielded IoU of 52.45%, but it was higher than individually trained data. We expect that multidirectional SRI plays an important role for training DL model.

The ultimate goal of DL is to expand trained models for universal use. However, the major problem of DL is data dependency. Geospatial data has a variety of regional properties; hence, it is a valuable task to provide pretrained models suitable for the specific regions (e.g., urban, residential, mountainous, and agricultural areas). In addition, flexible transfer learning that could avoid training DL models from scratch would make DL a more powerful tool in various fields. In addition, the integration strategy of various combinations of the multisource dataset might maximize the synergistic effect, for example, applying different priorities or weights for each dataset.

Data Availability

The datasets generated during the current study are available from the corresponding author on reasonable request. Vaihingen datasets are available in <http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html>.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1D1A1B07048732). The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [21]: <http://www.ifp.uni-stuttgart.de/dgpf/DKEPAllg.html>.

References

- [1] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943.
- [2] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: a survey," 2020, <https://arxiv.org/abs/2001.05566>.
- [3] D. Turcsány, *Deep Learning Models of Biological Visual Information Processing*, Doctor of Philosophy theses, The University of Nottingham, 2016.
- [4] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: a meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, 2019.
- [5] E. Maltezos, A. Doulamis, N. Doulamis, and C. Ioannidis, "Building extraction from LiDAR data applying deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 1, pp. 155–159, 2019.
- [6] N. Audebert, B. le Saux, and S. Lefèvre, "Beyond RGB: very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 20–32, 2018.
- [7] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.
- [8] F. H. Nahhas, H. Z. M. Shafri, M. I. Sameen, B. Pradhan, and S. Mansor, "Deep learning approach for building detection using LiDAR - orthophoto fusion," *Journal of Sensors*, vol. 2018, Article ID 7212307, 12 pages, 2018.
- [9] K. Zhou, D. Ming, X. Lv, J. Fang, and M. Wang, "CNN-based land cover classification combining stratified segmentation and fusion of point cloud and very high-spatial resolution remote sensing image data," *Remote Sensing*, vol. 11, no. 17, p. 2065, 2019.
- [10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [11] F. Alidoost and H. Arefi, "Knowledge based 3D building model recognition using convolutional neural networks from LIDAR and aerial imageries," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLI-B3, pp. 833–840, 2016.
- [12] L. Pibre, M. Chaumont, G. Subsol, D. Ienco, and M. Derras, "How to deal with multi-source data for tree detection based on deep learning," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1150–1154, Montreal, QC, Canada, November 2017.
- [13] P. Bodani, K. Shreshtha, and S. Sharma, "OrthoSeg: a deep multimodal CONVOLUTIONAL neural network architecture for semantic segmentation of orthoimagery," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-5, pp. 621–628, 2018.
- [14] O. Russakovsky, J. Deng, H. Su et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, Boston, MA, USA, June 2015.
- [16] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, pp. 1520–1528, Santiago, Chile, December 2015.
- [17] D. Lee, E. Cho, and D. C. Lee, "Semantic classification of DSM using convolutional neural network based deep learning," *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, vol. 37, no. 6, pp. 435–444, 2019.
- [18] D. C. Lee, *An Adaptive Texture Segmentation Approach for Applications in Digital Photogrammetry*, Ph.D. Dissertation, The Ohio State University, 1997.

- [19] F. Rottensteiner, G. Sohn, J. Jung et al., "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. I-3, pp. 293–298, 2012.
- [20] M. Cramer, "The DGPF-test on digital airborne camera evaluation - overview and test design," *Photogrammetrie - Fernerkundung - Geoinformation*, vol. 2010, no. 2, pp. 73–82, 2010.
- [21] C. Lemaire, "Aspects of the DSM production with high resolution images," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 37, Part B4, pp. 1143–1146, 2008.
- [22] P. A. Burrough, *Principles of geographical information systems for land resources assessment*, Clarendon Press, 1988.
- [23] B. K. P. Horn, "Hill shading and the reflectance map," *Proceedings of the IEEE*, vol. 69, no. 1, pp. 14–47, 1981.
- [24] V. Bruce and P. R. Green, *Visual perception: physiology, psychology and ecology*, Lawrence Erlbaum Associates, 1990.
- [25] R. Szeliski, *Computer vision: algorithms and applications*, Springer-Verlag, 2011.
- [26] D. C. Lee, D. H. Lee, and D. G. Lee, "Determination of building model key points using multidirectional shaded relief images generated from airborne LiDAR data," *Journal of Sensors*, vol. 2019, Article ID 2985014, 19 pages, 2019.
- [27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, <https://arxiv.org/abs/1409.1556>.
- [28] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, Lille, France, 2015.
- [29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," 2016, <https://arxiv.org/abs/1511.00561v3>.
- [30] M. Gerke, T. Speldekamp, C. Fries, and C. Gevaert, "Automatic semantic labelling of urban areas using a rule-based approach and realized with MeVisLab," 2015.
- [31] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [32] C. M. Gevaert, C. Persello, F. Nex, and G. Vosselman, "A deep learning approach to DTM extraction from imagery using rule-based training labels," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 142, pp. 106–123, 2018.
- [33] B. Benjdira, A. Ammar, A. Koubaa, and K. Ouni, "Data-efficient domain adaptation for semantic segmentation of aerial imagery using generative adversarial networks," *Applied Sciences*, vol. 10, no. 3, article 1092, 2020.
- [34] A. Garcia-Garcia, S. Orts-Escolano, S. O. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, <https://arxiv.org/abs/1704.06857v1>.
- [35] A. Hornberg, *Handbook of machine vision*, Wiley-Verlag, 2006.