



저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#) 

Master's Thesis of Engineering

Semi-supervised Learning
Framework for Very High
Resolution Image Classification
Using CycleGAN

초고해상도 영상 분류를 위한 순환 적대적 생성
신경망 기반의 준지도 학습 프레임워크

August 2021

Department of
Civil and Environmental Engineering
Seoul National University

Taehong Kwak

Semi-supervised Learning Framework for Very High Resolution Image Classification Using CycleGAN

Advisor Yongil Kim

Submitting a Master's Thesis of Science

June 2021

Department of
Civil and Environmental Engineering
Seoul National University

Taehong Kwak

Confirming the Master's Thesis written by
Taehong Kwak

August 2021

Chair _____
Vice Chair _____
Examiner _____

Abstract

Image classification of Very High Resolution (VHR) images is a fundamental task in the remote sensing domain for various applications such as land cover mapping, vegetation mapping, and urban planning. In recent years, deep convolutional neural networks have shown promising performance in image classification studies. In particular, semantic segmentation models with fully convolutional architecture-based networks demonstrated great improvements in terms of computational cost, which has become especially important with the large accumulation of VHR images in recent years.

However, deep learning-based approaches are generally limited by the need of a sufficient amount of labeled data to obtain stable accuracy, and acquiring reference labels of remotely-sensed VHR images is very labor-extensive and expensive. To overcome this problem, this thesis proposed a semi-supervised learning framework for VHR image classification. Semi-supervised learning uses both labeled and unlabeled data together, thus reducing the model's dependency on data labels. To address this issue, this thesis employed a modified CycleGAN model to utilize large amounts of unlabeled images.

CycleGAN is an image translation model which was developed from Generative Adversarial Networks (GAN) for image generation. CycleGAN trains unpaired dataset by using cycle consistency loss with two generators and two discriminators. Inspired by the concept of cycle consistency, this thesis modified CycleGAN to enable the use of unlabeled VHR data in model training by considering the unlabeled images as images unpaired with their corresponding ground truth maps.

To utilize a large amount of unlabeled VHR data and a relatively small amount of labeled VHR data, this thesis combined a supervised learning classification model with the modified CycleGAN architecture. The proposed framework contains three phases: cyclic phase, adversarial phase, and

supervised learning phase. Through the three phase, both labeled and unlabeled data can be utilized simultaneously to train the model in an end-to-end manner.

The result of the proposed framework was evaluated by using an open-source VHR image dataset, referred to as the International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen dataset. To validate the accuracy of the proposed framework, benchmark models including both supervised and semi-supervised learning methods were compared on the same dataset. Furthermore, two additional experiments were conducted to confirm the impact of labeled and unlabeled data on classification accuracy and adaptation of the CycleGAN model for other classification models. These results were evaluated by the popular three metrics for image classification: Overall Accuracy (OA), F1-score, and mean Intersection over Union (mIoU).

The proposed framework achieved the highest accuracy (OA: 0.796, 0.786, and 0.784, respectively in three test sites) in comparison to the other five benchmarks. In particular, in a test site containing numerous objects with various properties, the largest increase in accuracy was observed due to the regularization effect from the semi-supervised method using unlabeled data with the modified CycleGAN. Moreover, by controlling the amount of labeled and unlabeled data, results indicated that a relatively sufficient amount of unlabeled and labeled data is required to increase the accuracy when using the semi-supervised CycleGAN. Lastly, this thesis applied the proposed CycleGAN method to other classification models such as the feature pyramid network (FPN) and the pyramid scene parsing network (PSPNet), in place of UNet. In all cases, the proposed framework returned significantly improved results, displaying the framework's applicability for semi-supervised image classification on remotely-sensed VHR images.

Keyword : Semi-supervised Deep Learning, CycleGAN, Very High Resolution Image Classification, Semantic Segmentation

Student Number : 2019-24705

Table of Contents

Abstract	i
List of Tables	v
List of Figures	vi
1. Introduction	1
2. Background and Related Works	6
2.1. Deep Learning for Image Classification	6
2.1.1. Image-level Classification.....	6
2.1.2. Fully Convolutional Architectures	7
2.1.3. Semantic Segmentation for Remote Sensing Images.....	9
2.2. Generative Adversarial Networks (GAN).....	12
2.2.1. Introduction to GAN	12
2.2.2. Image Translation.....	14
2.2.3. GAN for Semantic Segmentation.....	16
3. Proposed Framework.....	20
3.1. Modification of CycleGAN.....	22
3.2. Feed-forward Path of the Proposed Framework	23
3.2.1. Cyclic Phase	23
3.2.2. Adversarial Phase.....	23
3.2.3. Supervised Learning Phase	24
3.3. Loss Function for Back-propagation.....	25
3.4. Proposed Network Architecture	28
3.4.1. Generator Architecture.....	28
3.4.2. Discriminator Architecture.....	29

4. Experimental Design	31
4.1. Overall Workflow	33
4.2. Vaihingen Dataset.....	38
4.3. Implementation Details	40
4.4. Metrics for Quantitative Evaluation.....	41
5. Results and Discussion	42
5.1. Performance Evaluation of the Proposed Feamework.....	42
5.2. Comparison of Classification Performance in the Proposed Framework and Benchmarks.....	45
5.3. Impact of labeled and Unlabeled Data for Semi-supervised Learning	52
5.4. Cycle Consistency in Semi-supervised Learning.....	55
5.5. Adaptation of the GAN Framework for Other Classification Models.....	59
6. Conclusion	62
Reference	65
국문 초록.....	69

List of Tables

Table 1. Comparison of the proposed framework and Mondal et al. (2019)'s method	35
Table 2. Real images of false color and ground truth maps in the three test sites: Area 1, Area 15, Area 23	39
Table 3. Learning rate schedules of the four networks.	40
Table 4. Confusion matrix including the number of the predicted and true pixels for each class with recall and precision scores.	44
Table 5. Overall classification accuracy in OA, F1-score of benchmarks and the proposed framework	46
Table 6. F1-score and OA in three test sites according to the number of labeled and unlabeled data.....	54
Table 7. The real image, fake class map, reconstructed image from semi-supervised method and fake class map from supervised method with ground truth accoring to epoch.....	56
Table 8. F1-score and mIoU of three classification models in supervised and semi-supervised learning	59

List of Figures

Figure 1. Model scaling for width, depth, resolution and compound scaling (Tan and Le, 2019).....	7
Figure 2. Depthwise separable convolution and inverted residual block in the MBConv of EfficientNet (Sandler et al., 2018).....	7
Figure 3. The architecture of UNet (Ronnerberger et al., 2015).....	9
Figure 4. The dense DownBlock and UpBlock.....	11
Figure 5. Overview of GAN structure.....	12
Figure 6. Example images for the image-to-image translation task (Isola et al., 2017).....	15
Figure 7. The structure of Pix2Pix to translate edges-to-photo (Isola et al., 2017).....	15
Figure 8. The workflows of CycleGAN (Zhu et al., 2017).....	16
Figure 9. Overview the semantic segmentation approach using adversarial networks (Luc et al., 2016).....	17
Figure 10. Overview of the semi-supervised semantic segmentation system with adversarial learning (Hung et al., 2018).....	19
Figure 11. The proposed semi-supervised framework using the modified CycleGAN.....	21
Figure 12. EfficientUNet architecture employed in this thesis.....	30
Figure 13. PatchGAN-based discriminator architecture employed in this thesis.....	30
Figure 14. Overall workflow of this thesis including data setting and three experiments.....	32
Figure 15. Overview of FPN architecture (Lin et al., 2017-b).....	36
Figure 16. Overview of PSPNet architecture (Zhao et al., 2017).....	37
Figure 17. The modified architectures of FPN and PSPNet employed in this thesis.....	37
Figure 18. The predicted maps by the proposed semi-supervised framework and ground truth map in three test sites.....	43
Figure 19. Close-up view of ground truth, the result map by benchmarks, and the proposed framework in Area 1.....	48
Figure 20. Close-up view of ground truth, the result map by benchmarks, and the proposed framework in Area 15.....	49
Figure 21. Close-up view of ground truth, the result map by benchmarks, and the proposed framework in Area 23.....	50
Figure 22. mIoU in three test sites according to the number of labeled and unlabeled data.....	53
Figure 23. The validation accuracy graph during model training of EfficientUNet (supervised) and EfficientUNet + modified CycleGAN (semi-supervised).....	55
Figure 24. The cycle consistency loss graph of two cycle parts.....	57

Figure 25. mIoU graph for each class by the three classification models and the proposed framework60

Chapter 1. Introduction

With recent advancements in remote sensing technology, Very High Resolution (VHR) images from various satellite and airborne sensors have become more accessible. VHR images provide detailed information on the observed land surface which can be used to improve our understanding of the Earth's environment. In particular, image classification using VHR images is a fundamental task to address many practical applications such as land-use/land-cover (LULC) mapping, urban planning, and vegetation mapping (Van de Voorde et al., 2007; Bellen et al., 2008; Feng et al., 2015). Given the rich spatial information and the fine-grained detail in VHR images, a diverse number of classification studies have been conducted. Stemming from this demand, there is a more pressing need to develop automatic VHR image classification algorithms in light of the rapidly increasing volume of VHR data amassed in recent years.

Recently, Deep Learning (DL) techniques have shown state-of-the-art performance in various Computer Vision (CV) tasks through the use of deeper network structures to hierarchically extract high-level features, while conventional image classification methods depend on hand-crafted features leading to poor performance of classification. Especially, Convolutional Neural Network (CNN) has yielded breakthrough results in CV-based image classification due to the use of deep convolutional layers that can capture spatial features effectively through weight sharing and sparse connections. Based on the promising results from CNN-based classification methods, CNN models have been readily applied for remotely-sensed VHR image classification with demonstrating superior classification accuracy.

Remote sensing image classification methods can be divided in terms of the model's output: traditional sliding window CNN-based and fully convolutional architecture-based (Viguera-Guillén et al., 2019). The sliding

window CNN-based method classifies each pixel by moving the window across the image, and the classification model outputs the input image as a single class. However, this approach is limited by expensive computational cost and redundancy issues due to the use of overlapping windows (Vigueras-Guillén et al., 2019). In contrast, the fully convolutional architecture-based approach can solve the computational cost of the sliding window CNN-based approach, since the architecture outputs a resulting image that is the same size as the input image without using the fully-connected layer included in conventional CNNs. This particular method can reduce redundant computation and exploit global information in an image, making this approach more computationally efficient than the sliding window CNN-based approach.

For VHR image classification, semantic segmentation based on fully convolutional models has been widely used. Semantic segmentation is a process to assign a specific semantic class label to each pixel in the image. For remotely-sensed VHR images, semantic segmentation using fully convolutional models is more efficient to process the large accumulation of VHR images in recent years, and has therefore been utilized in a wide range of applications such as road extraction (Zhang et al., 2018-b; Zhou et al., 2018), building footprint extraction (Van Etten et al., 2018; Shi et al., 2018; Bischke et al., 2019), change detection (Daudt et al., 2018; Peng et al., 2019), and multi-class image classification. In particular, since VHR image classification for multiple classes greatly increases data heterogeneity and required complexity of the used DL model, many sophisticated semantic segmentation models have been proposed in recent years (Kampffmeyer et al., 2016; Sherrah, 2016; Audebert et al., 2016; Iglovikov et al., 2017; Dong et al., 2019; Diakogiannis et al., 2020; Bai et al., 2021).

However, for achieving the stable classification results using DL-methods such as semantic segmentation, the models typically require a large amount

of labeled data, but acquiring the ground truth labels of multiple classes in remotely- sensed images is labor-intensive and expensive (Foody, 2002; Jin et al., 2014). Since an insufficient amount of training data can lower the performance and robustness of DL models, various strategies to overcome the high dependency on training data have been explored. Among these advancements, semi-supervised learning utilizes a small amount of labeled data and a relatively large amount of unlabeled data during model training. In semi-supervised methods, unlabeled data is used to provide additional information of the used dataset distribution or supplement pseudo labels, inducing the models to be regularized. Given the continuous accumulation of remotely-sensed VHR images being produced over time and the difficulty of manually labeling each image, semi-supervised learning can be a practically meaningful approach for image classification.

Meanwhile, Generative Adversarial Networks (GAN) were proposed by Goodfellow et al. (2014) for image generation and had been receiving extensive attention in various fields, especially for unsupervised learning. One of the representative applications of GAN is image-to-image translation which maps an image from the source domain to the target domain and produces a synthetic image. In particular, CycleGAN proposed by Zhu et al. (2017) translated “unpaired” images in the source and target domain. Here, “unpaired” refers to two image domains that are dissimilar and cannot be matched with each other. Unpaired images with spatially unmatched features cannot be trained directly using previous supervised learning approaches. To address this issue, cycle consistency loss was proposed by Zhu et al. (2017) and is based on the intuition that when an image is translated from one domain to another domain and back again to the initial domain, the resulting translated image is similar to the original image. To reduce the loss, two translators are trained toward preserving the input image’s features, while this process requires only input data without paired data to train the model.

Inspired by the cycle consistency, this thesis modified the CycleGAN architecture to enable the use of unlabeled VHR data for model training. This approach considered the two image domains as the original remotely-sensed image domain and the class map domain. From this perspective, unlabeled images can be considered as a VHR image unpaired with its corresponding ground truth map, which allows the unlabeled images to train the classification model.

This thesis aims to establish a semi-supervised learning framework to improve the classification accuracy of remotely-sensed VHR images. For this purpose, this thesis combines a supervised learning-based classification model with the modified CycleGAN architecture to utilize both a large amount of unlabeled data and a relatively small amount of labeled data. The classification model is trained to map an image from the remotely-sensed VHR image domain to the prediction map domain in the CycleGAN model by using labeled data in a supervised learning manner. At the same time, the cycle consistency and adversarial competition nature of CycleGAN are used to additionally train the model by using a large amount of unlabeled data in a semi-supervised learning manner. The main contributions of this thesis are as follows:

- An end-to-end semi-supervised learning framework is proposed which integrates the cycle consistency loss from CycleGAN, to enable the training of both labeled and unlabeled data. Through the application of CycleGAN's cycle consistency loss and reconstructed images for additional learning, this thesis confirms that applying unlabeled data in a semi-supervised learning manner aids model training by significantly improving classification accuracy.
- The proposed framework performed multi-class image

classification on 9cm very high resolution aerial images. To improve classification accuracy, this thesis replaced the translation model in CycleGAN with a UNet structure containing the pre-trained EfficientNet backbone. The classification results were evaluated with a comparison to five different benchmarks, which include both supervised and semi-supervised learning methods.

- This thesis conducted two in-depth experiments on data configuration and model implementation. First, the impact of different number of labeled and unlabeled data was investigated for the proposed semi-supervised framework. Second, the modified CycleGAN for semi-supervised classification was extended for use with three different classification models.

The remainder of this thesis is organized as follows: Chapter 2 introduces the background theory of this thesis with a review of related studies. Chapter 3 describes the proposed semi-supervised learning framework for VHR image classification, where the three phases of the framework, loss function for back-propagation, and network architecture are described. In Chapter 4, the detailed experimental design is explained. Chapter 5 provides the experimental results and discussion on the classification accuracy and resulting maps in comparison to benchmark studies. This section also includes an analysis of the impact of labeled and unlabeled data, the effect of cycle consistency loss, and the adaptation of the CycleGAN method for other classification models. Lastly, the conclusion of this thesis is given in Chapter 6.

Chapter 2. Background and Related Works

2.1. Deep Learning for Image Classification

2.1.1. Image-level Classification

In early deep learning-based approaches, deep neural networks were designed to obtain an image-level prediction corresponding to a single class. With the introduction of massive deep learning datasets such as ImageNet (Deng et al., 2009), more complex and deeper models have been proposed and reached higher classification accuracy. In 2016, the Residual Neural Network (ResNet) proposed by He et al. (2016) won the 2015 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by using shortcut connections through identity mapping to avoid the gradient degradation problem in deeper networks. This connection allowed models to be built deeper and to extract higher-level features without forfeiting a loss in accuracy.

More recently, Tan and Le (2019) proposed EfficientNet which has shown superior classification results through compound model scaling of three factors: network width, network depth, and resolution (figure 1). While conventional methods scale only one of these three factors or scale at random. Tan and Le (2019) argued that these factors could be uniformly scaled with a set of fixed coefficients. In more detail, the basic building blocks of EfficientNet consist of several mobile inverted bottleneck convolutions (MBConv), which greatly reduce the computational cost by employing depthwise separable convolution (figure 2 (a)) and inverted residuals (figure 2 (b)). EfficientNet controls the depth, width, and resolution of MBConv blocks with fixed scale and balances the network factors, enabling the network to achieve state-of-the-art image classification accuracy.

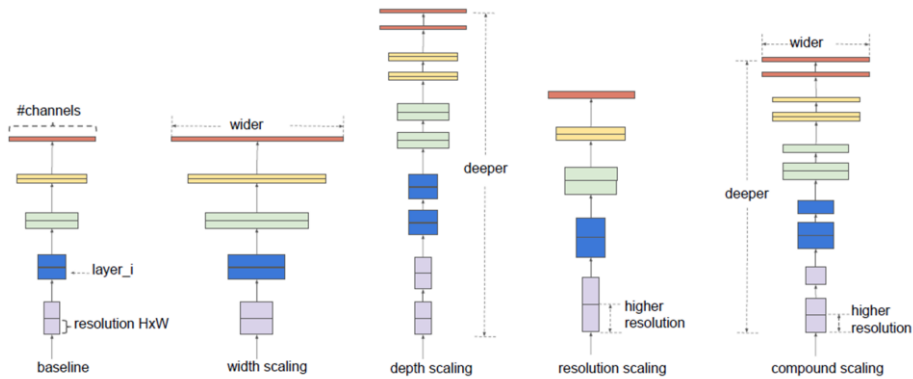


Figure 1. Model scaling for width, depth, resolution and compound scaling (Tan and Le, 2019).

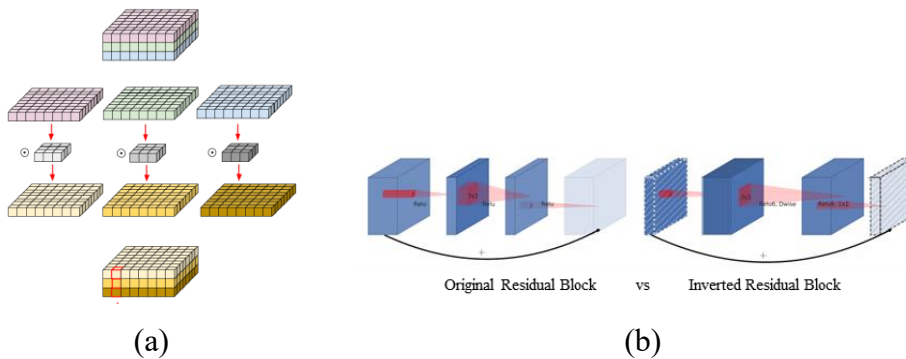


Figure 2. Depthwise separable convolution (a) and inverted residual block in the MBConv of EfficientNet (Sandler et al., 2018).

2.1.2. Fully Convolutional Architectures

To extend image-level prediction to pixel-level, early studies used the conventional CNN with a sliding window method so that the deep model predicts a class label for a target pixel from a patch of the image. However, this strategy is computationally expensive and faces pixel redundancy between overlapping patches (Viguera-Guillén et al., 2019). As a solution to this issue, Fully Convolutional Network (FCN) proposed by Long et al. (2015) performed pixel-level classification by replacing the fully connected layer of

conventional image classification networks with deconvolution layers. Unlike the dense layer, this fully convolutional structure enabled the preservation of the original image's spatial information and generated a segmented output image with the same size as an input image. To elaborate on the model architecture, FCN consists of downsampling and upsampling stages. Repetitive convolution and pooling layers in the downsampling reduce the size of the feature map, thus saving computational cost and preserving the size of the receptive field. Subsequently, the upsampling stage restores the resolution of the feature map to the size of the original input image. Although FCN has some limitations such as generating poor boundaries and losing detailed spatial information when producing feature maps, the downsampling-upsampling structure has served as the driving inspiration behind the design of most modern semantic segmentation models.

UNet was first proposed by Ronnerberger et al. (2015) for the segmentation of biomedical images. The structure of UNet can be divided into a contracting path and an expansive path, and is also referred to as an encoder-decoder which resembles a U-shaped architecture (figure 3). To address the blurring problem faced by the FCN, UNet employed skip connections to combine fine location information in shallow layers with global semantic features in the deep layers. Through the U-shaped architecture, UNet alleviated the trade-off between localization accuracy, while the use of context information was able to help yield more sophisticated segmentation results.

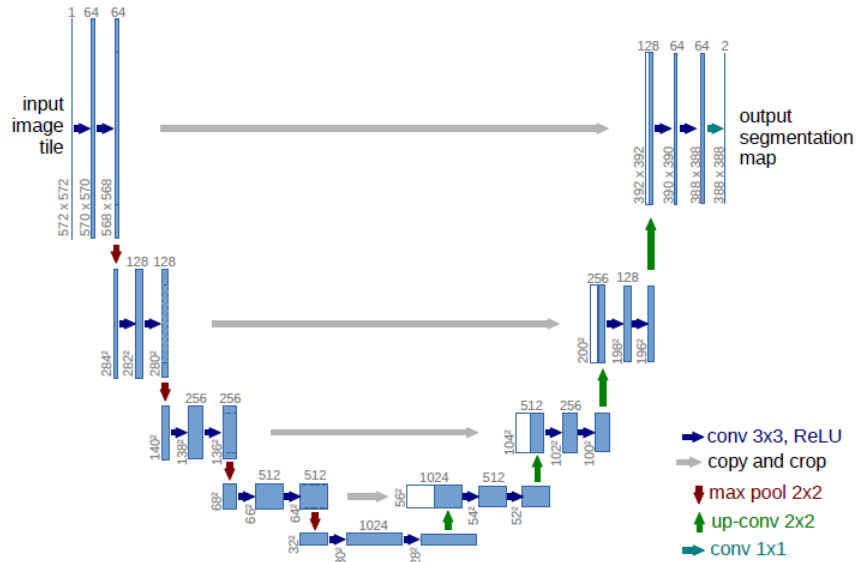


Figure 3. The architecture of UNet (Ronnerberger et al., 2015). Blue boxes indicate feature maps with different size

2.1.3. Semantic Segmentation for Remote Sensing Images

Semantic segmentation methods using the fully convolutional architecture are computationally more efficient compared to methods that use sliding window CNNs, especially for large VHR remote sensing images. Semantic segmentation models such as FCN and UNet have been implemented for multi-class classification using VHR images. Kampffmeyer et al. (2016) first applied FCN to perform multi-class semantic segmentation for VHR remote sensing images. Both sliding window-based and FCN-based classification approaches were used, and the study developed a combination of these models with median frequency balancing to achieve better overall classification accuracy for small objects. In addition, Iglovikov et al. (2017) employed UNet to classify VHR satellite multispectral images. To modify the model for VHR satellite images, Iglovikov et al. (2017) added a cropping layer to the output layers of U-Net to address boundary effects near the edge of each patch. The proposed UNet model was first applied to a WorldView-3

satellite image containing one panchromatic band, eight multispectral bands, and eight SWIR bands.

Recently, different feature extractors have been used in place of simple CNN-based encoders in the contracting path of segmentation networks as many sophisticated feature extractors based on image classification task have been developed. Firstly, Sherrah (2016) and Audebert et al. (2016) successfully applied VGG-16 (Simonyan and Zisserman, 2014) pre-trained by ImageNet for semantic segmentation of a remotely-sensed VHR image. Sherrah (2016) utilized a pre-trained VGG-16 network in FCN to make better use of robust features and achieved better VHR image classification accuracy compared to benchmark models. To address the limitation that pre-trained networks can only utilize three input channels, Audebert et al. (2016) combined two pre-trained encoders. In more detail, the first encoder receives three bands from the original image, and the second encoder receives additional input data, including DSM, nDSM, and NDVI as input bands. Following this dual-encoding path, two kinds of feature maps are merged together in the decoding path to output a single classified map.

However, employing the pre-trained model can face some limitations when the model is applied to domains that differ from the computer vision domain image used in pre-training. Due to this issue, feature extractors are often applied in the remote sensing domain without using pre-trained weights. Dong et al., (2019) employed the idea of dense connection in the encoding path of UNet structure. The encoder consists of several down-sampling blocks based on dense connection, which is illustrated in figure 4. In a dense block, all layers in a block are fully connected, or “densely” connected, to maximize the flow of information and gradient. After the dense connection, feature maps pass two paths composed of a max-pooling layer to reduce the dimension of feature maps and an up-sampling block of the decoder. In addition to improving the encoder’s structure, a focal loss function weighted

by median frequency balancing is employed to address class imbalance. Ultimately, Dong et al., (2019) demonstrated that the proposed methods achieved better accuracies compared to those of the original UNet and especially for small objects.

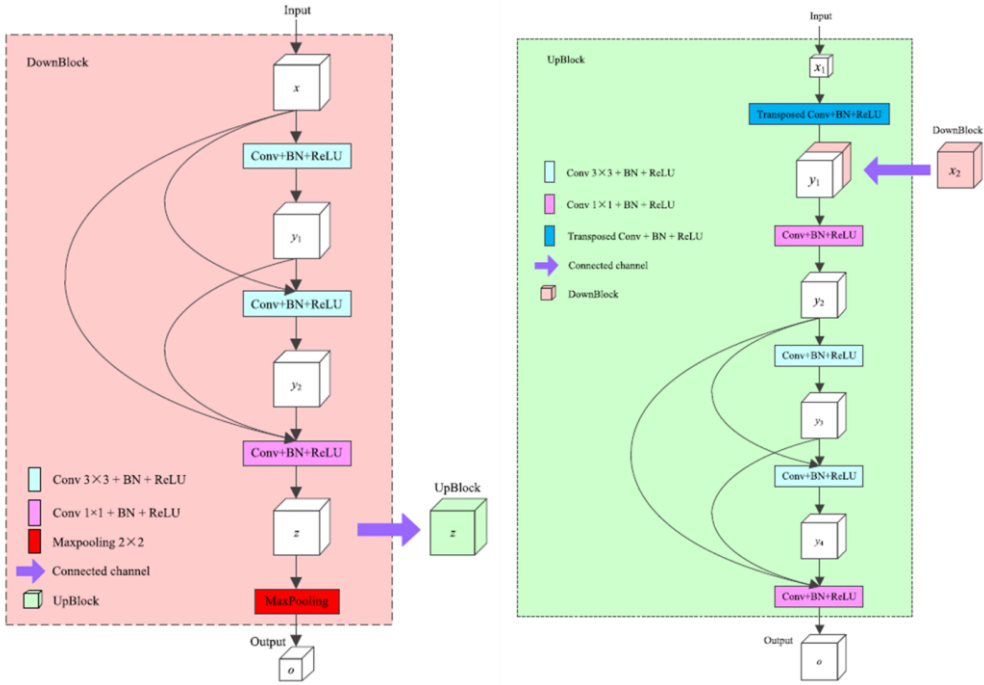


Figure 4. The dense DownBlock (left) and UpBlock (right).

Diakogiannis et al. (2020) proposed ResUNet-a using a UNet encoder-decoder backbone and the residual structure of ResNet. The deeper convolutional structures with residual connections allowed the model to achieve consistent training in the encoding path and remove gradient vanishing. With additional improvements such as applying an atrous convolution and a dice loss function, Diakogiannis et al. (2020) achieved top rank accuracy using VHR images, demonstrating significant improvement over state-of-the-art results. In this thesis, Sherrah (2016), Audebert et al. (2016), Dong et al., (2019), and Diakogiannis et al. (2020) are used as benchmarks to compare the proposed framework, especially under the condition of using limited labeled data.

2.2. Generative Adversarial Networks (GAN)

2.2.1. Introduction to GAN

GAN was proposed by Goodfellow et al. (2014) for image generation task and was trained by implementing a minimax game between the generator and discriminator (figure 5).

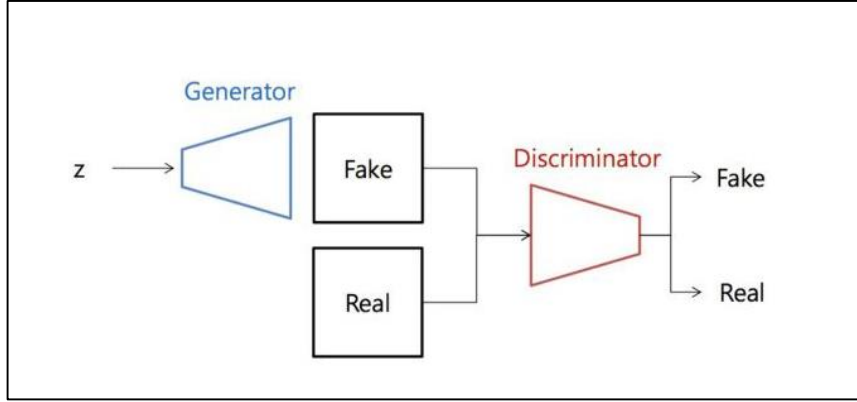


Figure 5. Overview of GAN structure. Fake data is generated from latent vector z by the generator. The discriminator tries to distinguish the fake data and real data as real or fake.

The objective function of GAN is defined as follows:

$$\min_G \max_D V_{GAN}(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (1)$$

where D and G denote the discriminator and generator, respectively. The generator map input variables z to fake image, and the fake image and real image x are fed into the discriminator. To maximize the objective function, the discriminator tries to map the real data to a value of one, and the fake image to a value of zero. In other words, the generator tries to make the discriminator fail by minimizing the objective function. In practice, the

function may not give sufficient gradient to train the generator, since the generator performs more poorly in comparison to the discriminator during the early few epochs. To address this problem, a modified function is practically employed for the generator:

$$\max_G V_{GAN}(G) = \mathbb{E}_{z \sim p_z(z)} \left[\log \left(D(G(z)) \right) \right] \quad (2)$$

Since GAN models are hard to train as evidenced by problems such as mode collapse, gradient vanishing and imbalanced training (Goodfellow, 2016), the modification of objective functions and network structure is a very active research area in an effort to overcome these problems. In particular, one representative research proposed alternate objective functions: Least Squares GAN (LSGAN) (Mao et al., 2017). LSGAN changes the original objective function based on binary cross entropy to a least squares-based function as follows:

$$\begin{aligned} \min_D V_{LSGAN}(D) = & \frac{1}{2} \mathbb{E}_{x \sim p_{data}(x)} [(D(x) - b)^2] \\ & + \frac{1}{2} \mathbb{E}_{z \sim p_z(z)} [(D(G(z)) - a)^2] \end{aligned} \quad (3)$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2} \mathbb{E}_{z \sim p_z(z)} [(D(G(z)) - c)^2] \quad (4)$$

where, a and b are the labels for fake data and real data, c is the value that the generator tries to make the discriminator “believe” in. The least squares function can penalize data to a greater extent from real samples, which pulls the data closer toward the decision boundary. This also allows the model to alleviate the gradient vanishing problem that occurs when learning with the original objective function.

In the original GAN model, the modes of the generated images cannot be controlled due to the unsupervised nature of GAN, thus resulting in poor quality and mode-collapsed outputs. In response to this problem, Mirza and

Osindero (2014) proposed Conditional GAN (CGAN) where additional information such as the class label of the image is added in the generator and the discriminator to control the generation process. The objective function of CGAN is defined as:

$$\min_G \max_D V_{CGAN}(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x|y)] + \mathbb{E}_{z \sim p_z(z)} \left[\log \left(1 - D(G(z|y)) \right) \right] \quad (5)$$

where y is ancillary information, inducing the generator to produce a fake image based on the extra information. CGAN introduced the first attempt to use class labels with GAN, consequently leading to the widespread usage of class information in many GAN applications including image translation.

2.2.2 Image Translation

GAN networks are continuously being used for a diversity of applications. One of the most popular tasks using GAN is image translation, which aims to map an image from a source domain to a target domain. For example, there are translations between edge-to-photo, aerial image-to-map, and real image-to-segment (figure 6). In image translation methods using GAN, there are two representative networks: the supervised learning-based Pix2Pix and the unsupervised learning-based CycleGAN. Isola et al. (2017) proposed Pix2Pix where the GAN architecture was applied for image translation in a supervised manner (figure 7). Since the conventional CNN-based image translation methods were used to formulate the task via a per-pixel regression, the outputs appeared to be blurred and unrealistic. To address this issue, Pix2Pix employed the adversarial loss of GAN to generate more realistic outputs. Also, supervised reconstruction loss using the L1 norm was employed with the adversarial loss to further train the model toward generating more similar results to the real samples.

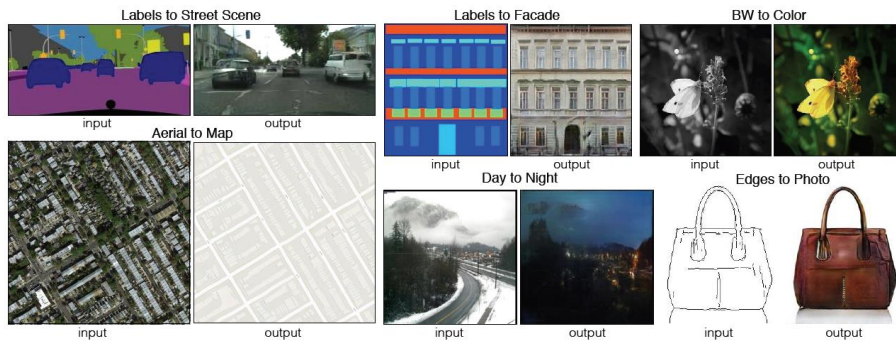


Figure 6. Example images for the image-to-image translation task (Isola et al., 2017).

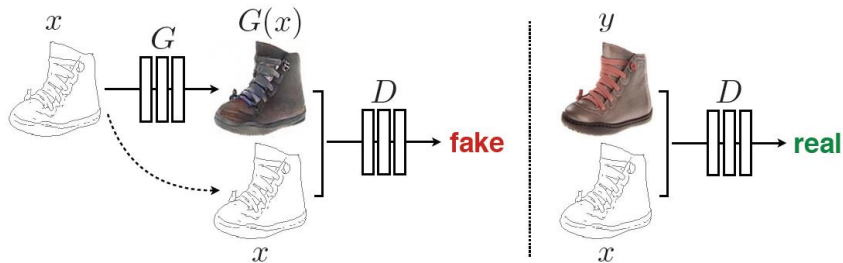


Figure 7. The structure of Pix2Pix to translate edges-to-photo (Isola et al., 2017). An image from the edges domain is translated to a fake photo image by the generator, while the discriminator tries to distinguish the two images.

One caveat is that Pix2Pix can only be trained by “paired” samples, but obtaining paired samples is difficult and sometimes impossible. To answer this issue, Zhu et al. (2017) proposed a novel method to train “unpaired” images using CycleGAN. To train an unpaired image for image translation, Zhu et al. (2017) exploited “cycle consistency loss”. The structure of CycleGAN is illustrated in figure 8. CycleGAN consists of two mapping functions ($G: X \rightarrow Y$, $F: Y \rightarrow X$) and two discriminators (D_X, D_Y). Zhu et al. (2017) introduced the cycle consistency loss from the intuition that if an image from a domain translated to the other and back again, the model should

be able to reconstruct the output at where the original image started. There are two cycle consistency losses: forward cycle consistency loss ($X \rightarrow G(X) \rightarrow F(G(x)) \approx X$) and backward cycle consistency loss ($Y \rightarrow F(Y) \rightarrow G(F(Y)) \approx Y$). The generators are trained toward reducing the cycle consistency losses, where paired images are not required. In addition, the adversarial loss of GAN is employed together for that the translated fake images cannot be distinguished from real images of the target domain. While CycleGAN outputs poor results when compared to supervised models, the model has inspired many researches in that unpaired images were utilized for model training in image translation studies.

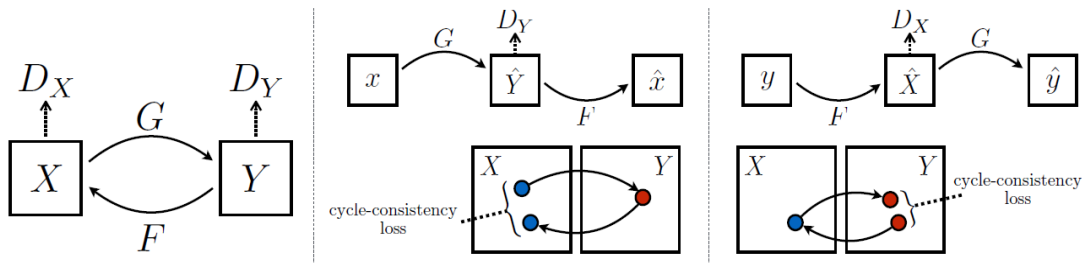


Figure 8. The workflow of CycleGAN (Zhu et al., 2017). CycleGAN contains two mapping functions (G and F) and two discriminators. The generators are trained so that the cycle consistency loss is minimized and the model calculates the difference between original data and reconstructed data.

2.2.3. GAN for Semantic Segmentation

The unique adversarial structure of GAN inspired many segmentation-based researches, since the generator (also referred to as the “translator”) can play a similar role to the classifier used in segmentation with models such as CNN. GAN has also been applied as an attached module in the adversarial function of conventional CNN-based classifiers or FCN-based segmentation models. Luc et al. (2016) first proposed a semantic segmentation model using

adversarial networks, where the discriminator encourages the generator to produce prediction maps that are hard to distinguish from ground truth maps (figure 9). Their results showed that the adversarial method improved semantic segmentation accuracy on the CV-based segmentation dataset. This segmentation method combining adversarial learning has also been applied to remotely-sensed imagery (Lin et al., 2017-a; Shi et al., 2018; Zhang et al., 2019;). In more detail, Shi et al. (2018) and Zhang et al. (2019) employed adversarial learning for binary segmentation of building footprint and roads, respectively, while Lin et al. (2017-a) conducted segmentation for a multi-class dataset. In these model architectures, the generator produces predicted images from remotely-sensed VHR images, while the discriminator tries to distinguish if the single class map is real with respect to the generated output and real label data.

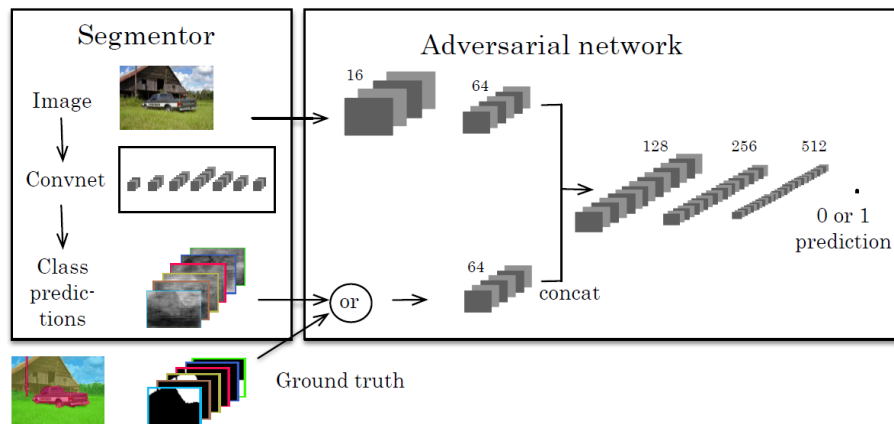


Figure 9. Overview the semantic segmentation approach using adversarial networks (Luc et al., 2016)

Beyond the simple application of the adversarial structure, modified GAN architectures have been studied for various applications of semantic segmentation. Zhang et al. (2018-a) proposed SegGAN where a pre-trained segmentation network was fitted into the GAN framework. They were

motivated by the fact that a predicted map from a segmentation network should have a strong correlation with the original image. For this purpose, the segmentation model predicted segmented images from the original image, and subsequently, the generator produced images from the predicted layers, while discriminator was used to distinguish original and fake images in competition with the generator. They originally utilized the GAN to reflect the correlation between the original image and label and achieved promising accuracy for semantic segmentation.

One of the main advantages of GAN is that the model can be trained using unlabeled images. Recent studies inspired by this advantage combined conventional supervised classification methods with the GAN structure to employ unlabeled data. Hung et al. (2018) proposed adversarial learning for semi-supervised semantic segmentation (figure 10). They designed a fully convolutional discriminator which enables semi-supervised learning through self-taught labeling. The semi-supervised loss is defined as:

$$\mathcal{L}_{semi} = - \sum_{\hat{h}, w} \sum_{c \sim \mathcal{C}} I(D(S(X_n))^{(\hat{h}, w)} > T_{semi}) \cdot \hat{Y}_n^{(\hat{h}, w, c)} \log(S(X_n)^{(\hat{h}, w, c)}) \quad (6)$$

where, I is an indicator for pseudo labeling based on threshold T_{semi} , and \hat{Y} is a self-taught label. The loss function is based on categorical cross-entropy for classification. Hung et al. (2018) argued that the confidence map produced by the discriminator from the unlabeled data provided a self-taught signal, and which led to refined segmentation results.

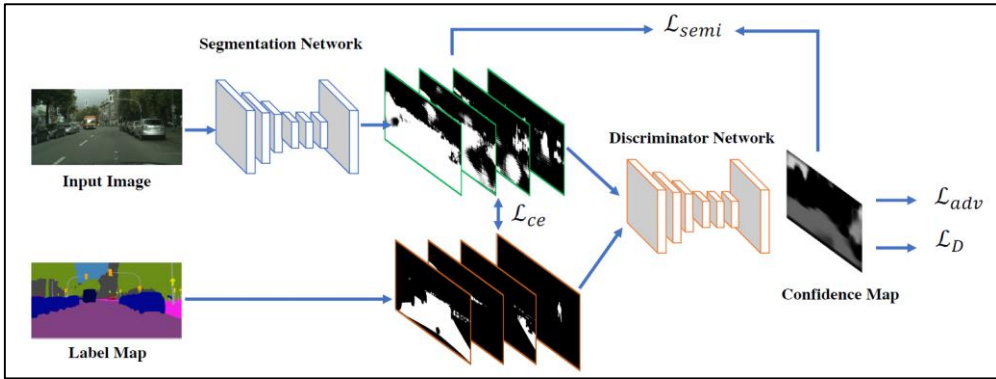


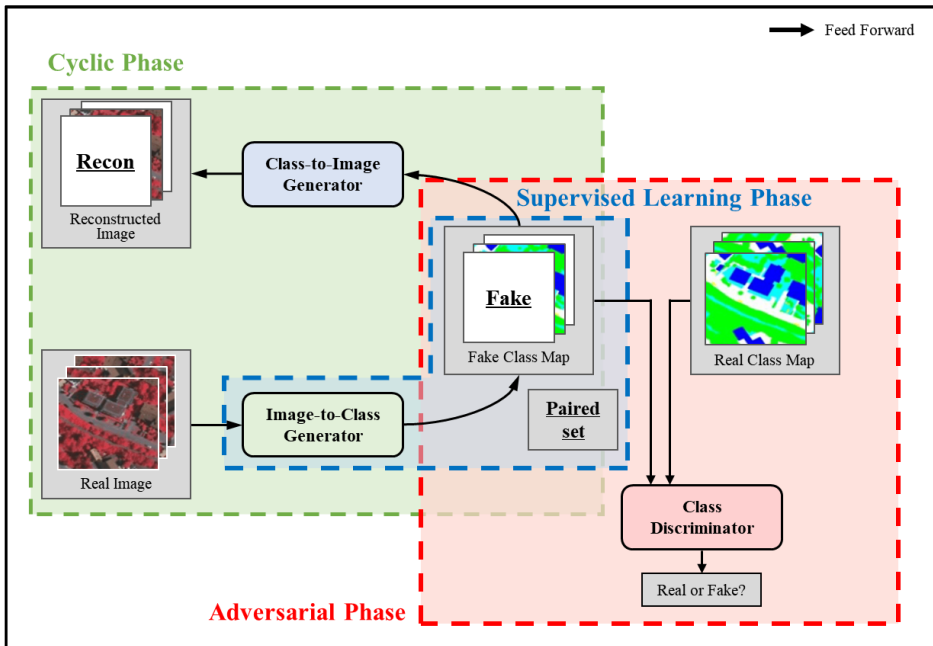
Figure 10. Overview of the semi-supervised semantic segmentation system with adversarial learning (Hung et al., 2018)

Meanwhile, Mondal et al. (2019) proposed semi-supervised segmentation algorithms using the unpaired image transfer capabilities of CycleGAN. This study employed the cycle consistency mapping for semantic segmentation and achieved improved segmentation performance in CV-based segmentation datasets such as PASCAL VOC and Cityscapes. However, in a study by Peng et al. (2020), the model by Mondal et al. (2019) was found to have some limitations that may lead to mode collapse problems leading to poor classification accuracy when applied to remote sensing image with different object properties. The semi-supervised framework proposed in this thesis modified the CycleGAN networks to adapt for remotely-sensed VHR images and the classification performance was compared to upper semi-supervised benchmarks (Hung et al., 2018 and Mondal et al., 2019).

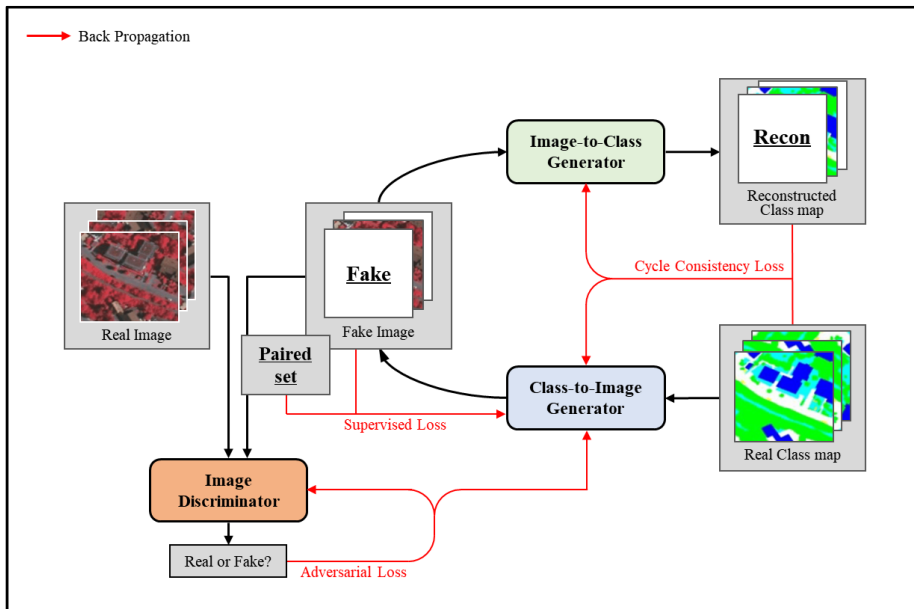
Chapter 3. Proposed Framework

This thesis proposed the VHR remote sensing image classification framework based on semi-supervised learning using a modified CycleGAN. The proposed models consist of two generators and two discriminators (figure 11): image-to-class generator G_{I2C} , class-to-image generator G_{C2I} , image discriminator D_I , class discriminator D_C . These four networks pass information of input data through feed-forward paths and are trained by back-propagation. The proposed framework can be divided into two parts including image-reconstructed cycle and class-reconstructed parts. In the respective parts, three phases are included: supervised learning phase, cyclic phase, and adversarial phase.

Section 3 is organized as follows: in Section 3.1., the modification of the original CycleGAN for semi-supervised VHR image classification is explained. In Section 3.2., the feed-forward path of the proposed framework is described. Section 3.3. presents the loss functions used to train the models through back-propagation. In Section 3.4., four network architectures using modified CycleGAN are explained.



(a)



(b)

Figure 11. The proposed semi-supervised framework using the modified CycleGAN. The framework includes two parts: (a) the image-reconstructed part and (b) class-reconstructed part. When displaying the structure, (a) shows the three phases of this framework and (b) shows the back-propagation process.

3.1. Modification of CycleGAN

This thesis was inspired by the cycle consistency concept of CycleGAN (Zhu et al., 2017) to establish a semi-supervised learning image classification framework. However, there are some gaps between the original CycleGAN and the purpose of this thesis. First, the original CycleGAN is only trained by unpaired data in an unsupervised manner, without employing the labeled data. While this can be a significant approach that the labeled data is not required, unsupervised learning for image classification is still challenging to obtain a stable accuracy. Second, the purpose of the original CycleGAN is image translation, while this thesis aims to perform image classification. Generators of CycleGAN are based on a style transform network that has different structure and loss function from image classification methods.

To tackle these two gaps, this thesis combined the supervised learning-based classification method and the structure of original CycleGAN based on cycle consistency employing unpaired images. In the cyclic and adversarial phases from the original CycleGAN, unlabeled images and unpaired class maps can be employed by an unsupervised manner. Additionally, in this thesis, labeled images and paired class map train the two generators in the supervised learning phase through the cross-entropy loss for image classification. With the modification in the framework's architecture and loss function, this thesis replaced the generator network which acts as an image classifier with a UNet-based segmentation network for image classification. Since remotely-sensed images contain many homogeneous and small-sized objects, skip-connection of UNet helps the generator to consider detailed spatial information leading to accurate classification in small objects. Consequently, supervised learning-based image classification network and cross-entropy loss were combined with the unpaired data-based CycleGAN for remotely sensed VHR image classification.

3.2. Feed-forward Path of the Proposed Framework

3.2.1. Cyclic Phase

First, in the image-reconstructed cycle part, both an unlabeled image set X_{unla} and a labeled image set X_{la} are used. Respectively, the real VHR images (x^{real}) are fed into G_{I2C} to generate fake class maps ($y^{fake} = G_{I2C}(x^{real})$), and thereafter, the fake class maps generated from G_{I2C} pass through G_{C2I} translating to reconstructed images ($x^{recon} = G_{C2I}(G_{I2C}(x^{real}))$). Since this cyclic feed-forward path does not require a ground truth map, unlabeled images can be fed into the two generators. Also, the labeled image set is used in this cycle without ground truth to retain more training images and to integrate with the supervised learning phase. Second, only the class map Y_{la} belonging to the labeled dataset are used in the class-reconstructed cycle path, because there is no class map in the unlabeled dataset. Reversely, G_{C2I} first feeds real class maps (y^{real}) generating fake images ($x^{fake} = G_{C2I}(y^{real})$), and then, the fake images are translated to reconstructed class map ($y^{recon} = G_{I2C}(G_{C2I}(y^{real}))$) through G_{I2C} .

3.2.2. Adversarial Phase

Fake images and class maps produced in the cyclic phase are used to train in an adversarial learning way with the real images and class maps. The adversarial phase aims to set the generator and discriminator so they compete against each other to produce more real samples. For the image domain, D_I receives both the real images (x^{real}) and the fake images from G_{C2I} (x^{fake}) to produce a prediction score that can be used to assess real or fake images. Likewise, receiving the real class maps (y^{real}) and the fake class maps from G_{I2C} (y^{fake}), D_C tries to distinguish whether inputs are real or fake class maps.

3.2.3. Supervised Learning Phase

In the supervised learning phase, only labeled images and their corresponding ground truth labels are used. G_{I2C} which acts as the main classifier receives the real images, generating the same prediction maps as the fake class maps ($\hat{y} = G_{I2C}(x_{la}^{real}) = y_{la}^{fake}$). Generated prediction maps \hat{y} and real class maps y_{la}^{real} are used in back-propagation by reducing supervised loss between the two maps. At the same time, G_{C2I} translates the real class maps to the fake images $x_{la}^{fake} = G_{C2I}(y_{la}^{real})$. This feed-forward path in the supervised learning phase is computationally included in the front section of the cyclic phase.

3.3. Loss Function for Back-propagation

Similar to the feed-forward path, the loss function for back-propagation can be divided into three parts: cycle consistency loss, adversarial loss, supervised learning loss. First, the supervised loss is defined as follows:

$$\mathcal{L}_{sup}(G_{I2C}, G_{C2I}) = \lambda_C^{sup} \mathcal{L}_C^{sup} + \lambda_I^{sup} \mathcal{L}_I^{sup} \quad (7)$$

$$\mathcal{L}_C^{sup}(G_{I2C}) = \mathbb{E}_{x \sim X_{Ia}, y \sim Y_{Ia}} [CEE(y, G_{I2C}(x))] \quad (8)$$

$$\mathcal{L}_I^{sup}(G_{C2I}) = \mathbb{E}_{x \sim X_{Ia}, y \sim Y_{Ia}} [|x - G_{C2I}(y)|_1] \quad (9)$$

where, λ is a weight coefficient for the corresponding loss term, and is determined experimentally. The total supervised loss \mathcal{L}_{sup} is defined as the weighted sum of two losses, \mathcal{L}_C^{sup} and \mathcal{L}_I^{sup} . Here, \mathcal{L}_C^{sup} is a supervised classification loss between the real class and predicted maps, where categorical cross-entropy is employed to compute the error between the two maps. Categorical cross-entropy is defined as:

$$CCE(y, \hat{y}) = - \sum_i^C y_i \log(\sigma(\hat{y}_i)) \quad (10)$$

where C is the number of classes and σ is the sigmoid activation function. \mathcal{L}_I^{sup} is the supervised image generation loss defined as L1 norm difference between the real images and generated images. Since the pixel values of generated and real images do not denote prediction probability like class maps but the brightness of the specific channel, the L1 norm difference is employed rather than categorical cross-entropy to measure the error between those two images. With the supervised loss, the two generators are trained toward generating results more similar to the reference data.

Second, cycle consistency loss is employed to train the two generators to

reconstruct the original image or class maps in an unsupervised manner. The loss is the weighted sum of the three losses with the labeled image set, unlabeled image set, and ground truth maps. The three losses are described as follows:

$$\mathcal{L}_{cycle}(G_{I2C}, G_{C2I}) = \lambda_C^{cycle} \mathcal{L}_C^{cycle} + \lambda_{la}^{cycle} \mathcal{L}_{la}^{cycle} + \lambda_{unla}^{cycle} \mathcal{L}_{unla}^{cycle} \quad (11)$$

$$\mathcal{L}_{la}^{cycle}(G_{I2C}, G_{C2I}) = \mathbb{E}_{x \sim X_{la}} \left[\|x - G_{C2I}(G_{I2C}(x))\|_1 \right] \quad (12)$$

$$\mathcal{L}_{unla}^{cycle}(G_{I2C}, G_{C2I}) = \mathbb{E}_{x \sim X_{unla}} \left[\|x - G_{C2I}(G_{I2C}(x))\|_1 \right] \quad (13)$$

$$\mathcal{L}_C^{cycle}(G_{I2C}, G_{C2I}) = \mathbb{E}_{y \sim Y_{la}} \left[CEE(y, G_{I2C}(G_{C2I}(y))) \right] \quad (14)$$

Cycle consistency loss aims to measure the error between original images or class maps and reconstructed images or class maps. Since the losses use original inputs and reconstructed outputs passing through two generators successively without ground truth data, unsupervised learning can be implemented.

To reduce the cycle consistency loss, the two generators try to preserve information from the original inputs though the corrected reference for respective networks is unknown. For error-measuring functions, categorical cross-entropy is used for class maps, while the L1 norm is used for the images.

Lastly, the adversarial loss is propagated backward in the networks for the generator and discriminator to compete with each other. The loss is described as follows:

$$\mathcal{L}_{adv}(G_{I2C}, G_{C2I}, D_C, D_I) = \lambda_C^{adv} \mathcal{L}_C^{adv} + \lambda_I^{adv} \mathcal{L}_I^{adv} \quad (15)$$

$$\begin{aligned}\mathcal{L}_C^{adv}(D_C) &= \mathbb{E}_{y \sim Y_{la}} \left[(1 - D_C(y))^2 \right] \\ &+ \mathbb{E}_{x \sim X_{la}, X_{unla}} \left[(D_C(G_{I2C}(x)))^2 \right]\end{aligned}\quad (16)$$

$$\mathcal{L}_C^{adv}(G_{I2C}) = \mathbb{E}_{x \sim X_{la}, X_{unla}} \left[(1 - D_C(G_{I2C}(x)))^2 \right] \quad (17)$$

$$\begin{aligned}\mathcal{L}_I^{adv}(D_I) &= \mathbb{E}_{x \sim X_{la}, X_{unla}} \left[(1 - D_I(x))^2 \right] \\ &+ \mathbb{E}_{y \sim Y_{la}} \left[(D_I(G_{C2I}(y)))^2 \right]\end{aligned}\quad (18)$$

$$\mathcal{L}_I^{adv}(G_{C2I}) = \mathbb{E}_{y \sim Y_{la}} \left[(1 - D_I(G_{C2I}(y)))^2 \right] \quad (19)$$

The adversarial losses are based on the least square loss function of LSGAN which tries to improve learning stability in the CycleGAN. To reduce the losses, discriminators are trained toward mapping the real inputs to a value of one and the fake inputs generated by generators to zero value. Simultaneously, the generators try to generate fake samples that are mapped to a value of one by the discriminators. The total adversarial loss can be divided into two losses in the class map domain and image domain. In the class map domain, D_C and G_{I2C} try to minimize the respective loss functions, allowing to generate more plausible prediction maps from the generator. Likewise, G_{C2I} tries to generate more realistic images to deceive D_I in the image domain. In addition, unlabeled images are used when D_I maps real images to a value of one and D_C maps fake class map from G_{I2C} to a value of zero. For unlabeled images, cycle consistency loss is used in model training and supervised loss cannot be used. Since the cycle consistency loss only considers errors from reconstructed data, the adversarial loss is employed to control fake data in the middle of the cycle path.

3.4. Proposed Network Architecture

3.4.1. Generator Architecture

The model architecture of an image-to-class generator which acts as a classifier is illustrated in figure 12. In this thesis, UNet is introduced in the G_{I2C} architecture and the backbone of the encoder is replaced with an EfficientNetB1 network. The EfficientNetB1-based encoder includes one stem block and seven MBConv blocks, containing the depthwise separable convolution and inverted residual block. After the contracting path through the EfficientNet-based encoder, the reduced resolution of the feature map is recovered using up-sampling operators with concatenating encoder blocks. Skip connections enable the combination of coarse high-level features and fine low-level features. The input size of G_{I2C} is defined as $(2, 256, 256, 3)$, which denotes a batch size of 2, the patch window size of 256, and channels corresponding to 3 bands. At the end of the encoder, input images are contracted to $(2, 64, 64, 1024)$ while the feature map is recovered to a size of $(2, 256, 256, n_c)$ through the decoder, where n_c represents the number of classes. The softmax activation function (Equation 20) is assigned to the last layer to output a classification probability vector using the features from the preceding layer.

$$\sigma_{softmax}(\vec{y})_i = \frac{e^{z_i}}{\sum_j^{n_c} e^{z_j}} \quad (20)$$

The overall architecture of the class-to-image generator is similar to the former generator. While G_{C2I} also introduces the EfficientUNet architecture, input class maps have six channels which is the same as the number of classes, but differs from the size of the input channel of G_{I2C} . Since the encoder uses pre-trained weights, the encoder is restricted to inputting only 3 channels. To address this limitation, two EfficientNet-based backbones are combined before decoding in G_{C2I} . On the contrary, the output size of G_{C2I} is $(2, 256,$

256, 3) to translate class maps to the real image domain having three bands. The sigmoid function (Equation 21) is employed as the last activation function to map values from a range of zero to one like real images.

$$\sigma_{sigmoid}(z) = \frac{1}{1 + e^z} \quad (21)$$

3.4.2. Discriminator Architecture

For the discriminators architecture, this thesis used the PatchGAN discriminator introduced in Pix2Pix as depicted in Figure 13. PatchGAN aims to distinguish if each $N \times N$ patch from an input image is real or fake, while the previous discriminator tries to distinguish if the whole image is real or fake. Instead of using the entire image, the discrimination of a patch unit allows the model to capture more high frequency information while using smaller training parameters. The two discriminators, G_C and G_I , possess the same architecture, but differ by their number of input channels, where $G_C: (2, 256, 256, 6)$, $G_I: (2, 256, 256, 3)$. The output size of the two discriminators is $(2, 30, 30, 1)$, which output a probability to determine whether each patch is real or fake.

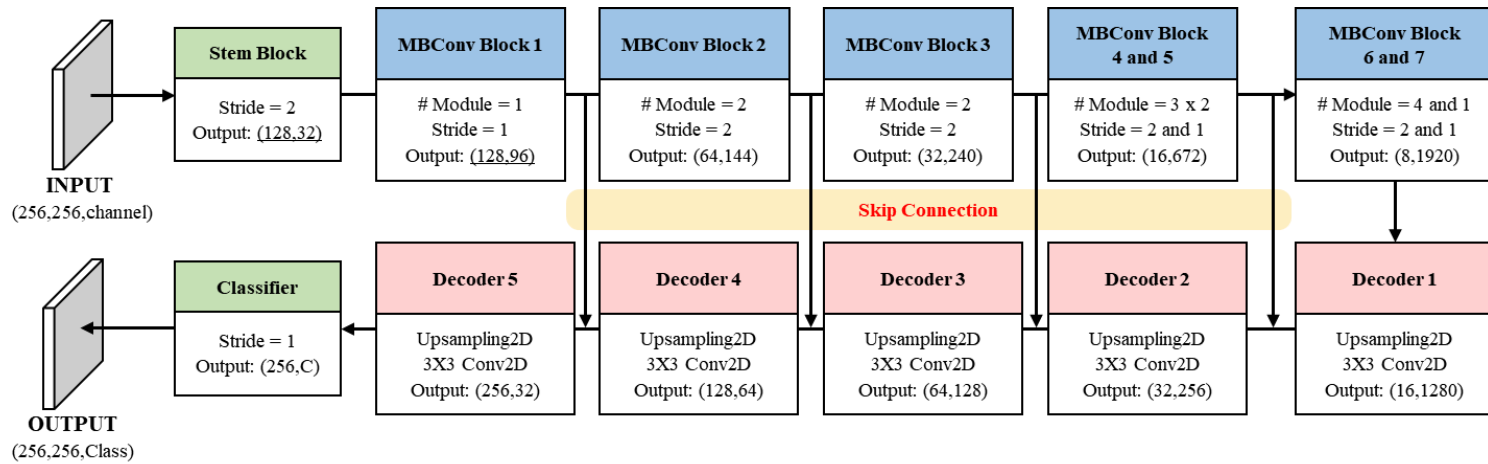


Figure 12. EfficientUNet architecture employed in this thesis. It includes EfficientNet-based encoder and upsampling-based decoder. The window size of input is 256 pixels, and the feature maps of encoder and decoder are connected by skip connection.

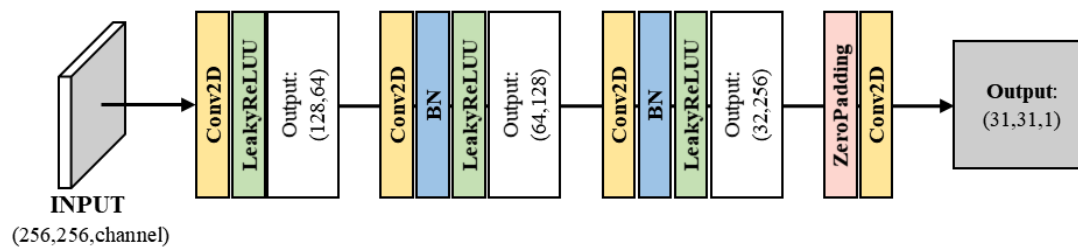


Figure 13. PatchGAN-based discriminator architecture employed in this thesis.

Chapter 4. Experimental Design

In this section, the detailed experimental design for the proposed framework is explained. First, in Section 4.1., An overall workflow (figure 14) is described including the explanation of three experiments. In Section 4.2., a description of the VHR image dataset used in the proposed framework is given. Section 4.3. outlines the details with regards to model and data implementation, including experimental data settings, data augmentation, and model training parameters. Lastly, Section 4.4. provides the metrics used to evaluate the accuracy of the proposed framework.

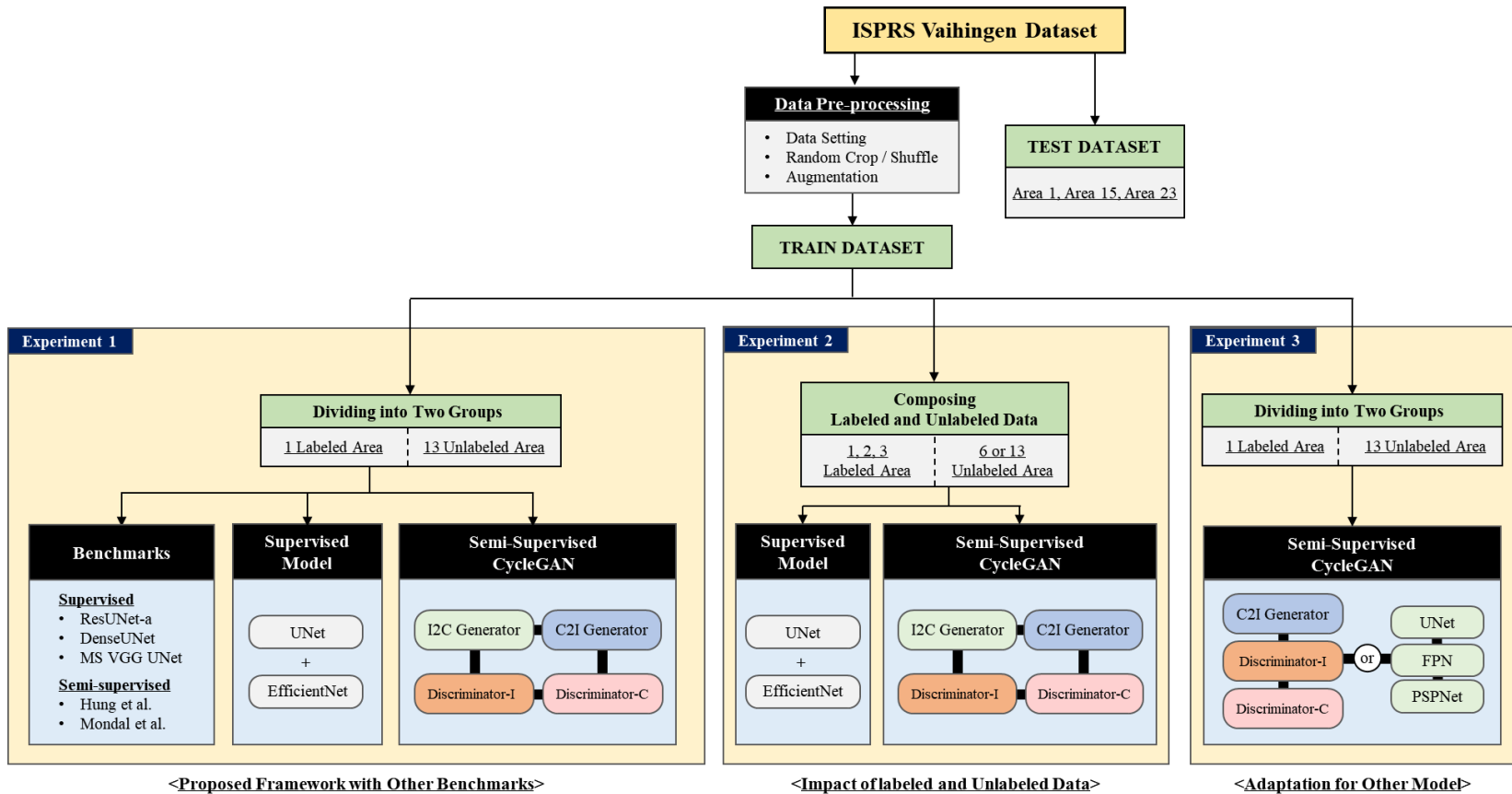


Figure 14. Overall workflow of this thesis including data setting and three experiments.

4.1. Overall Workflow

An overall workflow is described in figure 14. First, the full dataset is divided into train and test datasets based on fixed test sites. Thereafter, the train dataset goes through pre-processing including data setting, random crop, random shuffle, and data augmentation. Pre-processed train dataset is used in the tree experiments conducted in the proposed framework.

(1) Experiment 1: Classification Performance of the Proposed Framework with Other Benchmarks.

First, the proposed semi-supervised framework is trained and evaluated in comparison to five other benchmark models and the supervised learning method. In this experiment, a fixed number of labeled (one patch) and unlabeled data (thirteen patches) are used in the proposed framework and other benchmarks. The benchmarks consist of three supervised methods for remote sensing image classification and two semi-supervised classification methods based on GAN. The three supervised learning-based methods including multi-scale VGG with UNet (Audebert et al., 2016), DenseUNet (Dong et al., 2019), ResUNet-a (Diakogiannis et al., 2020) had used the same dataset as this thesis and performed the VHR image classification. For two semi-supervised benchmarks (Hung et al., 2018 and Mondal et al., 2019), GAN-based approach was utilized for image classification by semi-supervised learning manner.

While Hung et al. (2018) utilized GAN for additionally labeling to unlabeled data based on the discriminator’s confidence signal, Mondal et al. (2019) employed GAN for regularization of limited train dataset based on a large amount of unlabeled data. For this purpose, Mondal et al. (2019) used CycleGAN which is same approach as this thesis. However, there are some degradation in classification performance of this method when applied to remote sensing data due to its unique characteristics including small-sized

objects, inter-class similar spectral properties, and extremely fewer train samples. These features can lead to unstable model training and mode collapse in complex models such as GAN. This thesis is advantageous in terms of the model's efficiency and training stability than Mondal et al. (2019)'s method.

Table 1 shows the difference between the two methods utilizing CycleGAN. Mondal et al. (2019)'s study performed CV-based image classification, where the minimum number of train data is 899 patches. Since labeling ground truths for the small-sized and heterogeneous objects containing in VHR remote sensing imagery is very difficult and time-consuming, the minimum number of training data cannot be easily attained and much fewer training samples were usually available. The small amount of remote sensing training image leads the complex GAN model to be trained unstably and produce deteriorated classification results. To address this problem in remote sensing domain, this thesis replaced the ResNet-based image translation network of Mondal et al. (2019) with EfficientNet-based generator, which can greatly lower the model complexity through efficient model design. The total number of the EfficientNet-based network's parameters is 12,577,862 which is smaller than Mondal et al. (2019)'s generator (51,506,409). A Simple and efficient model is suitable for remote sensing images that are hard to obtain a large amount of data and have inter-class similar properties. With the high model efficiency, two detailed modifications are applied to improve training stability. In this thesis, labeled images were together used in the cyclic phase with unlabeled data. It allows assisting the cyclic training which is unguided by ground truth. In addition, image loss is calculated by L1 norm that is stable to outlier samples, while Mondal et al. (2019) uses L2 loss. It can alleviate the effect of errors in tree's texture, shadow, and details in small objects frequently generated by class-to-image translator.

Table 1. Comparison of the proposed framework and Mondal et al. (2019)’s method

	Mondal et al. (2019)	Proposed Framework
Generator	● ResNet-based transform network	● EfficientUNet
Network’s Parameter	● 51,506,409	● 12,577,862
Stability	● Unlabeled data only is used in cyclic phase. ● L2 loss in image loss	● Both Labeled and Unlabeled Data are used in cyclic phase. ● L1 loss in image loss

(2) Experiment 2: Impact of Labeled and Unlabeled Data.

In the second experiment, the impact of labeled and unlabeled data is confirmed by controlling the number of labeled and unlabeled data. For the train dataset containing sixteen patches, three patches are used with the paired ground truth maps, while the other thirteen patches are used for unlabeled image patches. The number of the labeled data increase from only one patch to three patches. At the same time, the number of unlabeled data is selected as the total patches of thirteen or half of them, six patches. The controlled train data is used to train the EfficientUNet (supervised) and EfficientUNet + modified CycleGAN (semi-supervised). This experiment aims to investigate the difference according to the various composition of labeled and unlabeled data for the proposed semi-supervised framework.

(3) Experiment 3: Adaptation for Other Classification Model.

Lastly, to see whether the proposed semi-supervised methods can improve classification performance of other segmentation networks besides UNet, two other networks, feature pyramid network (FPN) and pyramid scene parsing network (PSPNet), are also applied. The image-to-class translator of the modified CycleGAN is replaced with FPN and PSPNet, while EfficientNet backbone is fixed. The three results from UNet, FPN, and PSPNet using the

proposed semi-supervised framework are compared to the results of supervised UNet, FPN, and PSPNet. This experiment aims to confirm the adaptation of the proposed framework for other classification models, and consequently, to validate that the proposed semi-supervised CycleGAN module is not over-fitted to a specific UNet model and robust to the selection of classification networks.

FPN proposed in Lin et al. (2017-b) aims to leverage a pyramidal feature hierarchy of convolution layers. The construction of FPN includes a bottom-up path and a top-down path with connecting paths between them in order to link the bottom-up high-resolution features with the bottom-up high-level feature maps (figure 15). Feature maps of different resolutions in the top-down path are resampled to the same size and then merged into one prediction layer. This multi-scale kernel allows FPN to perform segmentation which is robust to the size of objects.

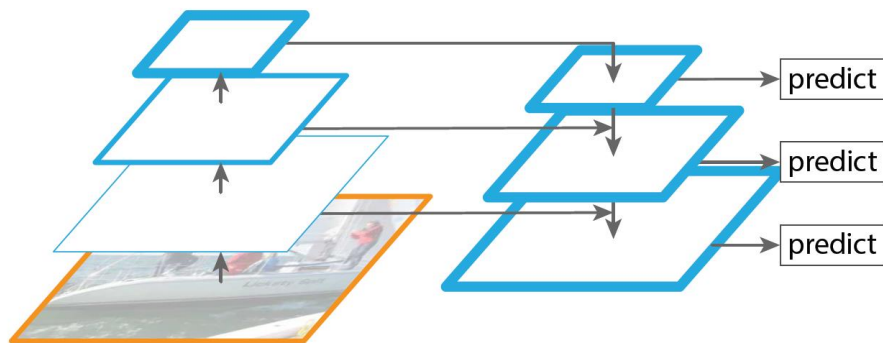


Figure 15. Overview of FPN architecture (Lin et al., 2017-b).

In Zhao et al. (2017), PSPNet was proposed to incorporate global features extending the pixel-level features to the global pyramid pooling features (figure 16). When pooling with a large size of kernel, the layer can extract global contextual information. By applying the pyramidal pooling, the size of the receptive field can be expanded, which can help to alleviate issues of

misunderstanding relationship and confusion categories.

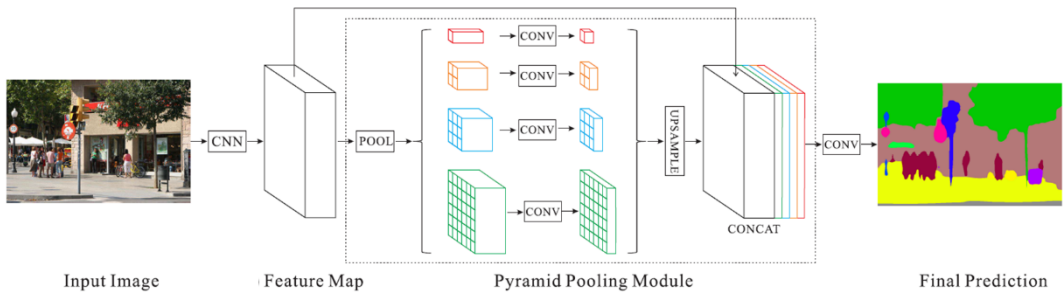
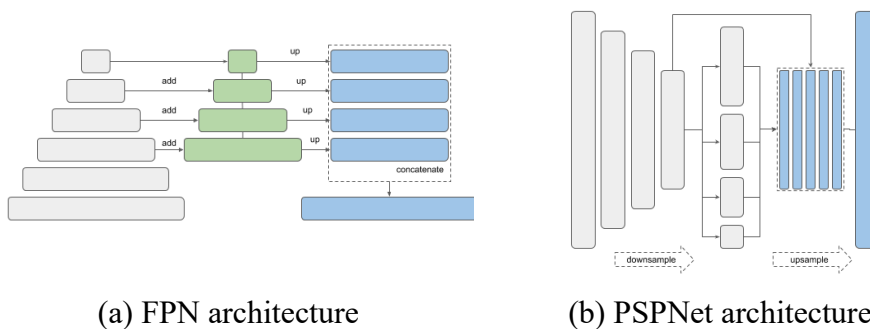


Figure 16. Overview of PSPNet architecture (Zhao et al., 2017).

In this thesis, FPN and PSPNet were also applied with U-Net in the semi-supervised CycleGAN architecture to validate the applicability of the proposed framework to other classification models. Figure 17 shows the modified classification models to apply the proposed GAN module. Like UNet, EfficientNet was used as the encoder for additional two classification models, and all models were compared using the supervised and semi-supervised methods.



(a) FPN architecture

(b) PSPNet architecture


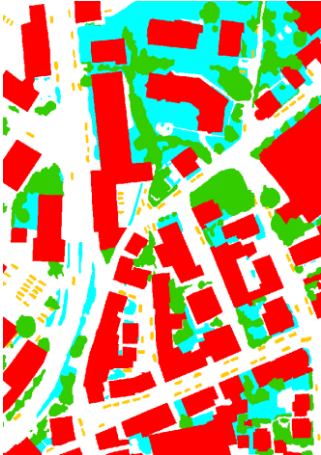
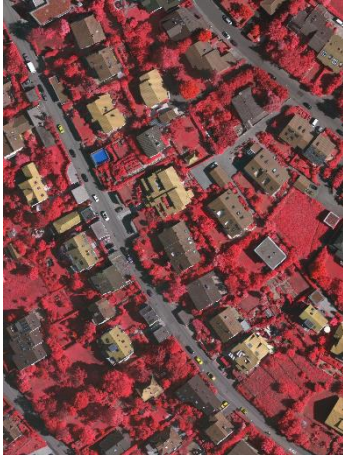
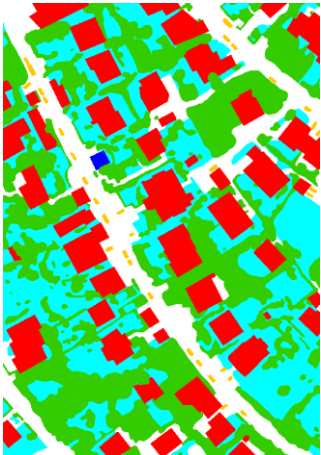

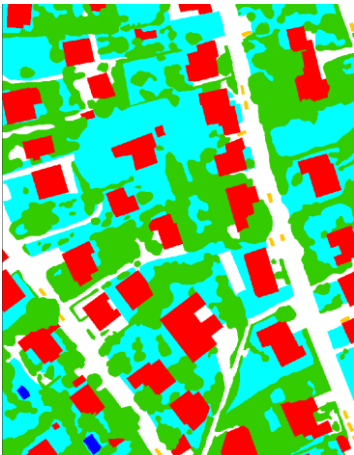
Figure 17. The modified architectures of FPN and PSPNet employed in this thesis. In the encoder of two models, EfficientNet is used like UNet.

4.2. Vaihingen Dataset

To evaluate the performance of the proposed framework, this thesis used a remotely-sensed VHR image dataset acquired using airborne sensors. The dataset was taken over the city of Vaihingen (Germany), which was disseminated for the 2D Semantic Labeling Contest by the International Society for Photogrammetry and Remote Sensing (ISPRS). The dataset is composed of ground truth maps and true ortho-photos with a spatial resolution of 0.09 m and three bands of near-infrared, red, and green channels. Ground truth data consists of six classes including impervious surfaces, building, low vegetation, tree, car, and clutter/background. The clutter/background class contains water bodies, containers, tennis courts, and swimming pools and is usually not of interest for segmentation in urban areas. Further, clutter/background and car classes are scarce compared to building, tree, and low vegetation classes.

The full dataset consists of nineteen image patches: areas 1, 3, 4, 5, 7, 10, 11, 13, 15, 17, 21, 23, 24, 26, 28, 30, 32, 34, 37. These patches are divided test dataset and train datasets. Among the patches, the three image patches of area 1, 15 and 23 are used for testing which contain different properties of objects with each other, especially in area 1. The false color images and ground truth maps of the test sites are illustrated in Table 2. Through the ground truth map of the Table 2, small-sized buildings are visible in area 15 and area 23, while area 1 contains many large-sized buildings. Tree and low vegetation classes constitute a large proportion of test sites, especially in area 15 and area 23. Also, in the middle left of area 15, tree and low vegetation classes appear to be heterogeneously mixed. While area 23 contains fewer samples from car class, area 1 includes the most cars and even a small parking lot.

Table 2. Real images of false color and ground truth maps in the three test sites: Area 1, Area 15, Area 23.

Sites	Real Image	Ground Truth
Area 1		
Area 15		
Area 23		

- Impervious Surface
- Building
- Low Vegetation
- Tree
- Car
- Clutter

4.3. Implementation Details

Data augmentation is used to increase the training dataset and improve the diversity of the dataset. Random scaling, flipping, brightness/contrast, and random cropping are performed. Since the generator model’s input size sets as 256 x 256 pixels, the original image needs to be cropped into training patches. Instead of simply dividing images in the training dataset to the identical 256 x 256 patches, this thesis crops the original image into 512 x 512 patches and then crops randomly to sub-patches of 256 x 256 pixels in the data augmentation process. As a consequence, models receive subtly different images and labels, leading to better model generalization and superior performance in the test step.

For this thesis, the configurations of the experimental environment are as follows: CPU: Intel® Core™ i5-6600 CPU @ 3.30GHz, GPU: an NVIDIA GeForce RTX 2070 SUPER with 8GB of memory, system: Cuda-10.0. and Cudnn-7.6.3, deep learning library: Tensorflow-gpu: 2.0.0. The adaptive moment estimation (Adam) optimizer was used. Different learning rates for generators and discriminators were used to balance the four networks with an exponential decay rate scheduling method (Table 3).

Table 3. Learning rate schedules of the four networks

Networks	Initial Learning Rate	Exponential Decay Rate	Decay Steps
Image-to-Class Generator	0.0005	0.96	500
Class-to-Image Generator	0.0003	0.96	500
Image Discriminator	0.0001	-	-
Class Discriminator	0.0001	-	-

4.4. Metrics for Quantitative Evaluation

This thesis used the overall accuracy (OA), intersection over union (IoU), and F1 score, which are conventional metrics for multi-class image classification. OA is defined as:

$$OA = \frac{\sum_i^C Pixels_i^{corrected}}{Pixels^{all}} \quad (22)$$

where, C is the number of classes. OA denotes the proportion of pixels that are correctly classified in the image. OA is an intuitive measurement and is easy to calculate, but is a poor indicator providing limited information for performance. This thesis, therefore, included F1-score and IoU which are popular metrics for evaluating semantic segmentation performance. F1-score and IoU are defined as follows:

$$F1 = \frac{2(Precision \cdot Recall)}{Precision + Recall} \quad (23)$$

$$IoU = \frac{Area(A \cap B)}{Area(A \cup B)} \quad (24)$$

where A and B refers to reference and predicted regions, respectively. Precision and recall for F1-score can be calculated as:

$$Precision = \frac{TP}{TP + FP} \quad (25)$$

$$Recall = \frac{TP}{TP + FN} \quad (26)$$

where TP, FP and FN denote the number of true positive pixels, false positive pixels and false negative pixels, respectively. F1-score and IoU were calculated for each class, and the class-weighted averages of F1-score and IoU were also measured.

Chapter 5. Results and Discussion

In this chapter, the results and discussion of the proposed framework from the comparative experiments are presented. Section 5.1 presents the classification results of the proposed framework in three test sites. In Section 5.2, the overall classification results are introduced to compare the proposed framework with other benchmarks including supervised and semi-supervised methods. Section 5.3 discusses the impact of labeled and unlabeled data composition for semi-supervised learning. In Section 5.4, the effect of the semi-supervised CycleGAN module is validated in terms of cycle consistency. Lastly, Section 5.4 investigates the influence of using the proposed GAN method for other common classification models including FPN and PSPNet.

5.1. Performance Evaluation of the Proposed Framework

Figure 18 shows the overall classified maps from the proposed semi-supervised framework and ground truth maps in the three test sites. In addition, Table 4 shows the confusion matrix including the number of the predicted and true pixels for each class with recall and precision scores. First, it was visually confirmed that the boundaries and shapes of buildings were classified properly in three test sites with showing the quantitatively highest recall and precision scores (0.866 and 0.877) among the six classes. Especially, even if there are unique buildings having various shapes and sizes in area 1, the predicted results were obtained stably regardless of the spatial features. In the Vaihingen dataset, buildings and their rooftop spectrally have high inner-class spectral homogeneity, resulting in consistent classification performance. Also, the most misclassified class of true building pixels was impervious surface which has more similar spectral information than other classes.

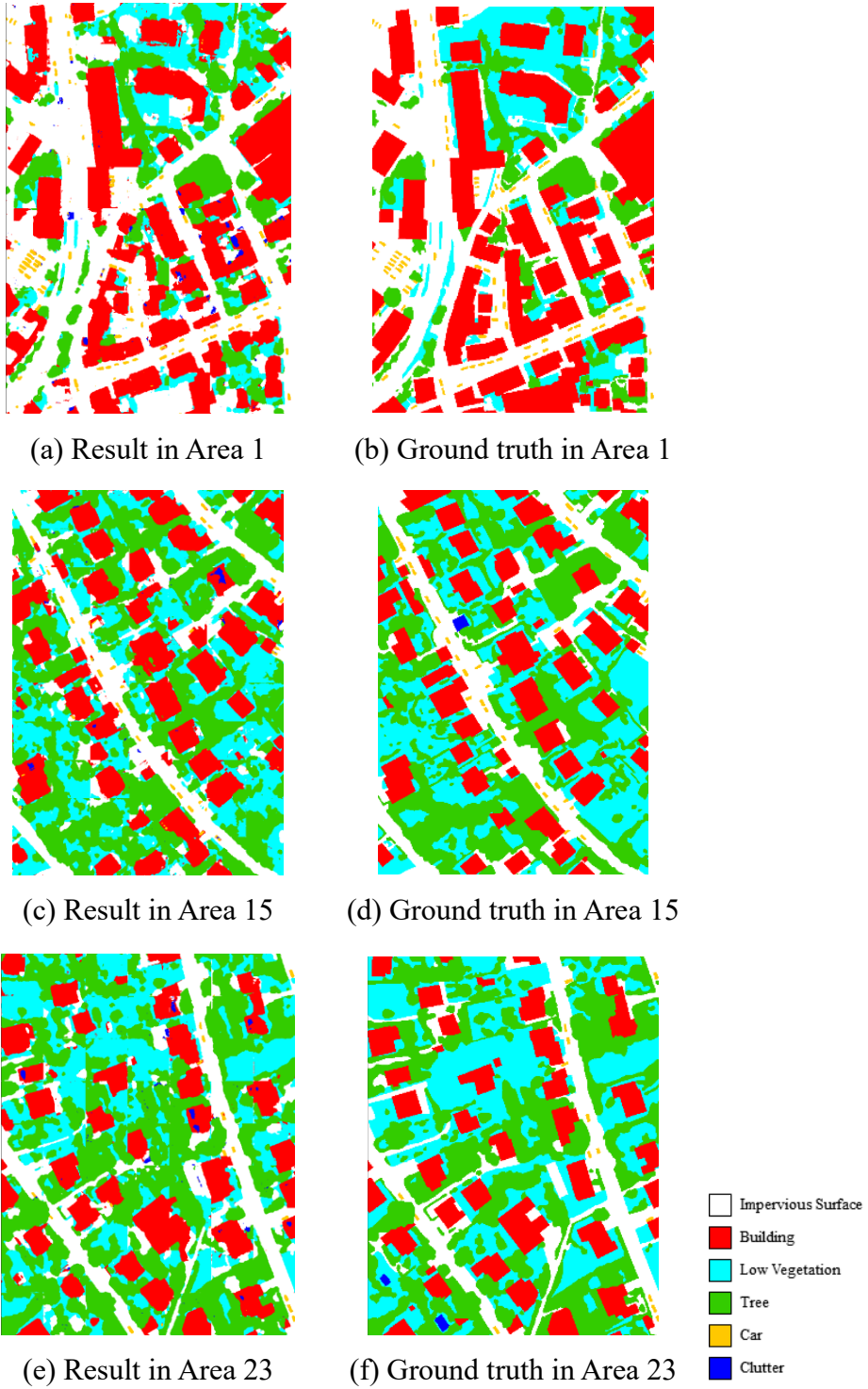


Figure 18. The predicted maps by the proposed semi-supervised framework and ground truth map in three test sites.

Table 4. Confusion matrix including the number of the predicted and true pixels for each class with recall and precision scores.

Predict \ True	Impervious surface	building	Low vegetation	Tree	Car	Clutter	Total	Recall
Impervious surface	2,548,676	207,281	149,829	132,257	14,859	8,607	3,061,509	0.832
building	333,227	2,951,023	59,723	30,413	4,007	30,154	3,408,547	0.866
Low vegetation	380,724	162,793	1,787,141	699,805	1,337	2,683	3,034,483	0.589
Tree	68,756	33,658	427,155	3,136,839	225	1,979	3,668,612	0.855
Car	31,763	4,658	931	953	75,162	2,249	115,716	0.650
Clutter	9,884	3,951	263	33	810	0	14,941	0.000
Total	3,373,030	3,363,364	2,425,042	4,000,300	96,400	45,672		
Precision	0.756	0.877	0.737	0.784	0.780	0.000		

Unlike the building class, results of low vegetation and tree class visually represented obscure boundaries with recording the lower precision scores (low vegetation: 0.737, tree: 0.784). In particle, two classes were mixed each other due to their similar spectral properties, also quantitatively confirmed in Table 4. Among the 637,901 pixels misclassified to low vegetation, 427,155 pixels (67.0%) were misclassified to tree. Also, Among the 863,461 pixels misclassified to tree, 699,805 pixels (81.0%) were misclassified to low vegetation. The heterogeneity and spectral similarity between the two classes lead to this mixed misclassification. While the proposed model is robust to the spatial properties such as shape and size, it is dependent on the spectral features resulting that the accurate decision boundary between low vegetation and tree is not obtained. Also, even if the car class has a very small size, many objects were properly classified. The main errors were omission errors to impervious surfaces, and it is induced by shadows of surrounding buildings leading the impervious surface and car to have similar spectral characteristics. The results of the clutter class were totally misclassified (precision: 0, recall: 0), since it has very low inner-class homogeneity and produces fewer train

samples. These unique true clutter pixels were mainly misclassified to impervious surfaces or buildings, and also disturbed the classification performance of the two classes leading to some commission errors.

5.2. Comparison of Classification Performance in the Proposed Framework and Benchmarks

The classification results of the proposed framework, supervised, and semi-supervised benchmarks were evaluated in the three test areas by three metrics: OA, F1-score and mIoU. Table 5 shows the classification results of ResUNet-a (Diakogiannis et al., 2020), DenseUNet (Dong et al., 2019), Multi-scale VGG with UNet (Audebert et al., 2016), EfficientUNet, EfficientUNet + CycleGAN, Hung et al. (2018), and Mondal et al. (2019). To elaborate on the configuration of the experiments, Audebert et al. (2016), Dong et al. (2019), Diakogiannis et al. (2020), and EfficientUNet methods are conducted using supervised learning, while the others are based on semi-supervised learning. In all experiments, the number of used labeled patches was fixed to one patch. And in the case of semi-supervised learning, thirteen unlabeled patches were used with the labeled patch. The highest score among the supervised or semi-supervised methods is highlighted in bold text, and the second is underlined.

Table 5. Overall classification accuracy in OA, F1-score, mIoU of benchmarks and the proposed framework

Leaning	Model	Area 1			Area 15			Area 23		
		OA	F1	mIoU	OA	F1	mIoU	OA	F1	mIoU
Supervised	ResUNet-a (Diakogiannis et al., 2020)	0.728	0.720	0.569	0.732	0.718	0.575	0.677	0.670	0.511
	DenseUNet (Dong et al., 2019)	0.737	0.729	0.582	0.721	0.717	0.568	0.715	0.715	0.564
	Multi-scale VGG with UNet (Audebert et al., 2016)	<u>0.731</u>	<u>0.724</u>	0.572	<u>0.735</u>	<u>0.723</u>	<u>0.582</u>	<u>0.727</u>	<u>0.719</u>	<u>0.570</u>
	EfficientUNet	0.730	<u>0.724</u>	<u>0.573</u>	0.767	0.759	0.625	0.743	0.737	0.594
Semi-supervised	EfficientUNet + CycleGAN	0.796	0.795	0.666	0.786	0.782	0.651	0.784	0.780	0.647
	Mondal et al., 2019	<u>0.777</u>	<u>0.778</u>	<u>0.642</u>	0.747	0.747	0.604	0.738	0.738	0.590
	Hung et al., 2018	0.763	0.760	0.617	<u>0.784</u>	<u>0.781</u>	<u>0.647</u>	<u>0.772</u>	<u>0.768</u>	<u>0.630</u>

In the three test sites, the proposed semi-supervised framework achieved the highest score of OA, F1-score, and mIoU in comparison to the benchmark models. The combined EfficientNet backbone and UNet structure proposed in this thesis for VHR image classification yielded an OA of 0.681 in area 1, 0.756 in area 15, and 0.723 in area 23. The baseline model was improved by the proposed semi-supervised CycleGAN method. The use of the proposed semi-supervised learning increased OA by 0.066 in area 1, 0.019 in area 15, and 0.041 in area 23.

Among the supervised methods, the highest OA was recorded by DenseUNet in area 1, while EfficientUNet was superior in area 15 and area 23. The poor results in supervised learning-based benchmarks were caused primarily by the limited amount of training data which is an inherent problem of VHR image classification, suggesting that the model may not have been trained sufficiently and could indicate overfitting. Since dense connections in the DenseUNet allow the model to be trained by connecting the gradient for back-propagation directly, relatively better classification results were achieved in area 1 having some different characteristics from the training dataset such as large-sized buildings well as a small amount of trees and low vegetation. For this problem, the use of semi-supervised learning vastly improved the accuracy in area 1 by regularizing the classifier through a large number of unlabeled data.

The three semi-supervised learning-based methods mostly yielded higher accuracy scores in the three test sites. In more detail, the method proposed by Mondal *et al.* achieved better classification results than that of Hung *et al.* when applied in area 1, whereas the opposite results occurred in area 15 and area 23. The method by Mondal *et al.* (2019) is based on consistency for regularization, while the method by Hung *et al.* (2018) employs pseudo labeling to create additional labels with high confidence. On the other hand, this thesis confirmed that the proposed semi-supervised framework achieved

the highest scores compared to the other two semi-supervised methods in all test sites.

Figure 19~21 respectively shows the close-up view of the three test sites. Upon visual inspection on buildings in the results, the predicted building segmentation results generated by the proposed framework are filled in and have sharper boundaries compared to other benchmark models. Besides generic building shapes in area 15 and area 23, buildings of unique shapes or sizes present in area 1 also were better segmented by using the proposed framework in comparison to the other benchmarks.

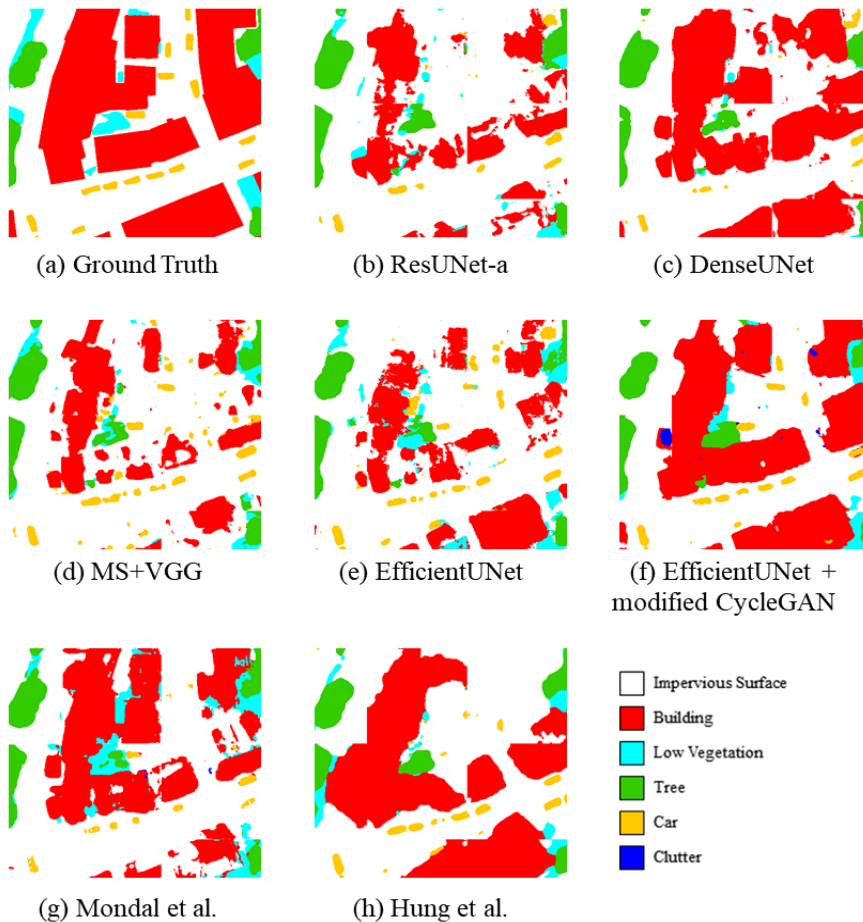


Figure 19. Close-up view of ground truth, the results map by benchmarks, and the proposed framework in Area 1

Additionally, while generic building shapes were better predicted by using the EfficientUNet method (figure 21), DenseUNet yielded more similar predictions of buildings in area 1 to the ground truth over the EfficientUNet methods due to the dense connections (figure 19).

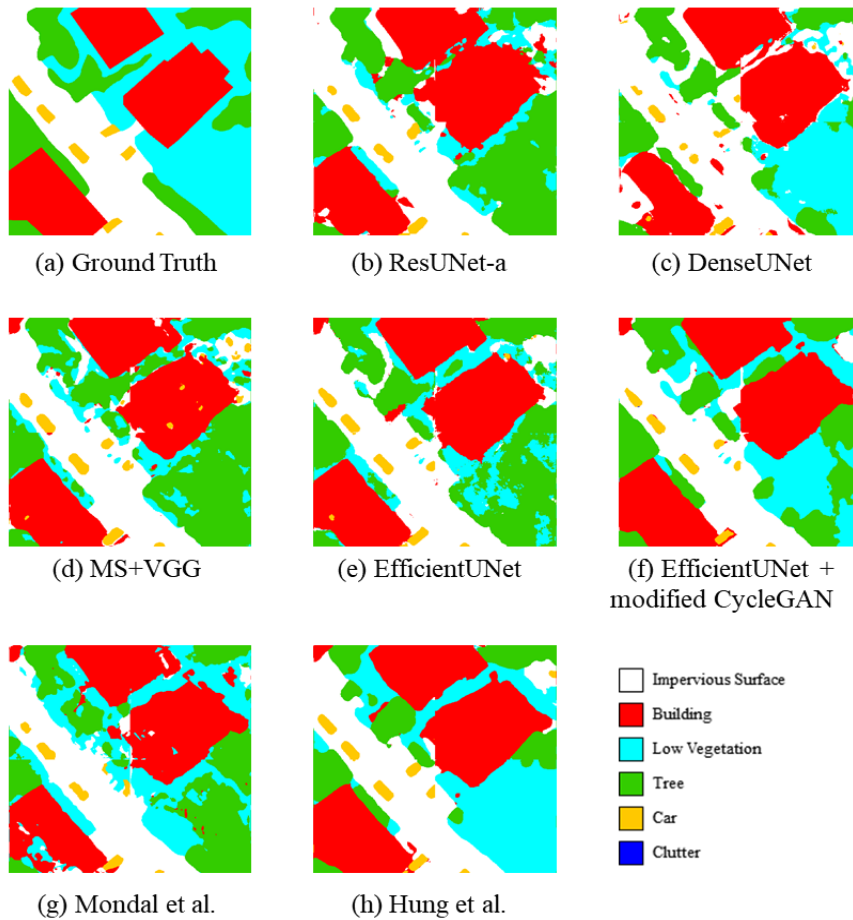


Figure 20. Close-up view of ground truth, the results map by benchmarks, and the proposed framework in Area 15

In the case of small-sized classes such as cars, MS+VGG, Hung et al, EfficientUNet, and the proposed semi-supervised framework generally exhibited satisfactory prediction maps. Moreover, DenseUNet yielded poorer results, and MS+VGG achieved better results than both the DenseUNet and ResUNet-a. This result can be attributed to the multi-scale ensemble in the

MS+VGG method, which helped to consider different sizes of spatial context and smoothen the shapes of segments. EfficientUNet also correctly extracted a large number of cars, also achieving similar results using additional CycleGAN method.

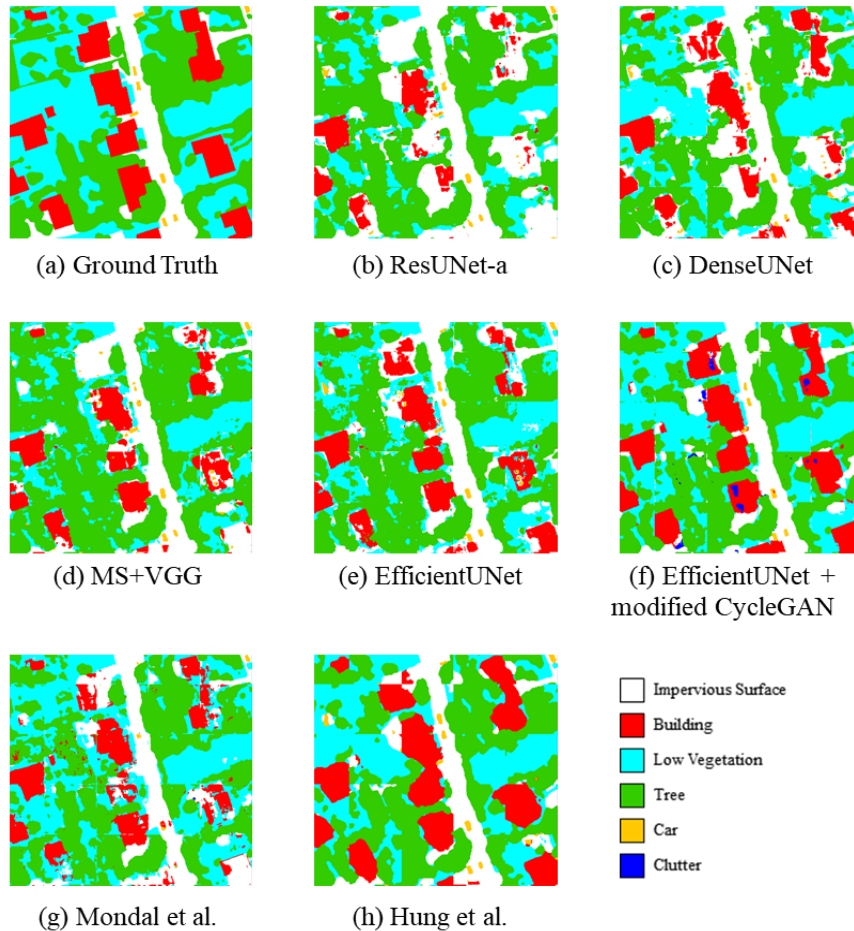


Figure 21. Close-up view of ground truth, the results map by benchmarks, and the proposed framework in Area 23

Since low vegetation and tree share many similar spectral features, those two classes were often misclassified for each other. Especially in the heterogeneous mixture of low vegetation and trees in area 23, the prediction

results returned several significant differences (figure 21). On the other hand, Hung et al. (2018)'s method and the proposed framework showed better classified maps in comparison to other benchmarks for regions of relatively homogeneous distribution of low vegetation and tree classes.

5.3. Impact of Labeled and Unlabeled Data for Semi-supervised Learning

Table 6 and figure 22 show the experimental results according to the composition of labeled and unlabeled data. The number of used labeled patches was set from one patch to three patches, while six or thirteen unlabeled patches were used. Figure 22 shows mIoU scores as a graph, and Table 4 shows F1-score and OA results from the experiments. In Table 4, “supervised” refers to the EfficientUNet model, while “semi-supervised” refers to the EfficientUNet + modified CycleGAN model. As illustrated in figure 22, the semi-supervised method achieved better mIoU scores than the supervised method. Among the semi-supervised methods, better results were obtained when using six unlabeled patches than when using 13 unlabeled patches. Notably, there was a smaller increase in mIoU when utilizing more unlabeled patches in comparison to when using more labeled data. When only one labeled patch was used, increasing in the number of unlabeled patches from 6 to 13 improved mIoU score by 0.030 (area 1), 0.023 (area 15) and 0.029 (area 23). In contrast, when 2 or 3 labeled patches were used, the increase in mIoU was relatively smaller from 0.007 to 0.018. This result showed that the proposed semi-supervised method achieves a clear increase in accuracy given the sufficient amount of unlabeled data compared to labeled data.

In terms of the number of labeled data used for the experiments, the classification accuracy can be assumed to increase as the number of labeled patches increases. However, in the case of using 2 to 3 labeled patches, especially in area 15, there was no increase in classification accuracy. This result can be explained since the third labeled patch was biased to other labeled patches, the result suffered from over-fitting with poor classification performance in some areas.

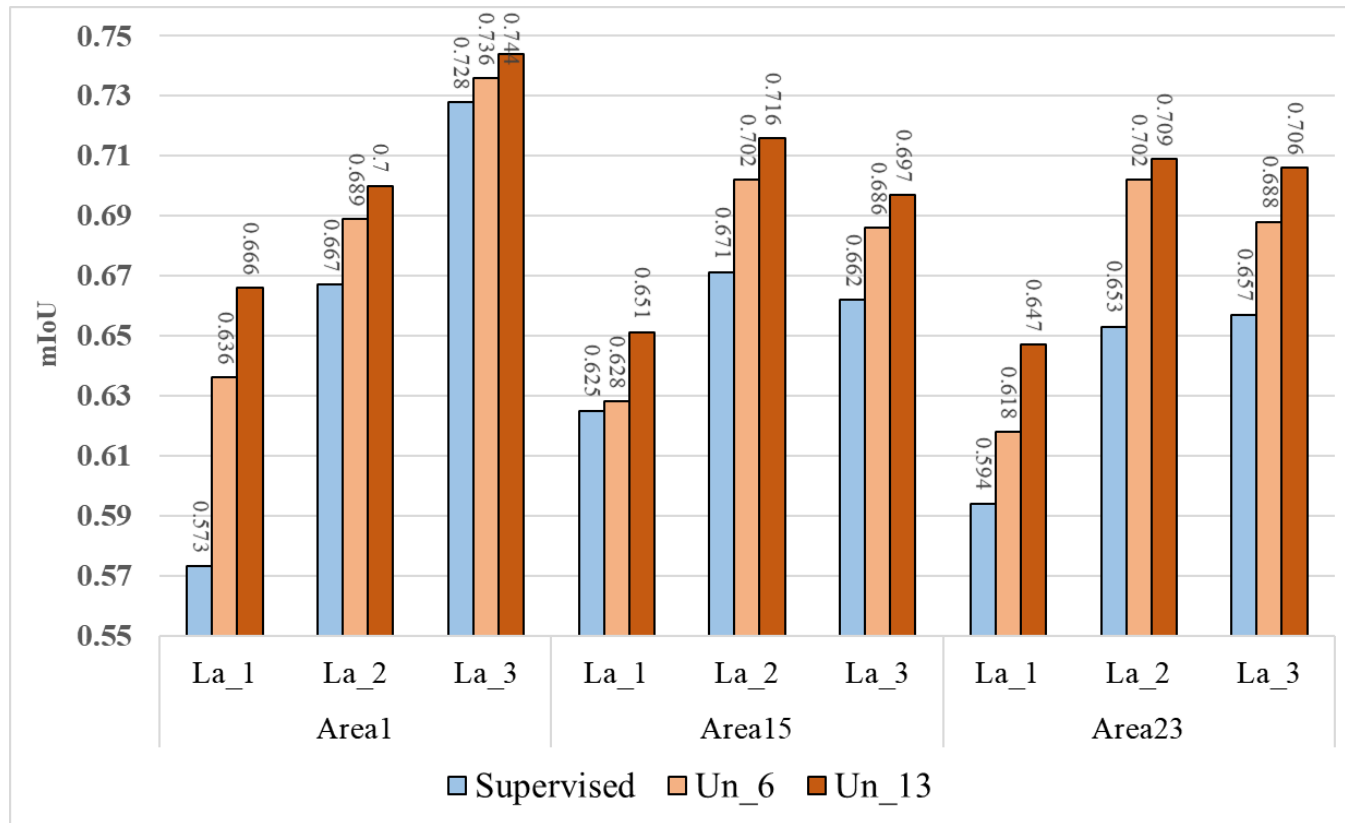


Figure 22. mIoU in three test sites according to the number of labeled and unlabeled data.

Table 6. F1-score and OA in three test sites according to the number of labeled and unlabeled data.

Area 1		Labeled 1 patches		Labeled 2 patches		Labeled 3 patches	
		F1	OA	F1	OA	F1	OA
Supervised		0.724	0.730	0.769	0.794	0.840	0.838
Semi Supervised	Unlabeled 6 patches	0.772	0.770	0.813	0.813	0.844	0.844
	Unlabeled 13 patches	0.795	0.796	0.820	0.819	0.851	0.850
Area 15		Labeled 1 patches		Labeled 2 patches		Labeled 3 patches	
		F1	OA	F1	OA	F1	OA
Supervised		0.759	0.767	0.797	0.798	0.789	0.792
Semi Supervised	Unlabeled 6 patches	0.764	0.766	0.821	0.822	0.808	0.807
	Unlabeled 13 patches	0.782	0.786	0.831	0.829	0.817	0.818
Area 23		Labeled 1 patches		Labeled 2 patches		Labeled 3 patches	
		F1	OA	F1	OA	F1	OA
Supervised		0.737	0.743	0.784	0.784	0.786	0.787
Semi Supervised	Unlabeled 6 patches	0.757	0.762	0.820	0.820	0.811	0.809
	Unlabeled 13 patches	0.780	0.784	0.827	0.825	0.824	0.824

5.4. Cycle Consistency in Semi-supervised Learning

Unlabeled images were mainly employed in the cyclic phase of recovering the information in original images through a sequence of two generators without labels. Figure 23 shows the difference between the supervised and semi-supervised methods during model training. X-axis means epoch number in the training process, and Y-axis is the validation accuracy of mIoU. The two graphs show a clearly different trend, where the accuracy of the semi-supervised method appears to fluctuate more, but reaches higher accuracy values than the supervised method. While the supervised model was more over-fitted to the training data yielding a limited and stable validation accuracy, the semi-supervised model strived to learn the properties of the numerous unlabeled data and reached higher accuracy maxima, albeit through a fluctuating trend.

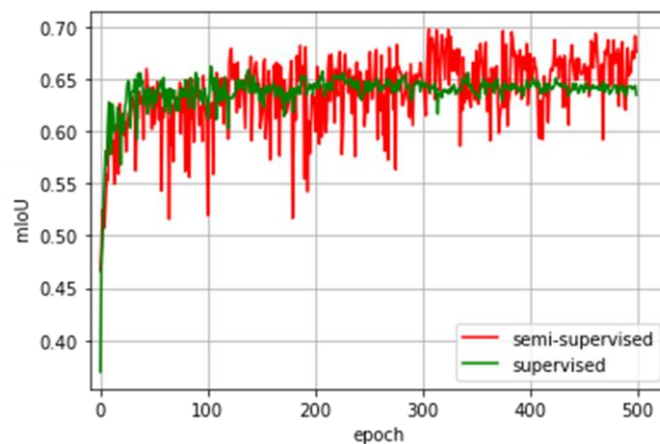
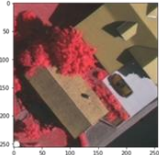
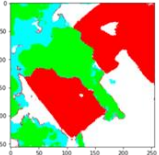
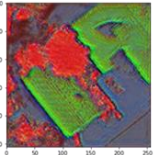
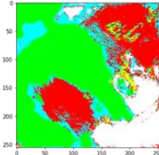
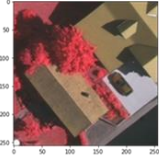
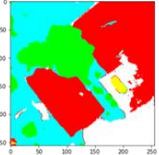
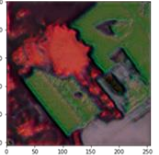
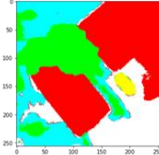
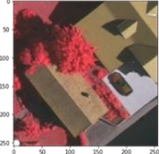
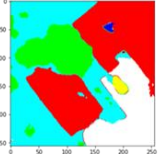
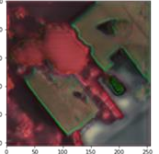
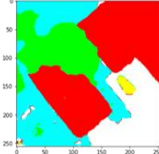
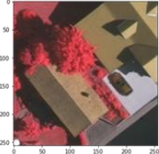
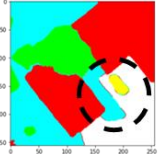
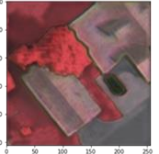
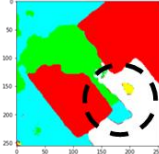
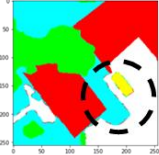




Figure 23. The validation accuracy graph during model training of EfficientUNet (supervised) and EfficientUNet + modified CycleGAN (semi-supervised).

Table 7. The real image, fake class map, reconstructed image from semi-supervised method, and fake class map from supervised method with ground truth according to epoch.

Epoch	Image-Class-Image Cycle			Supervised
	Real Image	Fake Class map	Reconstructed Image	Fake Class map
1				
10				
50				
Best				
Ground Truth				

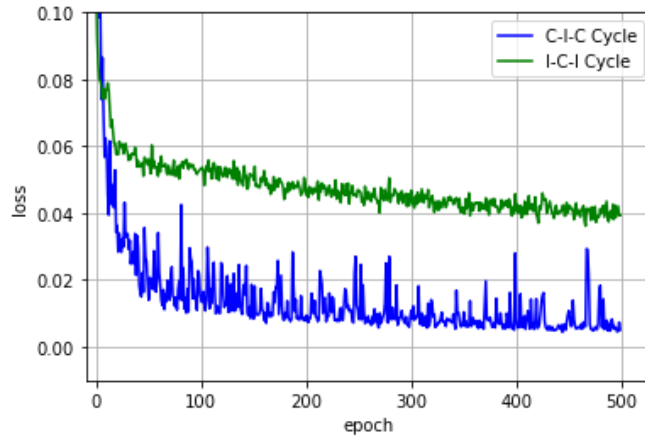


Figure 24. The cycle consistency loss graph of two cycle parts, image-class-image cycle (image-reconstructed cycle) and class-image-class cycle (class-reconstructed cycle) during model training of the proposed framework.

Table 7 shows the outputs of the image-class-image cyclic phase including real images, fake class maps, the reconstructed image of the semi-supervised method, and the fake class maps of the supervised method for several epochs. As model training progresses, the features of the reconstructed image such as color, shape, texture, and shadow grow to be more similar to those of the original image. Especially in the best epoch, the texture of trees, the color of building, and the shadows on the rooftop were better represented in comparison to the early epochs. For better recovery of the original image, the image-to-class generator was trained to preserve the information of the original image leading to highlighted edges of the segments and generated clearer segments. In table 7, the building segments were gradually filled with sharp edges, and the segmented results from the semi-supervised method were more similar to the ground truth in comparison to those produced from the supervised method. Similarly, when using the cycle consistency loss, it can be confirmed that both image-class-image and class-image-class cycle consistency losses gradually decreased during model training (figure 24),

meaning that both image and class map were increasingly better reconstructed by the two generators.

5.5. Adaptation of the GAN Framework for Other Classification Models

In this thesis, additional classification models, FPN and PSPNet, were applied to the proposed framework to replace the UNet backbone. Table 8 shows the F1-score and mIoU results of the three classification models and the results from adding the modified CycleGAN. The results demonstrated that the proposed framework can be adapted to the three classification models successfully and can improve classification accuracy. In particular, results from area 1 improved for all three models by about 0.090. In more detail, varying levels of increases in accuracy were observed based on the choice of the classification model. The increase in accuracy was less for PSPNet in comparison to that of UNet and FPN for area 15 and area 23.

Table 8. F1-score and mIoU of three classification models in supervised and semi-supervised learning.

	Area 1		Area 15		Area 23	
	F1	mIoU	F1	mIoU	F1	mIoU
UNet	0.724	0.573	0.759	0.625	0.737	0.594
+ modified CycleGAN	0.795	0.666	0.782	0.651	0.780	0.647
FPN	0.701	0.546	0.744	0.607	0.731	0.587
+ modified CycleGAN	0.774	0.638	0.772	0.635	0.764	0.625
PSPNet	0.703	0.547	0.734	0.589	0.719	0.569
+ modified CycleGAN	0.772	0.637	0.734	0.592	0.729	0.587

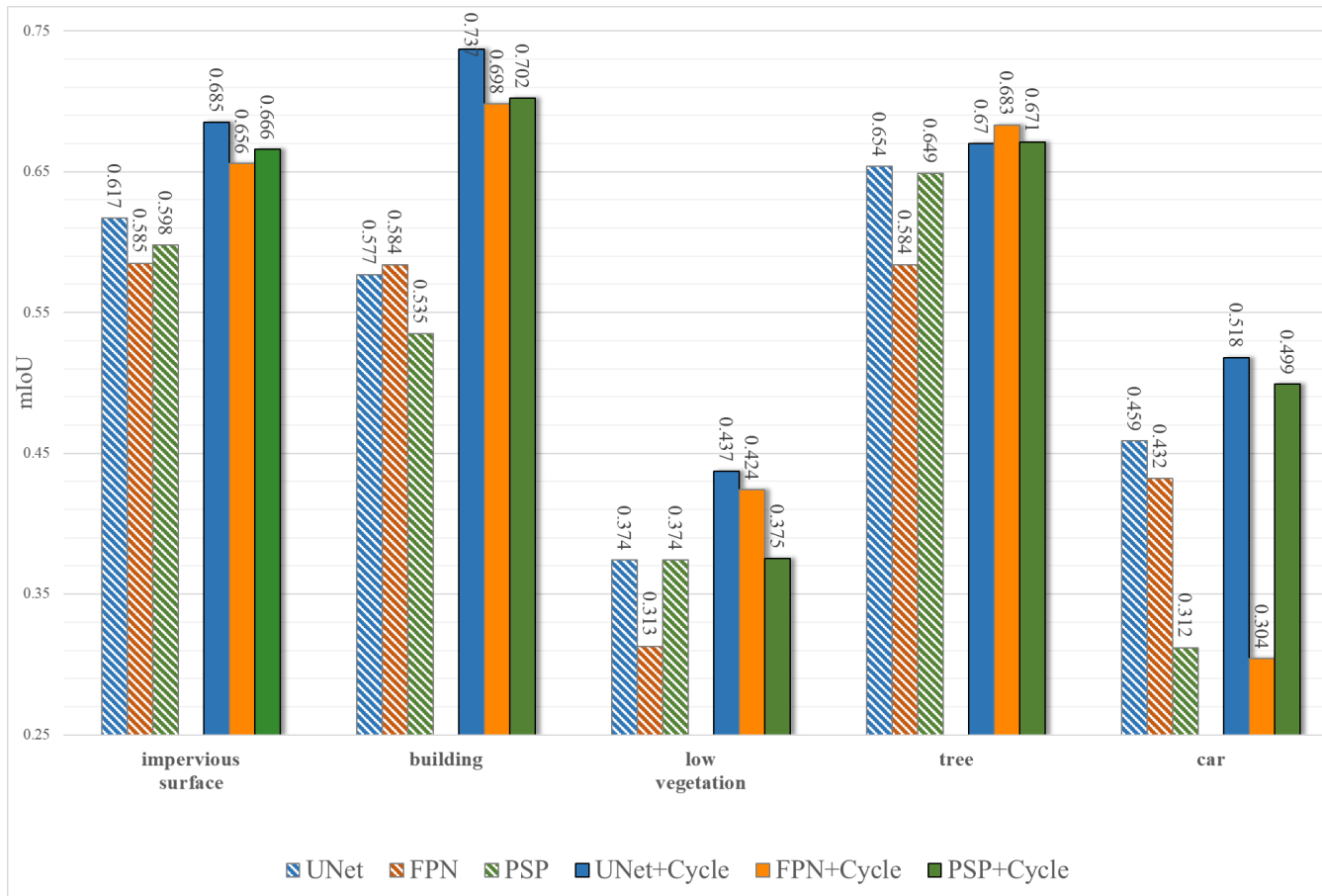


Figure 25. mIoU graph for each class by the three classification models and the proposed framework.

Figure 25 shows the mIoU scores by each class for the three classification models as well as for the inclusion of the proposed CycleGAN with the three classification models. Comparing results of the supervised classification models, UNet achieved the highest scores in all classes except for building. In the case of the building class, FPN recorded the best performance at 0.584, which can be explained by that multi-scale kernel in the model helps to capture information on the differently-sized buildings. PSPNet yielded better predictions than FPN for classes other than building and car. In particular, the pyramid pooling in PSPNet led the model to consider global contextual information, resulting in poor classification results of small-sized class like car. For the semi-supervised method, the mIoU score for the car class was greatly improved when using PSPNet by complementing insufficient local information derived from small objects with numerous unlabeled data. In general, the semi-supervised method improved the classification accuracy of the three classification models for the majority of the classes. In particular, the mIoU scores of the building class comprised of different sizes and shapes in test sites were greatly enhanced. However, since the CycleGAN module may lead to difficulty in model training, the accuracy results in some cases can be found to decrease. As an example, the mIoU score of the car class using FPN decreased from 0.432 to 0.304.

Chapter 6. Conclusion

This thesis proposed a semi-supervised framework using the EfficientNet backbone, UNet-based classification model, and a modified CycleGAN for remote sensing image classification. The proposed framework was established to address the problem of limited training data by employing the unlabeled images together with the labeled images. For this purpose, this thesis applied the cycle consistency loss of CycleGAN to help the generators preserve the information of the original image without the need for reference data. In addition, the recent EfficientNet was used based on its remarkable feature extraction performance and high training efficiency alleviating the mode collapse problems in limited remote sensing data. This semi-supervised framework proposed in this thesis can be a meaningful approach especially since remotely-sensed images are continuously being accumulated while manually labeling images is typically very difficult and expensive.

The proposed framework was evaluated by using three test sites from the ISPRS Vaihingen VHR dataset with five benchmarks composed of both supervised and semi-supervised methods. The highest accuracy (OA: 0.796 in area 1, 0.786 in area 15, 0.784 in area 23) in the test sites was achieved by the proposed framework when using semi-supervised learning with the EfficientUNet and the modified CycleGAN module. Especially, the largest increase in accuracy was observed in area 1 which contained objects with different properties due to the regularization effect for the unlabeled images. These results implied that a large amount of unlabeled data can be utilized to the generalization of the supervised model when biased and limited labeled data can be used, resulting in consistent segmentation of the remotely-sensed objects having various spatial properties.

In addition, this thesis analyzed the impact of labeled and unlabeled data and the cycle consistency loss on classification accuracy. By controlling the number of labeled image patches from one to three and the unlabeled image

patches from 6 and 13 patches, the results demonstrated that using more unlabeled patches improved the performance of the semi-supervised method, but the increase was smaller when a relatively sufficient amount of unlabeled data to labeled data are not available. Consequently, the improvement of the proposed framework is more dramatic when applied in an extreme condition of training data, such as when making a lot of ground truth data for remote sensing VHR image is extremely hard.

Lastly, to confirm the adaptation of the proposed framework for additional classification models, namely, FPN and PSPNet, were evaluated with UNet. For three models, segmentation results were improved when using the semi-supervised method. This third experiment showed that the proposed semi-supervised framework is not confined to a particle UNet model and independent to the backbone classification models, indicating the applicability to future classification backbone models.

Future works are necessary to tackle several limitations in this thesis. First, while the proposed framework was robust to spatial properties, objects containing similar spectral properties such as low vegetation and tree or objects in shadow were misclassified. Since the generators utilized the pre-trained weights by CV-based data, remote sensing data's unique spectral information could not be sufficiently considered and only three channels should be used in input data. If rich spectral information of remote sensing images is utilized through channel attention-based modules or multi-spectral data augmentation, results of the spectral-mixed objects can be improved. Second, the results of the clutter class are highly poor due to its inner-class heterogeneous and very limited samples. This problem can be critical in practical applications such as land cover classification, where class imbalance is often accompanied. Clutter's results emphasized that the modification of the proposed framework for class imbalance problem is necessary to apply to more realistic remote sensing data. Also, this thesis was restricted to a limited

dataset composed of similar ground properties used for training and testing. However, there is a clear need to apply the proposed framework to different domains containing various characteristics and classes. Moreover, for more practical applications, the semi-supervised framework should be extended toward employing multi-domain unlabeled images during model training. Lastly, the recent trend on semi-supervised learning is based on simultaneously combining several semi-supervised methods such as consistency loss, entropy minimization, and pseudo labeling (Berthelot et al., 2019). Stemming from this trend, the proposed semi-supervised framework in this thesis can be improved further to compare its applicability by combining additional semi-supervised methods together.

References

- Audebert, N., Le Saux, B., & Lefèvre, S. (2016). Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In Asian conference on computer vision (pp. 180-196). Springer, Cham.
- Bai, H., Cheng, J., Huang, X., Liu, S., & Deng, C. (2021). HCANet: A Hierarchical Context Aggregation Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*.
- Baheti, B., Innani, S., Gajre, S., & Talbar, S. (2020). Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 358-359).
- Bellens, R., Gautama, S., Martinez-Fonte, L., Philips, W., Chan, J. C. W., & Canters, F. (2008). Improved classification of VHR images of urban areas using directional morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 46(10), 2803-2813.
- Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. (2019). Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249.
- Bischke, B., Helber, P., Folz, J., Borth, D., & Dengel, A. (2019, September). Multi-task learning for segmentation of building footprints with deep neural networks. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 1480-1484). IEEE.
- Daudt, R. C., Le Saux, B., & Boulch, A. (2018). Fully convolutional siamese networks for change detection. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 4063-4067). IEEE.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
- Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162, 94-114.
- Dong, R., Pan, X., & Li, F. (2019). DenseU-net-based semantic segmentation of small objects in urban remote sensing images. *IEEE Access*, 7, 65347-65356.
- Feng, Q., Liu, J., & Gong, J. (2015). UAV remote sensing for urban vegetation mapping using random forest and texture analysis. *Remote sensing*, 7(1), 1074-1094.
- Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote sensing of environment*, 80(1), 185-201.

- Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial networks. arXiv preprint arXiv:1406.2661.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
- Hung, W. C., Tsai, Y. H., Liou, Y. T., Lin, Y. Y., & Yang, M. H. (2018). Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv:1802.07934.
- Iglovikov, V., Mushinskiy, S., & Osin, V. (2017). Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. arXiv preprint arXiv:1706.06169.
- Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).
- Jin, H., Stehman, S. V., & Mountrakis, G. (2014). Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado. *International journal of remote sensing*, 35(6), 2067-2081.
- Kampffmeyer, M., Salberg, A. B., & Jenssen, R. (2016). Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 1-9).
- Kirillov, A., He, K., Girshick, R., & Dollár, P. (2017). A unified architecture for instance and semantic segmentation.
- Lin, D. Y., Wang, Y., Xu, G. L., & Fu, K. (2017-a). Synthesizing remote sensing images by conditional adversarial networks. In 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 48-50). IEEE.
- Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017-b). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).
- Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).
- Luc, P., Couprie, C., Chintala, S., & Verbeek, J. (2016). Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408.
- Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017).

- Least squares generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2794-2802).
- Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
- Mondal, A. K., Agarwal, A., Dolz, J., & Desrosiers, C. (2019). Revisiting CycleGAN for semi-supervised segmentation. arXiv preprint arXiv:1908.11569.
- Peng, D., Zhang, Y., & Guan, H. (2019). End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sensing*, 11(11), 1382.
- Peng, D., Bruzzone, L., Zhang, Y., Guan, H., Ding, H., & Huang, X. (2020). SemiCDNet: a semisupervised convolutional neural network for change detection in high resolution remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).
- Sherrah, J. (2016). Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arXiv preprint arXiv:1606.02585.
- Shi, Y., Li, Q., & Zhu, X. X. (2018). Building footprint generation using improved generative adversarial networks. *IEEE Geoscience and Remote Sensing Letters*, 16(4), 603-607.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning* (pp. 6105-6114). PMLR.
- Van de Voorde, T., De Genst, W., & Canters, F. (2007). Improving pixel-based VHR land-cover classifications of urban areas with post-classification techniques. *Photogrammetric Engineering and Remote Sensing*, 73(9), 1017.
- Van Etten, A., Lindenbaum, D., & Bacastow, T. M. (2018). Spacenet: A remote sensing dataset and challenge series. arXiv preprint arXiv:1807.01232.
- Yakubovskiy, P. (2019). Segmentation Models. GitHub; https://github.com/qubvel/segmentation_models
- Vigueras-Guillén, J. P., Sari, B., Goes, S. F., Lemij, H. G., van Rooij, J., Vermeer, K. A., & van Vliet, L. J. (2019). Fully convolutional architecture vs sliding-window CNN for corneal endothelium cell

- segmentation. *BMC Biomedical Engineering*, 1(1), 1-16.
- Zhang, X., Han, X., Li, C., Tang, X., Zhou, H., & Jiao, L. (2019). Aerial image road extraction based on an improved generative adversarial network. *Remote Sensing*, 11(8), 930.
- Zhang, X., Zhu, X., Zhang, N., Li, P., & Wang, L. (2018-a). Seggan: Semantic segmentation with generative adversarial network. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)* (pp. 1-5). IEEE.
- Zhang, Z., Liu, Q., & Wang, Y. (2018-b). Road extraction by deep residual unet. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 749-753.
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890).
- Zhou, L., Zhang, C., & Wu, M. (2018). D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 182-186).
- Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

국문 초록

고해상도 영상 분류를 위한 순환 적대적 생성 신경망 기반의 준지도 학습 프레임워크

서울대학교 대학원

공과대학 건설환경공학부

곽 태 홍

고해상도 영상 분류는 토지피복지도 제작, 식생 분류, 도시 계획 등에서 다양하게 활용되는 대표적인 영상 분석 기술이다. 최근, 심층 합성곱 신경망 (deep convolutional neural network)은 영상 분류 분야에서 두각을 보여왔다. 특히, 심층 합성곱 신경망 기반의 의미론적 영상 분할 (semantic segmentation) 기법은 연산 비용을 매우 감소시키며, 이러한 점은 지속적으로 고해상도 데이터가 축적되고 있는 고해상도 영상을 분석할 때 중요하게 작용된다.

심층 학습 (deep learning) 기반 기법이 안정적인 성능을 달성하기 위해서는 일반적으로 충분한 양의 라벨링된 데이터 (labeled data)가 확보되어야 한다. 그러나, 원격탐사 분야에서 고해상도 영상에 대한 참조데이터를 얻는 것은 비용적으로 제한적인 경우가 많다. 이러한 문제를 해결하기 위해 본 논문에서는 라벨링된 영상과 라벨링되지 않은 영상 (unlabeled image)을 함께 사용하는 준지도 학습 프레임워크를 제안하였으며, 이를 통해 고해상도 영상 분류를 수행하였다. 본 논문에서는 라벨링되지 않은 영상을 사용하기 위해서 개선된 순환 적대적 생성 신경망 (CycleGAN) 방법을 제안하였다.

순환 적대적 생성 신경망은 영상 변환 모델 (image translation model)로

처음 제안되었으며, 특히 순환 일관성 손실 함수 (cycle consistency loss function)를 통해 페어링되지 않은 영상 (unpaired image)을 모델 학습에 활용한 연구이다. 이러한 순환 일관성 손실 함수에 영감을 받아, 본 논문에서는 라벨링되지 않은 영상을 참조데이터와 페어링되지 않은 데이터로 간주하였으며, 이를 통해 라벨링되지 않은 영상으로 분류 모델을 함께 학습시켰다.

수많은 라벨링되지 않은 데이터와 상대적으로 적은 라벨링된 데이터를 함께 활용하기 위해, 본 논문은 지도 학습과 개선된 준지도 학습 기반의 순환 적대적 생성 신경망을 결합하였다. 제안된 프레임워크는 순환 과정(cyclic phase), 적대적 과정(adversarial phase), 지도 학습 과정(supervised learning phase), 세 부분을 포함하고 있다. 라벨링된 영상은 지도 학습 과정에서 분류 모델을 학습시키는 데에 사용된다. 적대적 과정과 지도 학습 과정에서는 라벨링되지 않은 데이터가 사용될 수 있으며, 이를 통해 적은 양의 참조데이터로 인해 충분히 학습되지 못한 분류 모델을 추가적으로 학습시킨다.

제안된 프레임워크의 결과는 공공 데이터인 ISPRS Vaihingen Dataset을 통해 평가되었다. 정확도 검증을 위해, 제안된 프레임워크의 결과는 5개의 벤치마크들 (benchmarks)과 비교되었으며, 이때 사용된 벤치마크 모델들은 지도 학습과 준지도 학습 방법 모두를 포함한다. 이에 더해, 본 논문에서는 라벨링된 데이터와 라벨링되지 않은 데이터의 구성에 따른 영향을 확인하였으며, 다른 분류 모델에 대한 본 프레임워크의 적용가능성에 대한 추가적인 실험도 수행하였다.

제안된 프레임워크는 다른 벤치마크들과 비교해서 가장 높은 정확도 (세 실험 지역에 대해 0.796, 0.786, 0.784의 전체 정확도)를 달성하였다. 특히, 객체의 크기나 모양과 같은 특성이 다른 실험 지역에서 가장 큰 정확도 상승을 확인하였으며, 이러한 결과를 통해 제안된 준지도 학습이 모델을 우수하게 정규화(regularization)함을 확인하였다. 또한, 준지도 학습을 통해 향상되는 정확도는 라벨링된 데이터에 비해 라벨링되지

않은 데이터가 상대적으로 많았을 때 그 증가 폭이 더욱 커졌다. 마지막으로, 제안된 준지도 학습 기반의 순환 적대적 생성 신경망 기법이 UNet 외에도 FPN과 PSPNet이라는 다른 분류 모델에서도 유의미한 정확도 상승을 보였다. 이를 통해 다른 분류 모델에 대한 제안된 프레임워크의 적용가능성을 확인하였다

Keywords : 준지도 심층 학습, 순환 적대적 생성 신경망, 고해상도 영상 분류, 의미론적 영상 분할

Student Number : 2019-24705