Boundary Enhancement Semantic Segmentation for Building Extraction from Remote Sensed Image

Hoin Jung, Han-Soo Choi, Myungjoo Kang

***Abstract*—Image processing via convolutional neural network(CNN) has been developed rapidly for remote sensing technology. Moreover, techniques for accurately extracting building footprints from remote sensed images have attracted considerable interest owing to their wide variety of common applications, including monitoring natural disasters and urban development. Extraction of building footprints can be performed easily by semantic segmentation using U-Net-like CNN architectures. However, obtaining precise boundaries of segmentation masks remains challenging due to various impediments surrounding target objects. In this study, we propose a method to elaborate edges of buildings detected in remote sensed images to enhance the boundaries of segmentation masks. The proposed method adopts *holistically-nested edge detection(HED)*, which extracts edge features at an encoder of a given architecture. In the proposed *boundary enhancement(BE) module*, an extracted edge and segmentation mask are combined, sharing mutual information. To enable the proposed method efficiently to adapt to a wide variety of conditions, we design a distinctive approach adopting a HED-unit and BE module, which is applicable to various semantic segmentation networks containing encoder-decoder structures. Experiments were conducted on five different datasets (DeepGlobe, Urban3D, WHU(HR, LR), and Massachusetts). The results demonstrate that our proposed approaches improved on the performance of prior methods for extracting building footprints. Comparative experiments were conducted on various backbone architectures including U-Net, ResUNet++, TernausNet, and USPP to ensure the effectiveness of the proposed method. Based on various evaluation metrics and qualitative analysis, our results show that the proposed method achieved improved performance compared to prior methods for all datasets and backbone networks.**

***Index Terms*—Convolutional neural network (CNN), satellite imagery, remote sensing, semantic segmentation, building footprint extraction, boundary enhancement.**

I. Introduction

VARIOUS information from remote sensed images is widely used in the fields of environment[1, 2], and ocean sciences[3] as well as geology[4], forestry[5, 6], agriculture[7], and meteorology[8]. Satellite image and aerial image processing, called remote sensing, refers to any activity that grasps the characteristics of objects on the ground without physical contact or exploration. To this end, various studies have been conducted to obtain and analyze information on land, environments, and resources by using multi-spectral sensors[9] or optical cameras installed on the ground, in aircraft[10], and in satellites.

This remote sensing technology is advantageous in terms of the time necessary to acquire data. Thus, in well-developed remote sensing systems, the amount of effort required to obtain data becomes much lower than that of ground measurement by conventional land surveying techniques[11]. In addition, such technologies enable the acquisition of data from a wide range of areas, including remote or inhospitable areas that humans may not be able to travel. Therefore, image processing techniques using high-resolution remote sensed images[12, 13] are needed, enabled by the fact that it has become possible in recent years to secure large amounts of such data[14]. Furthermore, the accumulated database on remote sensed images enables predicting future states of a target area by modeling, analyzing, and monitoring objects almost in real-time[15] by utilizing both current and historical information.

Recently, remote sensing technology based on image processing by deep learning has been rapidly developed. Several deep-learning-based methodologies in the remote sensing area have been suggested to analyze various types of data, including hyperspectral images(HSI), optical images, LiDAR data, and integrated multi-sensor data[16, 17]. Above all, the convolutional neural networks (CNNs) methodology has proven effective in classifying and segmenting each pixel in a given image into a semantic label[18]. Among the many object segmentation and landcover classification tasks in remote sensing, building extraction by binary semantic segmentation methods[19–27] is one of the most fundamental challenges because of its extensive practical application. The capability to obtain accurate and immediate building footprint information has significant benefits in monitoring urban development or detecting natural disasters[28].

For these reasons, numerous CNN architectures have been proposed to extract building footprints from the remote sensed image of the target areas. However, some extracted buildings still tend to have inaccurate edges because the encoder-decoder networks downsample the input image as a compressed feature map, and upsample the feature map again as the same size as input images to classify objects pixel-wise as true or false. In this procedure, there exists a possibility that boundary information might be lost. In addition, building boundaries can be easily disturbed by other objects such as shadows, trees, or various objects near a building[29]. As a result, the classified pixels near the actual boundaries become fragmented and anfractuous because of their ambiguity.

This phenomenon usually appears in encoder-decoder structures like the U-Net[30] CNN architecture, which is known as one of the most fundamental network in semantic segmentation. Because most U-Net-like networks include skip connections to preserve comprehensive semantic information of input images, such networks are not sufficiently deep to detect detailed edge features. Therefore, the proposed method aims to straighten the uncertain edge effectively while improving the semantic segmentation performance. Furthermore, this approach can be attached to any architecture with a U-Net-like structure, regardless of the depth of the backbone network, as described in the following steps.

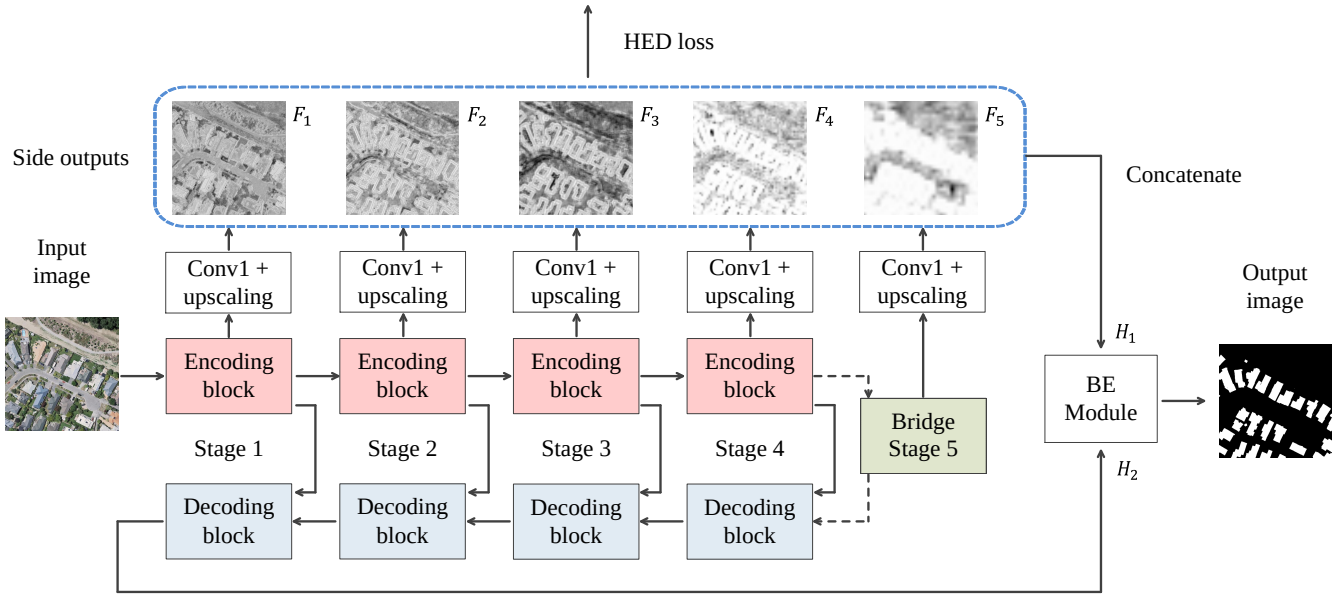

Fig. 1: HED-unit is attached to a given backbone's encoder. Side outputs are extracted from the last feature maps in each stage being stretched as the same size as the input shape. The side outputs become an auxiliary prediction to guide the network to detect edges of the target object as being corrected by HED loss, L_{HED}. Consequently, the concatenated side outputs will be input to BE module. Although this figure is an example based on U-Net, the HED-unit can be applied to any backbone containing encoder-decoder architecture.

First, the edge features of detected buildings are extracted by a *holistically-nested edge detection (HED)*[31] unit attached to a given backbone encoder. The edge features are then used as inputs for the proposed *boundary enhancement (BE) module*. This BE module includes two parallel sub-units detecting a boundary and segmentation mask, respectively. Each sub-unit shares information by a *combined probability map* to represent the likelihood of a target object. Finally, the segmentation mask feature map passes through a *positive replacement* to enhance its boundary.

In this study, experiments for various backbone networks and datasets were conducted to estimate the performance enhancement of the proposed HED-unit and BE module over prior methods. U-Net[30], ResUNet++[32], TernausNet[19], and USPP[33] were used as backbone networks in the proposed method. The DeepGlobe[34], Urban3D[35], WHU high-resolution(HR) datasets[36], WHU low-resolution(LR) datasets[36], and Massachusetts Buildings Datasets[37] were used in the experiments because of their widespread use in this field.

Consequently, by adopting the proposed approach, the extracted building results can have crisp and delicate boundaries compared with existing semantic segmentation methods. In addition, the proposed HED-unit and BE module can be used widely with various U-Net-like backbones without a significant increase in network parameters or computational cost.

The rest part of this paper is organized as follows. In Section II, we introduce several studies related to semantic segmentation and building footprint extraction using remote sensed imagery. In Section III, the process and characteristics of our proposed method are introduced in detail, including the HED-unit and BE module. The designed combination of a loss function and a training strategy is also explained. In Section IV, the feasibility of our proposed method is verified by the results of several experiments. In Section V, we present our final conclusions.

II. Related Work

Recently, the semantic segmentation by CNNs has rapidly developed, showing dramatic increases in capability. A number of semantic segmentation methodologies are adopting a deep encoder-decoder architecture achieving end-to-end and pixel-to-pixel network[30]. Also, several U-Net-like architectures have been proposed for remote sensed image, especially for building footprint extraction. TernausNet[19], which is widely used for building extraction, adopted the VGG11[38] pretrained model as an encoder block to increase its pixel-wise classification ability. [33] proposed a bridge module between an encoder and decoder, U-shape spatial pyramid pooling (USPP), to learn a multiple spatial scales and global contextual information. MC-FCN[21] used a multi-constraint FCN[39] structure employing subsampled ground truth. ResUNet-a[22] included a residual block structure similar to ResUNet++[32], considering more detailed information on distance transform and color space via a multi-task layer. Residual refine module was proposed by BRR-Net[40], containing a large receptive field and deeper layer to refine the input image and produce accurate building extraction.

On the other hand, few non-U-Net-like backbone networks for building extraction have been developed, such as MAP-Net[23] and DE-Net[24]. In MAP-Net, a multi-parallel path

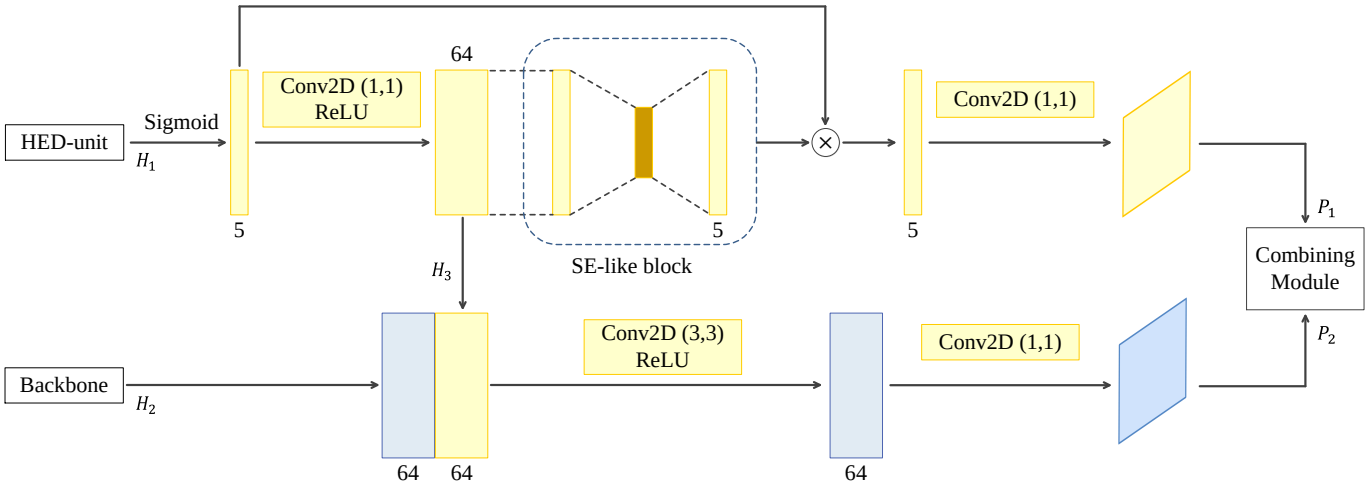

Fig. 2: The figure shows the structure of BE module. H_1 is the concatenated output of the HED-unit containing boundary information. H_2 is the segmentation mask produced by the backbone network. Both H_1 and H_2 are adopted to sub-units of BE module in parallel. Two parallel sub-units extract and enhance both boundary and segmentation masks, producing each probability map P_1 and P_2. The two probability maps will be integrated by *combining probability map & positive replacement.*

and an attention module were used to preserve the spatial localization of multi-scale features. DE-Net developed an upsampling method called densely upsampling convolution(DUC) to recover high-resolution prediction maps. Nevertheless, U-Net-like architectures remain a major methodology for detecting semantic building footprint labels.

Meanwhile, some studies to utilize edge information for semantic segmentation[41], salient object detection[42], and building footprints extraction[25–27] have been conducted. A multi-scale aggregation fully convolutional neural network (MA-FCN)[25] regularized the polygonized irregular boundaries of a segmentation mask using the Douglas-Peucker algorithm for postprocessing. BR-Net[26] proposed a two-way ground truth strategy, combining segmentation loss and boundary loss. In addition, to obtain more detailed information on the boundaries of buildings, [27] proposed a boundary loss based on differentiable surrogate metric. However, many attempts to extract edge information from segmentation network are operated on decoder and tail parts, while encoder-based feature extractions are also effective[31, 42]. To simplify our proposed system and reduce its computational cost, we adopted HED to extract edges by inserting it into a backbone network.

III. Proposed Method

In this section, we propose a method to enhance the performance of semantic segmentation in remote sensed images by obtaining crisp and straightened edges of extracted buildings. The proposed method involves two main steps. The first is a HED-unit, inspired by the fundamental edge detector, HED[31], attached to a backbone encoder. Using this approach, the backbone network is enabled to extract both edges and segmentation masks. The second approach is a boundary enhancement(BE) module to support the ability to detect building footprints by combining the extracted edge and segmentation mask from the HED-unit and backbone, respectively. A detailed description of this process is presented as follows.

A. HED-Unit

Fig. 1 shows the first part of the proposed boundary enhancement approach, the HED-unit, attached to each stage of the encoder part in a backbone network. The last feature map of each stage in the encoder is extracted and compressed into a 1-channel feature map by a 1×1 convolution layer and then upsampled bilinearly with the same size as the input image. We denote these results as $F_i, i \in \{1, 2, 3, 4, 5\}$ in Fig. 1.

The stretched side outputs from each stage are utilized in two ways. First, the side outputs are connected with the loss function, *HED loss*, based on the edges of the ground truth. They are considered as auxiliary outputs to guide the encoder to detect the boundaries of the target objects. Second, the side outputs are concatenated as H_1 to be used as an input to the BE module, containing a holistically nested feature map as shown in equation (1)

$$H_1 = \text{concat}\left([F_1, F_2, \cdots, F_N]\right) \tag{1}$$

where N is the number of the stage the architectures have. $N = 5$ is set in Fig. 1 since the exemplified architecture, U-Net, has five stages structure. Using this procedure, the BE module can include sufficient information on a target's boundary to encourage the tail part of the entire network to produce a smooth segmented boundary.

B. Boundary Enhancement(BE) Module

The proposed boundary enhancement module is shown in Fig. 2. To control both information about the boundary and segmentation masks of target objects, the BE module includes two-way input, H_1, H_2, which is produced by the previous network, as shown in Fig. 1. H_1 is the concatenated edge output from the HED-unit, which contains abundant boundary

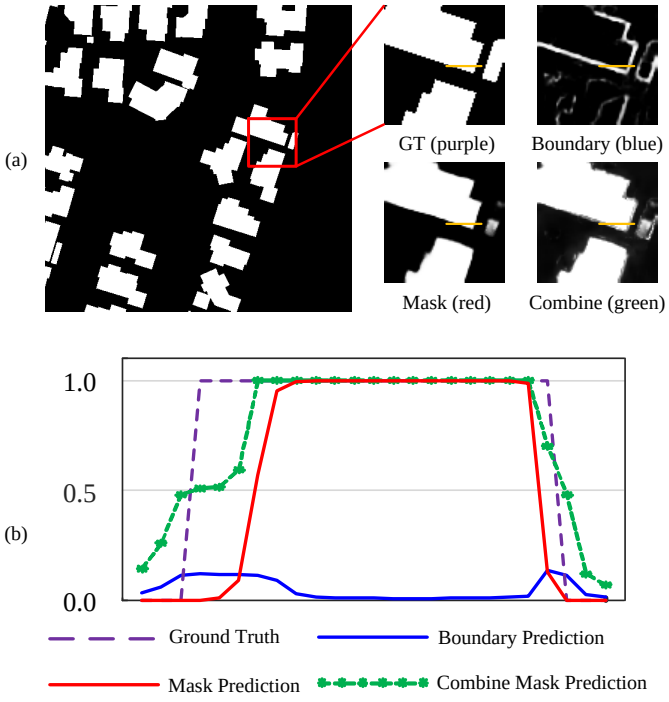

Fig. 3: Fig. (a) shows the predicted result of probability of boundary and mask exist. The predicted mask is not sufficiently covering the ground truth, and the predicted boundary can supplement the deficient area. The combining method is shown in Fig. (b), merging two probability maps for segmentation mask($P_m(x_m)$) and boundaries($P_b(x_b)$) as followed the expressed yellow lines in Fig. (a). Missed pixel prediction of mask or irregular boundary of buildings can be supplemented by the combined probability map($P_s(x)$).

information. The other input, H_2, is the last feature map of the backbone network containing information of the segmentation mask with not sufficiently accurate boundary footprints. The two inputs from the HED-unit and backbone share their information to strengthen their feature prediction.

The BE module has a parallel sub-unit structure responsible for detecting boundary and segmentation masks, respectively. The boundary detection sub-unit includes a component similar to the SE block[43] to boost the representative power of the boundary feature map. The boundary information H_1 can be reinforced by calibrating the interdependency response between channels. The output of the SE-like block is be compressed into 1-channel mask by a 1×1 convolutional layer, while the last feature map is used as a probability map P_1.

On the other hand, the segmentation mask detection sub-unit is the principal prediction unit of the entire network, which is used as the final segmentation output. This sub-unit receives two inputs, H_2 and H_3, from the backbone network and boundary detection sub-unit. Sufficient boundary information is shared by H_3 from the boundary detection sub-unit to the segmentation mask detection sub-unit. Moreover, H_2 and H_3 are concatenated to utilize both information sources on the

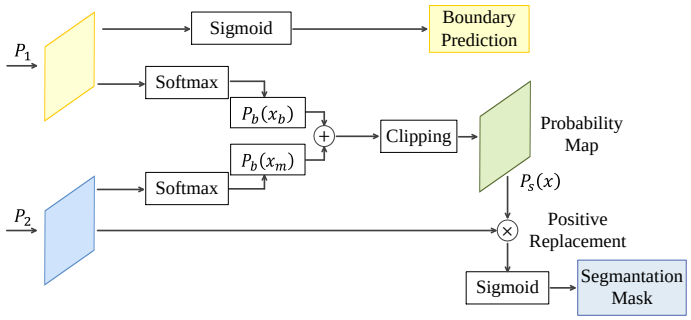

Fig. 4: Combining module consists of probability map & positive replacement. The probability map enhances the edge of segmentation mask directly, and the combined result will replace the original segmentation mask maintaining its scale by positive replacement.

segmentation mask and boundary. By passing through several convolutional layers, the sub-unit produces a probability map P_2.

Finally, the segmentation mask feature map is enhanced to contain precise boundary prediction by the combining module, consisting of *combining probability map* and *positive replacement*.

1) Combining Probability Map : Each parallel sub-unit in the BE module produces 1-channel feature maps, P_1 and P_2, for both the boundary and segmentation masks. P_1 and P_2 become probability maps, $P_b(x_b)$ and $P_m(x_m)$, individually by a softmax activation function, indicating the likelihood of the target object existing in a given pixel. Simultaneously, P_1 is regarded as the boundary prediction to calculate the loss function with the boundary ground truth passing through the sigmoid activation function.

Each probability map, $P_b(x_b)$ and $P_m(x_m)$, is combined by the pixel-wise addition of the probability values, as shown in Fig. 3. When the added probability values exceed 1.0 because of the overlapped prediction of the boundary and segmentation mask, the results were considered to be simply 1.0, as shown in equation (2). In the inference step, the pixel-wise addition is conducted on five times amplified boundary probability map, $5P_b$. Although the predicted segmentation mask could not sufficiently cover the footprint representation, the proposed method makes it possible to suggest a precise boundary to the deficient area of segmentation mask prediction.

$$P_s(x) = \min\left(P_b(x_b) + P_m(x_m), 1\right) \tag{2}$$

2) Positive replacement: The proposed combined probability map produces probability values [0,1]. However, the pixel signal values of the raw segmentation masks, P_2, are allocated in a broad range between $(-\infty, \infty)$, before the softmax function is applied.

To obtain the refined boundary while retaining the scale of the response, P_2 is reused by employing the provided probability maps $P_s(x)$, as shown in Fig. 4 and equations (3) and (4).

$$x_{output} = x_m - \sigma(x_m) + P_s(x)\lambda(x_m) \tag{3}$$

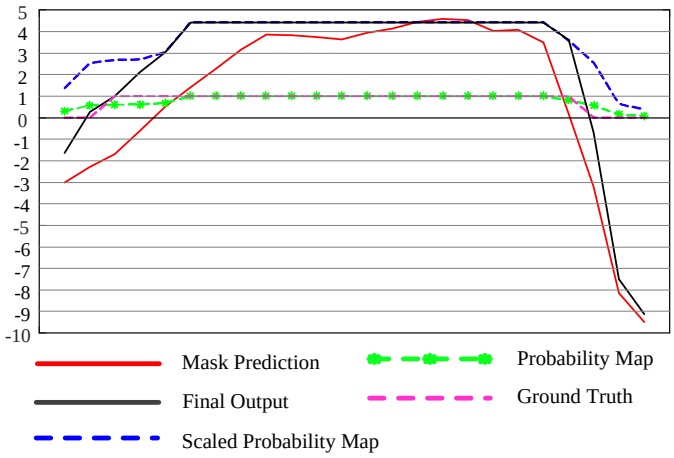

Fig. 5: The raw signal values of the mask prediction(x_m) are modified by the combined probability map($P_s(x)$). Positive signals($\sigma(x_m)$) in x_m are removed and replaced with the scaled combined probability map($P_s(x)\lambda(x_m)$) while the negative signals are supplemented. The adjusted prediction, final output(x_{output}), shows improved results fitting well with the ground truth.

$$\lambda(x_m) = \frac{1}{N_{\text{nonzero}}} \sum_{k=0}^{N-1} \sigma(x_m)_k \tag{4}$$

where x_m is signal values of P_2 and $\sigma(x)$ is an activation function. In these experiments, the activation function is set to a rectified linear unit(ReLU): i.e. $\sigma(x) = \max(0, x)$.

In equation (3) and (4), the positive parts of responses in P_2, $\sigma(x_m)$, are removed while the negative ones are maintained: $x_m - \sigma(x_m)$. Instead, the scaled probability map $P_s(x)\lambda(x_m)$ replaces the positive parts of the x_m as shown in equation (3), where the scale factor, $\lambda(x_m)$, is the average value obtained from $\sigma(x_m)$, the positive parts of P_2. The renewed probability map, $P_s(x)\lambda(x_m)$, has sufficient information about both boundary and segmentation mask while maintaining the scale of original response of P_2. Finally, the final output prediction mask x_{output} can be obtained representing boundary enhanced segmentation mask, as shown in Fig. 5. x_{output} passes through the sigmoid activation function to compute the loss function with the ground truth of the segmentation mask.

C. Loss Functions

The entire network should effectively detect both boundary and segmentation masks to acquire a well-regularized segmentation mask. For this reason, three different loss functions were designed for each output, situated separately as follows. The total loss will be discussed in training strategy.

$$L_{total} = f(L_{HED}, L_{boundary}, L_{mask}) \tag{5}$$

1) Boundary Loss: To obtain errors related to detecting the edges of objects, the binary cross-entropy(BCE)[44] loss is applied to compute the loss between predictions and ground truth for a given boundary. In the HED-unit, multi-level loss is acquired by calculating the loss functions for each side output F_k, originating in the encoder of the backbone network; $L_k = \text{BCE}(y_{\text{boundary}}, F_k)$, $F_k \in \{1, 2, \cdots, N_{stage}\}$, where y_{boundary} is ground truth of building's boundary. The summation of each loss function becomes L_{HED}, in equation (6).

$$L_{HED} = \sum_{k=1}^{N_{\text{stages}}} L_k \tag{6}$$

where N_{stages} means the number of encoding stage of the backbone networks. Moreover, $L_{boundary}$ is obtained by calculating the loss function between the ground truth and the P_1, output of the boundary detection sub-unit in the BE module: $L_{boundary} = \text{BCE}(y_{\text{boundary}}, \hat{y}_{\text{boundary}})$, where $\hat{y}_{\text{boundary}}$ is predicted boundary from P_1,

2) Segmentation Mask Loss: The segmentation mask loss is obtained by combining the focal loss[45] and MS-SSIM loss[46], in equation (7).

$$L_{mask} = L_{focal} + L_{MS-SSIM}. \tag{7}$$

Because the proposed method adopts the *combined probability map* and *positive replacement* to accurately detect both boundary and segmentation masks, structural imbalance might occur in prediction near the edge of a target. Focal loss can solve this problem by weighting the negative term in the cross-entropy loss with $\gamma = 2$. Furthermore, to preserve the structural characteristics of the target, MS-SSIM loss considers contrast, luminance, and structure at various scales to prevent the segmentation mask from being fragmented.

The MS-SSIM loss can be obtained by the following equation.

$$L_{MS-SSIM} = 1 - \prod_{j=1}^{M} \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \right)^{\beta_j} \left(\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \right)^{\gamma_j}$$

for $\mathbf{x} = \{x_i \mid i = 1, 2, \cdots, M\}$ and $\mathbf{y} = \{y_i \mid i = 1, 2, \cdots, M\}$, where M is the total number of scale layers, and β_j and γ_j are parameters defining the relative importance of the components. Let μ_x, σ_x^2, and σ_{xy} be the mean of x, the variance of x, and the covariance of x and y, respectively, between the ground truth and prediction. The small constants $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$ are added to prevent the equation from being divided by zero, where L is the dynamic range of the pixel value, 1.0. In our experiments, we set $M = 5$, $K_1 = 0.1$, $K_2 = 0.3$, and $\beta_j = \gamma_j = [0.0448, 0.2856, 0.3001, 0.2363, 0.1333]$ as denoted in [46].

D. Training strategy

The main purpose of our network is ultimately to obtain a high-quality segmentation mask. In equation (5), the segmentation mask loss has equal importance with the boundary loss. In addition, the value of the boundary sub-unit loss in the BE module, $L_{boundary}$, is similar to the individual loss, L_k, in the HED-unit. This situation is not suitable because if the importance of L_{HED} is N_{stages} times larger than $L_{boundary}$ and L_{mask}, the total loss drives the entire network's emphasis

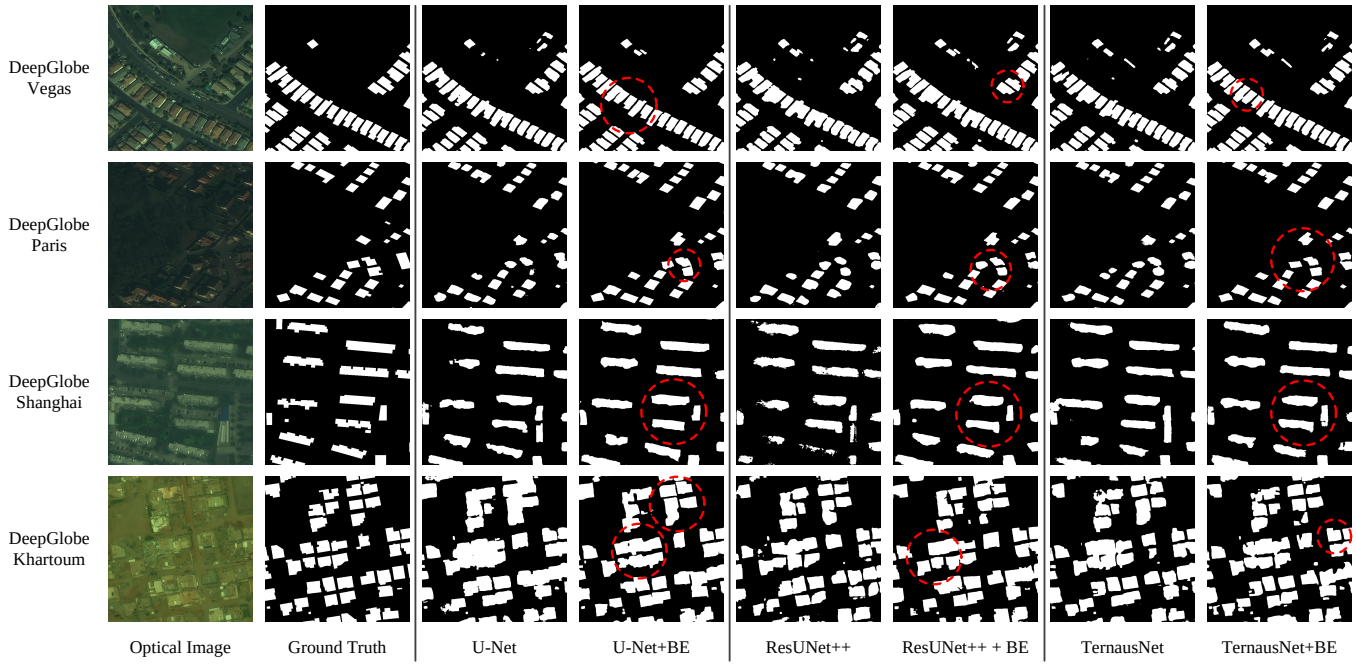

Fig. 6: Experimental result of DeepGlobe dataset for each area of interest and backbone networks. The boundaries of the building footprint are straightened and manifested, while adjacent buildings are distinguished well.

on edge detection rather than segmentation mask.

We designed the training procedure as two steps to deal with this conflict by supposing that each step has an intrinsic function. In the first few steps, the total loss weights in the HED loss to guide the network to learn multi-scale boundary features in the encoding section, as mentioned in equation (8). All individual loss functions have equal importance to the HED loss L_k of each stage.

Second, the total loss becomes having more weight in L_{mask}, as shown in equation (9). The entire HED loss of each stage is regarded as a single loss value that has the same weight as L_{focal} and $L_{MS-SSIM}$. Moreover, $L_{boundary}$ has more weight because the output of the boundary detection subunit is connected with the possibility map and affects the final segmentation mask output. In this study, equation (8) is applied for the first two epochs, and equation (9) is the dominant total loss function. Here N is the number of stage in backbone network.

$$L_{total} = \begin{cases} L_{HED} + L_{boundary} + \dfrac{L_{mask}}{2} & (8) \\ \dfrac{1}{N} \times L_{HED} + L_{boundary} \times N + L_{mask} & (9) \end{cases}$$

IV. Experimental Results

A. Datasets

In this study, extensive experiments were conducted to evaluate the proposed method for five open datasets, including DeepGlobe Dataset[34], Urban3D Challenge Dataset[35], WHU Building Dataset(HR, LR)[36], and Massachusetts Buildings Dataset[37]. In these experiments, all the training and validation subsets for each datasets were merged as a single dataset because some provided datasets were divided into train/validation/test subsets, whereas others comprised integrated data.

The DeepGlobe dataset, which was used in the DeepGlobe Building Detection Challenge and SpaceNet Building Detection Challenge, covers $3,011km^2$ of urban and suburban land area, including four regions: Las Vegas, Paris, Shanghai, and Khartoum. The source images came from the WorldView-3 satellite sensor, which produces RGB and 8-band multispectral data containing $302,701$ building footprints with a 30 cm ground sample distance(GSD). The dataset contains 650×650 pixel image size for all areas of interest with $9,004$ and $1,589$ images as the training set and testing set, respectively.

The USSOCOM Urban3D Challenge dataset was also produced by the WorldView-3 satellite containing 2D orthorectified RGB images and 3D digital surface models with 50 cm GSD. The source images cover over $360km$ of terrain and contain roughly $157,000$ building footprints with a 2048×2048 resolution image. In these experiments, the high-resolution images are splitted into 512×512 pixel images with 2,912 and 672 images as the training and testing sets, respectively.

The WHU building dataset includes two different subdatasets, aerial and satellite subsets. The aerial subset, called WHU high-resolution(HR) dataset in this paper, covers $450km^2$ of Christchurch, New Zealand, with 30 cm GSD, containing more than 187,000 building footprints. The source images were provided with a 512×512 pixel size with 5,772 training sets and 2,416 testing sets.

The satellite subset, named WHU low-resolution(LR) subset in this paper, consists of satellite images covering $550km^2$ of

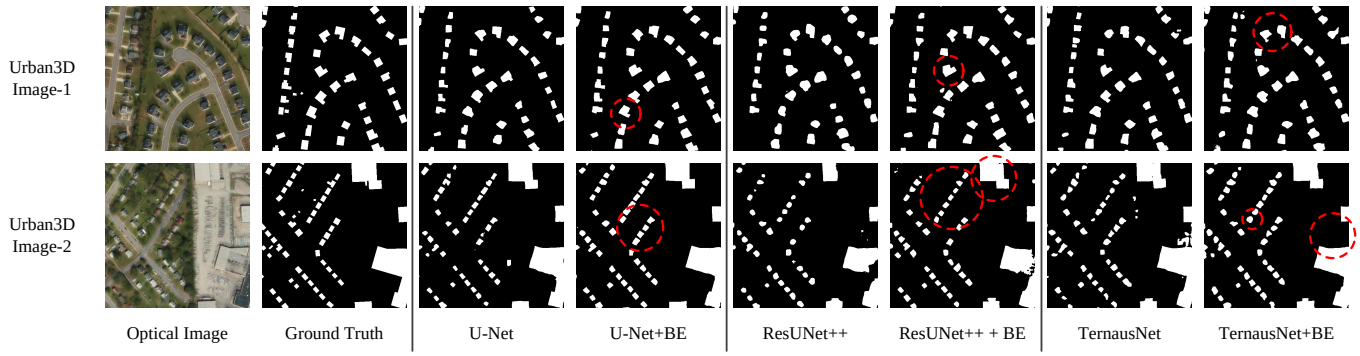

Fig. 7: Experimental result of Urban 3D Dataset for each backbone networks. The boundaries of building footprint are straightened and manifested, while adjacent buildings are distinguished well.

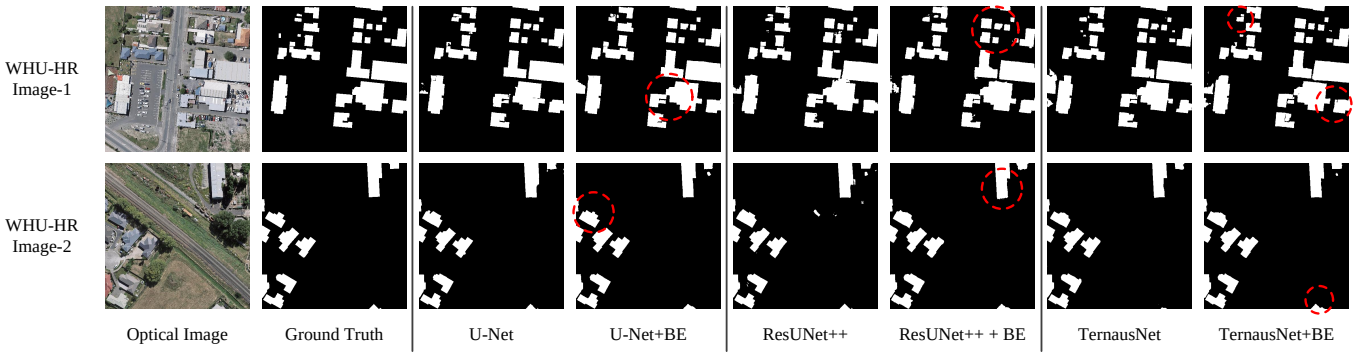

Fig. 8: Experimental result of WHU-HR Dataset for each backbone networks. The boundaries of building footprint are straightened and manifested, while adjacent buildings are distinguished well.

East Asia with 2.7m GSD. The LR subset includes $17,388$ tiles with 512×512 pixel size containing $29,085$ buildings that are separated into 13,662 training sets and 3,726 testing sets.

The Massachusetts Buildings Dataset consists of 151 aerial images with 1500×1500 resolution and 1.0m GSD covering the Boston area. The entire datasets are split into 137, 10, and 4 images as the training, testing, and validation sets, respectively. Besides, merging and cropping were conducted to use the dataset properly. The validation set is assigned to the training set to enhance the deep learning model strictly. The 1500×1500 pixel images are cropped into 9 smaller images with 512×512 resolution with overlapping, while some blank areas are abandoned. As a result, we can get 1,100 images for the training set and 90 images for the testing set with 512×512 pixel size and 1.0m GSD.

We used only 2D RGB satellite or aerial images for all datasets, although the offered dataset contains other types of materials such as 3D images or multi-spectral data. Furthermore, although some source images contain incorrect ground truth or blank areas, all images are used to train and test without abandoning. Moreover, to estimate the ability of enhancement strictly, all datasets are used as given by providers without any augmentation.

B. Implementation details

In these experiments, we used various U-Net-like backbone network. U-Net[30], ResUNet++[32], and TernausNet[19] were adopted to validate our proposed approach. U-Net includes an encoder-decoder architecture with 5-stages. Each stage has a typical structure with two convolution filters and two activation functions, alternately. ResUNet++[32] is an advanced architecture for semantic segmentation containing SE blocks [43], residual blocks[47], ASPP [48], and attention blocks[49]. ResUNet++ has 4-stage structure with the same number of downsampling and upsampling sequences. However, ResUNet++ is regarded as having five stages because it contains an ASPP bridge part acting as a fifth downsampling layer. For the proposed method, both U-Net and ResUNet++ include five side outputs, whereas the TernausNet[19], based on VGG11[38] encoder, contains a sixth stage and the same number of side outputs.

Our proposed HED-unit connected to the encoder part of a U-Net-like network would obtain five or six side outputs depending on the backbone architecture. As a result, both outputs of the HED-unit and decoder in the backbone network are be used as two parallel inputs of the BE module. The structure of the BE module maintains equally regardless of the type of backbone, without a significant increase in the calculation cost. The amount of floating point operations per second(FLOPs) and calculation parameters used are listed in Table I, as followed by the 512×512 size of input image.

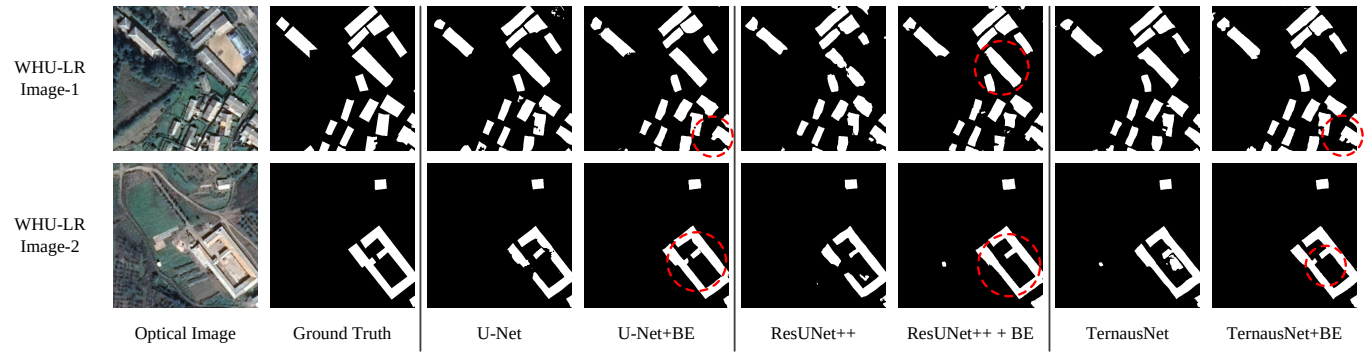

Fig. 9: Experimental result of WHU-LR Dataset for each backbone networks. The boundaries of building footprint are straightened and manifested, while adjacent buildings are distinguished well.

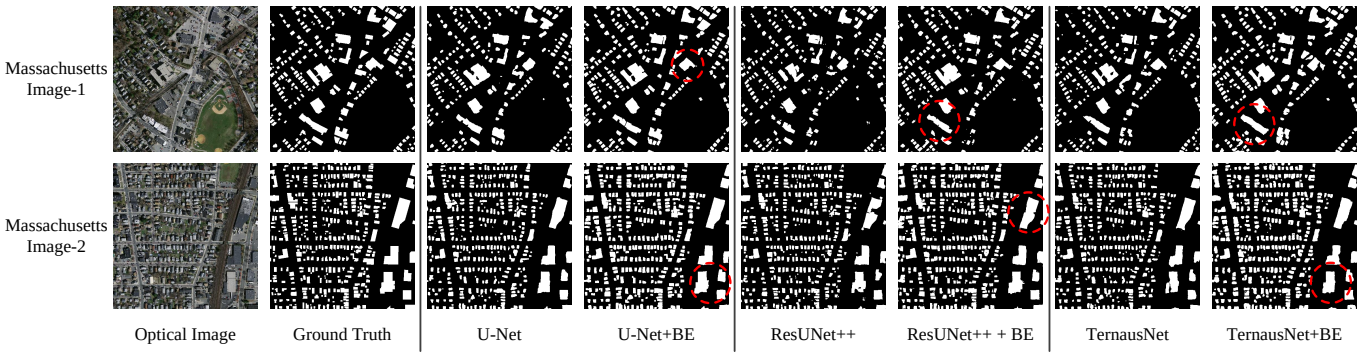

Fig. 10: Experimental result of Massachusetts Dataset for each backbone networks. The boundaries of building footprint are straightened and manifested, while adjacent buildings are distinguished well.

Model	Params (M)	FLOPs (G)
U-Net	34.52	130.90
U-Net + BE	34.59	140.64
ResUNet++	14.48	141.71
ResUNet++ + BE	14.52	146.64
TernausNet	22.93	84.24
TernausNet +BE	22.97	89.14
USPP	18.81	97.58
USPP +BE	18.93	112.14

TABLE I: Comparison of parameters and FLOPs on each backbones shows that the addition of proposed approach causes marginal increment.

The Adam[50] optimizer was used with $\beta_1 = 0.9, \beta_2 = 0.999$ as weight decay. The initial learning rate was set to $lr = 10^{-4}$ without a learning scheduler. The batch size was set to 4 because of a limitation on GPU usage. All hyperparameters were maintained in the same manner to conduct recursive experiments for all datasets and backbones in equivalent conditions. In the training procedure, all outputs of the network are passed through the sigmoid activation function to apply loss functions.

C. Evaluation

1) Inference: In the inference procedure, the output of the *positive replacement* in the BE module is used as the final prediction. When this segmentation mask representation passes through the sigmoid activation function in the training process, the raw results from *positive replacement* are used as a prediction result without any activation function. The positive part of the output indicates the extracted building footprints. Moreover, the negative values of the segmentation mask representation are considered as background.

2) Metric: To evaluate the performance of the proposed method, pixel-wise estimations were conducted, indicating the quality of each experiment. F1 score and Jaccard coefficient were used as evaluation metrics.

F1 score is defined as the harmonic mean of a model's precision and recall by classifying each pixel as true or false. F1 score is widely used to evaluate binary classification systems, commonly in semantic segmentation tasks, because it indicates a well-balanced prediction between precision and recall.

The Jaccard coefficient is also derived from precision and recall. The Jaccard coefficient is useful for estimating semantic segmentation performance because it shows the similarity between the two structures of prediction and ground truth intuitively.

D. Results Analysis

The results of the experiments are presented in Tables II, III, IV, V, and VI. The results show that our proposed method can improve the ability to extract building footprints from optical images for all the experimental datasets and backbone networks. The F1 score and Jaccard scores for all

Model	Precision	Recall	F1 Score	Jaccard
U-Net	0.82576	0.80547	0.79807	0.68606
U-Net + BE	**0.82677**	**0.80813**	**0.80403**	**0.69218**
ResUNet^{++}	0.81605	0.76335	0.76977	0.65272
ResUNet^{++}+ BE	**0.82760**	**0.78717**	**0.78986**	**0.67647**
TernausNet	**0.84196**	0.77473	0.78740	0.67618
TernausNet + BE	0.82874	**0.80531**	**0.80436**	**0.69257**
USPP	0.80923	0.76833	0.76970	0.65336
USPP + BE	**0.83140**	**0.80805**	**0.80499**	**0.69476**
Mean : base	0.82325	0.77797	0.78123	0.66708
Mean : base + BE	**0.82862**	**0.80216**	**0.80081**	**0.68899**

TABLE II: Comparison of applying proposed method for the various backbone network on DeepGlobe dataset.

Model	Precision	Recall	F1 Score	Jaccard
U-Net	**0.86880**	0.73122	0.78720	0.66150
U-Net + BE	0.85624	**0.78489**	**0.81350**	**0.69670**
ResUNet^{++}	0.83855	0.72414	0.77028	0.64240
ResUNet^{++}+ BE	**0.84410**	**0.75251**	**0.78905**	**0.66244**
TernausNet	0.84441	0.74676	0.78606	0.65788
TernausNet + BE	**0.85607**	**0.75970**	**0.79902**	**0.67683**
USPP	0.89237	0.75147	0.80908	0.69150
USPP + BE	**0.89301**	**0.76556**	**0.81868**	**0.70401**
Mean : base	0.86103	0.73840	0.78816	0.66332
Mean : base + BE	**0.86236**	**0.76567**	**0.80506**	**0.68500**

TABLE III: Comparison of applying proposed method for the various backbone network on Urban 3D dataset.

Model	Precision	Recall	F1 Score	Jaccard
U-Net	**0.93399**	0.92463	0.92396	0.86911
U-Net + BE	0.92926	**0.93667**	**0.92793**	**0.87456**
ResUNet^{++}	0.91847	0.90808	0.90645	0.84237
ResUNet^{++}+ BE	**0.92230**	**0.91035**	**0.90870**	**0.84581**
TernausNet	**0.92903**	0.91349	0.91558	0.85547
TernausNet + BE	0.91114	**0.93838**	**0.91991**	**0.86006**
USPP	**0.93949**	0.91584	0.91964	**0.86554**
USPP + BE	0.90755	**0.94953**	**0.92241**	0.86540
Mean : base	**0.93025**	0.91551	0.91641	0.85812
Mean : base + BE	0.91756	**0.93373**	**0.91974**	**0.86146**

TABLE IV: Comparison of applying proposed method for the various backbone network on WHU-HR dataset.

Model	Precision	Recall	F1 Score	Jaccard
U-Net	0.86461	0.70668	0.75567	0.62906
U-Net + BE	**0.86525**	**0.73607**	**0.78178**	**0.65530**
ResUNet^{++}	**0.85672**	0.68817	0.74195	0.60969
ResUNet^{++}+ BE	0.83958	**0.70782**	**0.75089**	**0.61731**
TernausNet	0.83262	**0.76335**	0.78493	0.65801
TernausNet + BE	**0.84453**	0.76126	**0.78645**	**0.66162**
USPP	**0.88645**	0.67198	0.74283	0.61374
USPP + BE	0.83906	**0.72449**	**0.75800**	**0.62819**
Mean : base	**0.86010**	0.70755	0.75635	0.62763
Mean : base + BE	0.84711	**0.73241**	**0.76928**	**0.64061**

TABLE V: Comparison of applying proposed method for the various backbone network on WHU-LR Asia dataset.

experiments were improved, as shown in Tables II, III, IV, V, and VI. The average F1 scores of the four backbones for each dataset(DeepGlobe, Urban3D, WHU-HR, WHU-LR, Massachusetts) were increased by about 1.43%p, 1.93%p, 0.35%p, 1.19%p, and 1.22%p, respectively, compared to baseline and proposed approaches. Similarly, the Jaccard scores also increased by approximately 1.54%p, 2.47%p, 0.45%p, 1.30%p, and 1.25%p, respectively. Furthermore, the improvement of the performance was relatively higher than the variation of computational parameters, about 0.32% on average, as shown in Table I.

The results are shown in Figs. 6, 7, 8, 9, and 10. As shown by the red dashed circles in each experimental figure, the improved method could classify pixels more precisely, even if the target buildings had a complicated shape. In particular, although impediments may disturb the ability to detect the edge area of a target, the results show that the proposed method can maintain the rectilinear boundaries of buildings detected in remote sensing applications. Moreover, the proposed method helps classify adjacent buildings separately, which tend to be detected as single buildings. These results show that our proposed method can be used widely, regardless of area of interest and backbone network.

E. Robustness

In order to investigate the robustness of enhancement of the proposed method, repetitive experiments were conducted ten times with all backbone networks on the Urban3D and Massachusetts dataset representing high-resolution and low-resolution, respectively. Table VII shows the evaluation result of repetitive experiments.

The combination of the HED unit and BE module improve the F1-score for each experiment compared to baseline without exception. Furthermore, the standard deviation results are lowered in modified networks. It shows that the proposed approach can stabilize the network to extract the precise building segmentation mask, regardless of the target image's resolution.

F. Comparison of Recent Method

The comparison results are shown in Table VIII. The evaluation results are improved in boundary-enhanced method for all U-Net-like architectures, including state-of-the-art(SOTA) U-Net-like methodologies, USPP and BRR-Net. Moreover, the USPP network with the proposed method achieved a higher F1-score and Jaccard score than recent Non-U-Net-like methods, MAP-Net and DE-Net. The Urban3D and Massachusetts datasets are adopted to prove the ability of boundary reinforcement for the high-resolution image(0.3m GSD) and low-resolution image(1.0m GSD).

The comparison results in Table VIII show that any U-Net-like network can adopt our proposed approach. Furthermore, if a well-performing backbone is given, the HED unit and BE module can supplement it to outperform other SOTA architectures.

G. Ablation study

To investigate the contribution of the proposed approach and MS-SSIM[46] loss, we designed ablation experiments for all backbone networks on the Urban3D dataset. Similar to previous experiments, the F1 score and Jaccard score were used as evaluation metrics to estimate accuracy.

In addition to Table III, we conducted a further experiment for the baseline with the proposed HED-unit and BE module without MS-SSIM loss. Unlike equation (7), the loss function

Model	Precision	Recall	F1 Score	Jaccard
U-Net	**0.83242**	0.78581	0.80358	**0.67663**
U-Net + BE	0.79366	**0.82605**	**0.80446**	0.67587
ResUNet^{++}	**0.81641**	0.66833	0.72758	0.57764
ResUNet^{++}+ BE	0.79855	**0.72761**	**0.75738**	**0.61007**
TernausNet	**0.82742**	0.74835	0.78222	0.64514
TernausNet + BE	0.78833	**0.79740**	**0.78721**	**0.65249**
USPP	**0.85439**	0.77684	0.80884	0.68312
USPP + BE	0.80159	**0.83593**	**0.81237**	**0.68829**
Mean : base	**0.83244**	0.74483	0.78056	0.64563
Mean : base + BE	0.79553	**0.79675**	**0.79035**	**0.65668**

TABLE VI: Comparison of applying proposed method for the various backbone network on Massachusetts dataset.

Dataset	F1-score			
	Urban3D		Massachusetts	
	Mean	Std. Dev.	Mean	Std. Dev.
U-Net	0.79315	0.00667	0.80079	0.01243
U-Net + BE	**0.81578**	**0.00167**	**0.80265**	**0.00399**
ResUNet++	0.76797	0.00632	0.73760	0.00484
ResUNet++ + BE	**0.78818**	**0.00433**	**0.75048**	**0.00471**
TernausNet	0.78173	0.00361	0.78051	0.00619
TernausNet +BE	**0.80342**	**0.00348**	**0.78886**	**0.00304**
USPP	0.80498	0.01362	0.79652	0.01452
USPP + BE	**0.81878**	**0.00787**	**0.81147**	**0.00441**

TABLE VII: Repetitive experiments were conducted ten times with each backbone network.

Dataset		Urban3D		Massachusetts	
Model		F1-score	Jaccard	F1-score	Jaccard
Baseline U-Net -like	U-Net	0.78720	0.66150	0.80358	0.67663
	ResUNet++	0.77028	0.64240	0.72758	0.57764
	TernausNet	0.78606	0.65788	0.78222	0.64514
	USPP	0.80908	0.69150	0.80884	0.68312
	BRR-Net	0.79960	0.67716	0.79291	0.66138
Non-U-Net -like	MAP-Net	0.81615	0.70397	0.79638	0.66657
	DE-Net	0.81694	0.70446	0.81179	0.68630
Modified U-Net -like (+BE)	U-Net	0.81350	0.69670	0.80446	0.67587
	ResUNet++	0.78905	0.66244	0.75738	0.61007
	TernausNet	0.79902	0.67683	0.78721	0.65249
	USPP	**0.82163**	**0.70901**	**0.81237**	**0.68829**
	BRR-Net	0.80893	0.69258	0.79676	0.66635

TABLE VIII: Comparison of most recent building extraction methods on Urban3D dataset and Massachusetts dataset.

Model	Precision	Recall	F1 Score	Jaccard
U-Net	**0.86880**	0.73122	0.78720	0.66150
U-Net+BE+BCE	0.86190	0.76008	0.80211	0.68082
U-Net+BE+MS	0.85624	**0.78489**	**0.81350**	**0.69670**
ResUNet^{++}	0.83855	0.72414	0.77028	0.64240
ResUNet^{++}+BE+BCE	0.83414	0.74910	0.77992	0.65088
ResUNet^{++}+BE+MS	**0.84410**	**0.75251**	**0.78905**	**0.66244**
TernausNet	0.84441	0.74676	0.78606	0.65788
TernausNet+BE+BCE	0.83851	**0.76641**	0.79493	0.67055
TernausNet+BE+MS	**0.85607**	0.75970	**0.79902**	**0.67683**

TABLE IX: Ablation study results for Urban3D datasets. To convince the performance of HED-Unit and BE Module, additional experiments were executed using Binary Cross-Entropy(BCE) Loss only, without proposed combination loss function as shown in the above table as MS.

of the mask was set as a binary cross-entropy loss. The weights of each loss function were maintained equally, 1:1:1. The results are listed in Table IX. The F1 score and Jaccard score results for the additional experiments were improved in comparison to the baseline model for all backbone networks.

The scores tended to be lower than the proposed combination of the loss function. The ablation study shows that the HED-unit and BE module were able to enhance the performance of semantic segmentation sufficiently, and the proposed loss function guarantees the stable improvement of performance.

V. Conclusion

We have proposed a method to enhance the ability of automated systems to detect the boundaries of semantic segmentation masks in building extraction from remote sensed imagery. The method is composed of two major steps: a HED-unit and a BE module, which can be adopted in various U-Net-like convolutional neural networks containing an encoder-decoder architecture. The HED-unit is connected to an encoder directly to detect the boundary of a target object well. The BE module is a tail part of the entire network and produces the final output by combining the information of the edge and segmentation masks by probability map and positive replacement.

For this boundary enhancement method, the boundary of the extracted building footprints can be smoother than the existing semantic segmentation without a considerable increase in calculation cost. The evaluation metrics with F1 score and Jaccard coefficient also show that this boundary enhancement method produced improved output in terms of the accuracy of segmentation and refined the sensitivity and specificity. Moreover, the new combination of loss functions guarantees the network to generate a segmentation-mask-preserving structure.

Also, the experimental results show that the proposed method is feasible to both high-resolution image (DeepGlobe, Urban3D, WHU-HR) and low-resolution dataset(WHU-LR, Massachusetts). Consequently, the limit resolution of the proposed approach is not found through these experiments. Though, we suppose that the proposed approach may not enhance the performance of extraction when the boundary and mask are not clearly distinguished due to the low-resolution or damage in given image. Nevertheless, the combination of HED unit and BE module surely improve the extracting ability regardless of the resolution, target location, and backbone network.

In general, our research provides a new approach for amending the shape of extracted buildings while improving accuracy metrics. Currently, the proposed method is implemented in building extraction with U-Net-like CNN architectures. In future works, we intend to further study various architectures such as DeepLabv3[51], Mask R-CNN[52], and tasks such as multi-class extraction, road detection, and land cover classification.

Acknowledgment

Myungjoo Kang was supported by the National Research Foundation grant of Korea (2015R1A5A1009350, 2021R1A2C3010887) and the ICT R&D program of MSIT/IITP(No. 1711117093).

REFERENCES

[1] T. Blaschke, S. Lang, E. Lorup, J. Strobl, and P. Zeil, "Object-oriented image processing in an integrated gis/remote sensing environment and perspectives for environmental applications," *Environmental information for planning, politics and the public*, vol. 2, pp. 555–570, 2000.

[2] K. Whitehead and C. H. Hugenholtz, "Remote sensing of the environment with small unmanned aircraft systems (uass), part 1: A review of progress and challenges," *Journal of Unmanned Vehicle Systems*, vol. 2, no. 3, pp. 69–85, 2014.

[3] J. M. Almendros-Jimenez, L. Domene, and J. A. Piedra-Fernandez, "A framework for ocean satellite image classification based on ontologies," *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 6, no. 2, pp. 1048–1063, 2012.

[4] B. C. Gallo, J. A. Demattê, R. Rizzo, J. L. Safanelli, W. d. S. Mendes, I. F. Lepsch, M. V. Sato, D. J. Romero, and M. P. Lacerda, "Multi-temporal satellite images on topsoil attribute quantification and the relationship with soil classes and geology," *Remote Sensing*, vol. 10, no. 10, p. 1571, 2018.

[5] D. J. Marceau, D. J. Gratton, R. A. Fournier, and J.-P. Fortin, "Remote sensing and the measurement of geographical entities in a forested environment. 2. the optimal spatial resolution," *Remote Sensing of Environment*, vol. 49, no. 2, pp. 105–117, 1994.

[6] J. L. Ohmann, M. J. Gregory, and H. M. Roberts, "Scale considerations for integrating forest inventory plot data and satellite image data for regional forest mapping," *Remote sensing of environment*, vol. 151, pp. 3–15, 2014.

[7] M. Wójtowicz, A. Wójtowicz, J. Piekarczyk *et al.*, "Application of remote sensing methods in agriculture," *Communications in Biometry and Crop Science*, vol. 11, no. 1, pp. 31–50, 2016.

[8] A. Zhang and G. Jia, "Monitoring meteorological drought in semiarid regions using multi-sensor microwave remote sensing data," *Remote sensing of Environment*, vol. 134, pp. 12–23, 2013.

[9] F. Paul, C. Huggel, and A. Kääb, "Combining satellite multispectral image data and a digital elevation model for mapping debris-covered glaciers," *Remote sensing of Environment*, vol. 89, no. 4, pp. 510–518, 2004.

[10] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 165, pp. 108–119, 2020.

[11] S. Liang, *Comprehensive Remote Sensing*. Elsevier, 2017.

[12] Y. Chen, D. Ming, and X. Lv, "Superpixel based land cover classification of vhr satellite image combining multi-scale cnn and scale parameter estimation," *Earth Science Informatics*, vol. 12, no. 3, pp. 341–363, 2019.

[13] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *ISPRS TC VII Symposium-100 Years ISPRS*, vol. 38, 2010, pp. 298–303.

[14] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016.

[15] F. Petitjean, J. Inglada, and P. Gançarski, "Satellite image time series analysis under time warping," *IEEE transactions on geoscience and remote sensing*, vol. 50, no. 8, pp. 3081–3095, 2012.

[16] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and lidar data using ip-cnn," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[17] X. Liu, L. Jiao, L. Li, L. Cheng, F. Liu, S. Yang, and B. Hou, "Deep multiview union learning network for multisource image classification," *IEEE Transactions on Cybernetics*, 2020.

[18] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[19] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.

[20] W. Kang, Y. Xiang, F. Wang, and H. You, "Eu-net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sensing*, vol. 11, no. 23, p. 2813, 2019.

[21] G. Wu, X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu, and R. Shibasaki, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sensing*, vol. 10, no. 3, p. 407, 2018.

[22] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[23] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "Mapnet: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[24] H. Liu, J. Luo, B. Huang, X. Hu, Y. Sun, Y. Yang, N. Xu, and N. Zhou, "De-net: Deep encoding network for building extraction from high-resolution remote sensing imagery," *Remote Sensing*, vol. 11, no. 20, p. 2380, 2019.

[25] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2178–2189, 2019.

[26] G. Wu, Z. Guo, X. Shi, Q. Chen, Y. Xu, R. Shibasaki, and X. Shao, "A boundary regulated network for accurate roof segmentation and outline extraction," *Remote Sensing*, vol. 10, no. 8, p. 1195, 2018.

[27] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation," in *International Symposium on Neural Networks*. Springer, 2019, pp.

388–401.

[28] A. Van Etten, D. Hogan, J. M. Manso, J. Shermeyer, N. Weir, and R. Lewis, "The multi-temporal urban development spacenet dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6398–6407.

[29] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask r-cnn with building boundary regularization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 247–251.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[31] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[32] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019, pp. 225–2255.

[33] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128 774–128 786, 2019.

[34] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 172–181.

[35] H. R. Goldberg, S. Wang, G. A. Christie, and M. Z. Brown, "Urban 3d challenge: building footprint detection using orthorectified imagery and digital surface models from commercial satellites," in *Geospatial Informatics, Motion Imagery, and Network Analytics VIII*, vol. 10645. International Society for Optics and Photonics, 2018, p. 1064503.

[36] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[37] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, University of Toronto, 2013.

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

[39] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[40] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai, "Brrnet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sensing*, vol. 12, no. 6, p. 1050, 2020.

[41] Y. Kim, S. Kim, T. Kim, and C. Kim, "Cnn-based semantic segmentation using level set loss," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1752–1760.

[42] Z. Chen, H. Zhou, J. Lai, L. Yang, and X. Xie, "Contour-aware loss: Boundary-aware learning for salient object segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 431–443, 2020.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[44] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[46] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

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

[48] 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, 2017.

[49] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[51] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

Hoin Jung received the B.S. degree in Aircraft System Engineering in Aerospace & Mechanical Engineering Department from the Korea Aerospace University in 2014, Goyang, Korea. From 2014 to 2017, he was a mechanical engineer in the Korea Air Force as an officer. From 2017 to 2020, he was a mechanical engineer in the Department of Digital appliances, Samsung Electronics Co., Ltd. He is now in an M.S. course in the Department of Computational Science and Technology at Seoul National University. His research interests include remote sensed image processing using deep learning.

Han-Soo Choi received athe B.S. degree in Mathematics from the Kookmin University in 2011, Seoul, Korea, and an M.S. degree in Mechatronics at the Korea University from 2013-2015 in Seoul, Korea, and a undertook a Ph.D course in the Department of Computational Science and Technology at Seoul National University from 2016-2020. He is now a training researcher at the Research Institute of Mathematics, Seoul National University. His research interests include image processing using deep learning.

Myungjoo Kang received a B.S. degree in mathematics from Seoul National University, Seoul, Korea, and his Ph.D. degree in Mathematics from the University of California, Los Angeles, in 1996. He was with the Electrical and Computer Engineering Department from the University of California, San Diego, as a Postdoctoral Researcher, from to 1996-2000. He has been an Assistant Professor, from 2003-2007, and an Associate Professor, from 2008-2013, and a Pprofessor, from 2014-present at the Department of Mathematical Sciences, Seoul National University. His research interests include in mathematical image processing, as well as numerical schemes and computational fluid dynamics. His current research interests are focused on image processing using deep learning.