remote sensing

MDPI

Article

Multi-Scale Feature Map Aggregation and Supervised Domain Adaptation of Fully Convolutional Networks for Urban Building Footprint Extraction

Jagannath Aryal *,† **and Bipul Neupane** †

Department of Infrastructure Engineering, Faculty of Engineering and IT, The University of Melbourne, Melbourne, VIC 3010, Australia

* Correspondence: jagannath.aryal@unimelb.edu.au

† These authors contributed equally to this work.

Abstract: Automated building footprint extraction requires the Deep Learning (DL)-based semantic segmentation of high-resolution Earth observation images. Fully convolutional networks (FCNs) such as U-Net and ResUNET are widely used for such segmentation. The evolving FCNs suffer from the inadequate use of multi-scale feature maps in their backbone of convolutional neural networks (CNNs). Furthermore, the DL methods are not robust in cross-domain settings due to domain-shift problems. Two scale-robust novel networks, namely MSA-UNET and MSA-ResUNET, are developed in this study by aggregating the multi-scale feature maps in U-Net and ResUNET with partial concepts of the feature pyramid network (FPN). Furthermore, supervised domain adaptation is investigated to minimise the effects of domain-shift between the two datasets. The datasets include the benchmark WHU Building dataset and a developed dataset with 5× fewer samples, 4× lower spatial resolution and complex high-rise buildings and skyscrapers. The newly developed networks are compared to six state-of-the-art FCNs using five metrics: pixel accuracy, adjusted accuracy, F1 score, intersection over union (IoU), and the Matthews Correlation Coefficient (MCC). The proposed networks outperform the FCNs in the majority of the accuracy measures in both datasets. Compared to the larger dataset, the network trained on the smaller one shows significantly higher robustness in terms of adjusted accuracy (by 18%), F1 score (by 31%), IoU (by 27%), and MCC (by 29%) during the cross-domain validation of MSA-UNET. MSA-ResUNET shows similar improvements, concluding that the proposed networks when trained using domain adaptation increase the robustness and minimise the domain-shift between the datasets of different complexity.

Keywords: deep learning; building footprint extraction; multi-scale feature aggregation; supervised domain adaptation; U-Net; ResUNET

Citation: Aryal, J.; Neupane, B. Multi-Scale Feature Map Aggregation and Supervised Domain Adaptation of Fully Convolutional Networks for Urban Building Footprint Extraction. *Remote Sens.* **2023**, *15*, 488. https://doi.org/10.3390/rs15020488

Academic Editor: Benoit Vozel

Received: 1 December 2022
Revised: 6 January 2023
Accepted: 12 January 2023
Published: 13 January 2023

1. Introduction

Urban remote sensing for feature extraction is the data source of spatial information. The extracted spatial information is used for infrastructure analytics and planning. With the availability of analysis-ready very-high-resolution (VHR) Earth observation (EO) remote sensing images, it is possible to retrieve spatial information that is fit for purpose. However, the retrieved information needs quality assurance in terms of efficiency, robustness, accuracy and precision. These aspects are addressed to some extent by utilising evolving intelligent approaches such as DL. CNNs are fundamental to extracting context information from images for object detection and semantic segmentation. Per-pixel semantic segmentation is the current state-of-the-art (SOTA) DL method for extracting urban features such as buildings. The method is evolving and has several known limitations [1]. This paper focuses on two limitations:

1. Loss of context information at different scales of feature maps in CNNs.

2. The domain-shift problem due to the difference in training and testing data.

The first limitation is faced by all CNN-based DL networks. The current SOTA networks for the semantic segmentation of images come from the FCN [2] family [1]. The widely common FCNs such as U-Net [3], ResUNET [4], and SegNet [5] are symmetrical networks with an encoder and a decoder. From the explanation of Ronneberger et al. (2015) [3], an encoder is essentially a CNN that extracts context features at multiple scales of feature maps. A decoder ensures the precise localisation of the object (building in our focus). Context refers to information that helps CNNs accurately interpret an object. Both context information and precise localisation are fundamental for urban feature extraction from EO images. However, the known problem of CNNs is that the context information is lost because of the downsampling operations used in them. The operation provides multi-scale feature information necessary to extract an object. However, at the same time, they reduce the size of the feature maps, resulting in the loss of context information. The popularity of U-Net comes from a trade-off between context and precise localisation. Recent studies have shown that the loss of context can be compensated by the aggregation and fusion of multi-scale feature maps (also known as features) [6]. FPN [7] is a common example of multi-scale feature aggregation. In this paper, commonly used U-Net and ResUNET networks of FCN family are integrated with the partial concepts of an FPN. The proposed two networks named Multi-Scale Aggregation U-NET (MSA-UNET) and MSA-ResUNET are developed to investigate whether the multi-scale feature aggregation technique inspired by FPN can improve the precision and accuracy of U-Net and ResUNET. These multi-scale features are crucial in extracting size-variant urban building footprints on EO images.

The second limitation, i.e., the domain-shift problem, is a common yet less-studied limitation of CNN-based DL methods. Domain-shift occurs when the training and test data are of different settings. A general example of the limitation is that a DL network trained to extract the buildings from India fails to extract the buildings from Australia. The limitation not only occurs due to the difference in the architectures of buildings in two places, but also occurs due to several other factors such as the different types of buildings (low-rise, mid-rise, and high-rise), the spatial resolution of EO images, time of image collection, shadows, and other spatial variabilities between train and test environment. These problems are unavoidable when it comes to using EO images for nationwide building footprint extraction. Therefore, the domain-shift problem must be reduced to increase the robustness of the DL network in cross-domain settings. The two proposed networks are therefore evaluated in two experimental settings: (i) same-domain validation; and (ii) cross-domain validation. The same-domain validation refers to the setting of having the training and validation data from the same dataset. Similarly, the cross-domain refers to having the training and validation data from two different datasets. To tackle the domain-shift problem in cross-domain settings, a supervised domain adaptation (DA) based on *inductive transfer learning* [8] is applied between two datasets: one from Christchurch, New Zealand, and another from Melbourne, Australia. The first one is a benchmark dataset called the Wuhan University (WHU) Building dataset [9], and the second one is prepared from the case study area of the City of Melbourne. Four settings of supervised DA are experimented to see its effects on different scales of feature maps in the proposed networks.

The major contributions of this study are:

1. Development of MSA-UNET and MSA-ResUNET networks to tackle the loss of context information at different scales of feature maps in CNNs.
2. Development of a high-resolution dataset for an experimental design within DL framework.
3. Evaluation of four settings of supervised domain adaptation to tackle the problem of domain-shift in different datasets.
4. A comprehensive evaluation approach in cross-domain settings.

The rest of the paper is structured as follows: Section 2 presents the related literature; Section 3 describes the datasets and method; Section 4 presents the experiments and results,

Section 5 provides critical discussions with more experiments; and Section 6 concludes the paper.

2. Related Works

2.1. Dl-Based Urban Feature Extraction from VHR EO Imagery

The building footprint extraction falls under a broader research topic of urban feature extraction, which includes a set of features to extract: impervious surface [10], buildings, low vegetation [11,12], tree [13], car [14], etc. Building footprints are among the most prominent features in urban settings [15]. With the availability of VHR imagery, the research paradigm of urban feature extraction has shifted from object-based image analysis to pixel-based semantic segmentation [1]. The neural networks developed for medical image segmentation have been adopted in the segmentation of VHR EO images to gain from the reduction in pixel size. Some also improved these adopted networks to gain efficiency in the context of EO-based methods for urban feature extraction [16]. Among the commonly experimented networks, CNNs [17] such as AlexNet [18], VGGNet [19], GoogleNet [20], ResNet [21], Xception [22], and RefineNet [23] are frequently used for urban feature extraction including buildings. However, with the superior performance of FCNs [2] including encoder–decoder networks such as U-Net or SegNet [5] in semantic segmentation, CNNs are now used as a feature extractor of encoder–decoder networks. The evolution of CNNs and FCNs leading towards encoder–decoder architectures that are used to extract urban building footprints are presented in the next section. The architectures formulate the proposed networks and the method design of this research.

2.1.1. CNNs and FCNs for Building Footprint Extraction

In one of the earliest studies of building footprint extraction, [24] trained a randomly initialised Restricted Boltzmann Machine (RBM) to extract the roads on aerial images. Volodymyr Mnih (2013) [25] later proposed a Massachusetts Buildings Dataset while studying machine learning (ML) methods including RBM to label buildings and roads on aerial image patches. Saito et al. (2015) [26] and Saito et al. (2016) [27] improved upon Mnih's RBM [25] and propose a five-layered CNN. Vakalopoulou et al. (2015) [28] trained a CNN to produce a building/not-building binary classification on red–green–blue (RGB) images. The output is used as a feature vector to train a support vector machine (SVM) classifier, followed by Random Forest (RF) during post-processing. These early studies used CNN and post-processing methods assisted by a conventional ML to improve the final segmentation.

In contrast to the patch-based segmentation using CNN, FCNs perform building extraction at the per-pixel level. Several studies at the time concluded that the pre-trained FCNs such as FCN-2s, FCN-4s, and FCN-8s when fine-tuned over new building and roads dataset provide higher performance over CNNs [29,30]. With the ongoing investigation over finding the superiority of CNNs and FCNs, a dual-stream CNN (VGG-Net and AlexNet) trained simultaneously over smaller and larger patch-size was proposed by Marcu et al. (2017) [31] to outperform the existing CNNs [25,26]. Zhao et al. (2018) [32] extracted buildings using a combination of Mask R-CNN (with the ResNet-101 feature extractor and a head of Faster R-CNN) and building boundary regularisation with a combination of Douglas–Peucker algorithm, set of hypothesis, and Minimum Description Length (MDL) framework. Yang et al. (2018) [33] compared nine CNNs, an FCN, a CRF as Recurrent Neural Network (CRFasRNN), and a SegNet for building segmentation across 78 locations to conclude the superior performance of SegNet. Griffiths et al. (2019) [34] trained Mask R-CNN and RetinaNet on two sets of labels that are improved using morphological operations. After the continuous improvement of CNNs and FCNs, the encoder–decoder networks such as SegNet and U-Net changed the SOTA in building footprint extraction. The structural configuration is presented in the next section.

2.1.2. Encoder–Decoder Network Architectures

The architecture of encoder–decoder networks consists of an encoder and a decoder. An encoder is a feature extractor CNN that turns the input image space into the depth of feature maps. A decoder then maps the low-resolution feature maps from the encoder to the high-resolution feature maps of input size for pixel-wise classification. For example, SegNet consists of an encoder that is similar to a 13-layer VGG-16 CNN, and a corresponding decoder that upsamples the encoder's lower-resolution input feature maps. U-Net consists of the encoder and decoder with the addition of "skip connections". These connections pass the low-level information from the encoder to the decoder. This makes the utmost use of the encoder–decoder structure. Unlike CNNs, encoder–decoder networks, therefore, utilise both low- and high-resolution features, conserving the spatial integrity of objects that is crucial in the semantic segmentation of features in EO data. However, the maximum potential of these multi-scale features is still being explored in recent studies [35,36].

With the success of U-Net in medical image segmentation, it is also adopted for EO-based building footprint extraction. U-Net has seen several re-iterations such as ResUNET, U-Net++ [37], and U-Net3+ [38]. Xu et al. (2018) [39] trained ResUNET on multi-modal images obtained from hand-crafted features such as NDVI, digital surface model (DSM), normalised DSM (nDSM), and principal component analysis (PCA). They pre-processed the hand-crafted features using edge enhancement to reduce noise and pre-process the output of ResUNET using guided filters. Liu et al. (2019) [40] trained a U-Net on a multi-channel image stack of red, red-edge, coastal, blue, and green image bands of WorldView-3 satellite images and labels from Google maps, OSM, and MapWorld. The final segmented feature map is binarised with a threshold of 0.5 to obtain building polygon vectors. Yi et al. (2019) [41] proposed DeepResUNET for effective urban building segmentation on ultra-high resolution (UHR) images of 0.075 m and compare to FCN-8s, SegNet, DeconvNet, U-Net, ResUNET, and DeepUNet. Bischke et al. [42] modified SegNet by adding layers to predict the distance to the border of buildings and introduce a multi-task loss to address the boundary segmentation problem. Qin et al. (2019) [43] tested the efficacy of Gaofen-2 imagery for the semantic segmentation of building roofs in dense urban environments using an FCN with a VGG-16 backbone. Abdollahi et al. (2020) [44] combined SegNet and U-Net to form Seg-Unet. Pan et al. (2020) [45] trained a U-Net to classify four types of buildings. Sariturk et al. [46] compared SegNet and FCN. Ayala et al. (2021) [47] fused synthetic aperture radar (SAR) image from Sentinel-1 and multi-spectral image (RGB-NIR-NDVI at 10 m) of Sentinel-2 together with OSM label vectors to train a modified U-Net with ResNet-34 as an encoder. It can be seen over the years that the complexity of experimental data has changed from sparse building environments to dense urban environments, and the FCNs and encoder–decoder networks are evolving to tackle several problems associated with DL-based semantic segmentation.

The proposed networks in this study start with an encoder–decoder architecture of U-Net and ResUNET, upon which multi-scale feature aggregation (MSA) with a partial concept of FPN is added. The related works on the aggregation and fusion of multi-scale features on encoder–decoder networks are presented in the next section.

2.1.3. Multi-Scale Feature Aggregation on Encoder–Decoder Networks

The fusion of feature maps is an advancing work [48]. The aggregation and fusion of the multi-scale features in encoder–decoder networks such as U-Net is a trending approach to exploit the multi-scale features in recent years [49–51]. Liu et al. (2018) [52] proposed a self-cascaded network (ScasNet) that aggregates the multi-scale feature maps of image size obtained from its encoder to extract global-to-local contexts. The coarse feature maps are then refined using the fine-structure object refinement method. Yang et al. (2018) [53] proposed a dense-attention network (DAN) to utilise the multi-level features for building segmentation. DAN consists of an encoder of light DenseNets as a feature extractor and decoder with a spatial attention fusion module to utilise both low- and high-level feature maps. Ji et al. (2018) [9] proposed a Simamese U-Net that learns from the original image

as well as downsampled image patch, and one input in one of two branches of the U-Net with shared weights. This allows learning multi-scale and multi-source information. The outputs of the branches are then concatenated for the final output. They also prepared the WHU building dataset of 0.3 m resolution. Wu et al. (2018) [54] proposed a multi-constraint FCN (MC–FCN) that adopts the basic structure of U–Net and adds three extra multi-scale constraints between upsampled layers and their corresponding ground truths. Chen et al. (2018) [55] automated the mapping of buildings through roof segmentation and compared three networks with ResNet-101 backbone: FPN, FPN with multi-scale feature fusion (MSFF), and pyramid scene parsing network (PSPNet). The PSPNet shows the highest performance. Wei et al. (2019) [6] proposed a multi-scale aggregation FCN (MA-FCN), which is a combination of a VGG-16 encoder as a feature extractor, a corresponding decoder, and an FPN that aggregates the multi-scale feature outputs from the decoder. Two post-processing methods are introduced to refine the final segmentation map, and overlapping tiles are used to remove incorrect predictions at the edges of image tiles. The network is trained on the WHU dataset and is compared to SiU-Net [9], U-Net, DeepLabv3+, and Mask R-CNN. Ji et al. (2019) [56] used Atrous convolutions, multi-scale aggregation, combined data augmentation, and relative radiometric calibration method for multi-source building extraction. They compared their method to Deeplabv3+ [57], C-Unet (MC-FCN from [54]), U-Net, FCN-8s and a 2-scale FCN [30]. The proposed aggregation of multi-scale feature maps in this study comes from a partial concept of FPN. Unlike in FPN, the aggregation utilises both higher- and lower-resolution feature maps as both are crucial to preserving the spatial relationship among the features in EO images.

After integrating the multi-scale feature aggregation in the proposed networks, supervised DA techniques are investigated to minimise the effects of the domain-shift problem. The related works on the problem and the solution of DA are presented in the next section.

2.2. The Domain-Shift Problem, Transfer Learning, and Domain Adaptation

The domain-shift problem is a common problem in DL methods. The problem occurs when there exists a difference in training (source domain) and test (target domain) data. In VHR EO data, the shift is generally caused by the difference in imaging sensors, spatial resolution, complexity, and class representation. Transfer learning (TL) [8,58] is a common method to reduce the impact of domain-shift and increase the performance of a DL network with fewer training samples and less computational power. It allows the knowledge gained from solving one problem to be transferred and used for similar problems, making it popular in studies that lack enough training samples [59]. Pan et al. (2009) [8] categorised the TL into three settings: inductive TL, transductive TL, and unsupervised TL. According to their definition, a transductive TL setting is called DA when the label samples are only available in the source domain and the task is the same in both the source and target domain. However, over the years, the definition of DA has changed. DA can be of three types [60,61]: supervised, unsupervised, and semi-supervised. The transductive TL setting where only the source data are labelled is now called unsupervised DA. Conversely, DA is supervised if both the source and target domain are labelled. Finally, a semi-supervised DA has labelled data in the source and both labelled and unlabelled data in the target.

The unsupervised and supervised DA is found in the recent literature on building footprint extraction. The instance transfer is a DA (whether supervised or unsupervised) that is performed to re-weight some labelled data from the source domain to be used in the target domain. Panboonyuen et al. (2019) [62] performed a domain-specific TL to transfer the knowledge of a graph convolution network (GCN) trained on VHR satellite images to a GCN trained on medium-resolution images. They perform instance transfer to re-label weights from the source domain to the target domain utilising the backpropagation algorithm to improve the target learning task. In a similar work, Wurm et al. (2019) [63] used TL for their DL networks trained on training samples created from different sensors (QuickBird, Sentinel-2, and TerraSAR-X) with varying spatial resolution. Some methods based on the training and fine-tuning of FCNs [29,30] were already presented in the

previous section. Some have used TL in their experiments for the ablation study [9,56]. In most studies, the most common TL approach is to transfer the weights of an off-the-shelf pre-trained CNN such as ResNet, VGGNet, Xception, Inception, etc., as a feature extractor that is previously trained on a generic dataset such as Visual Object Classes (VOCs) [64] and the ImageNet [65] dataset. The prior studies performed supervised DA. Liu et al. (2020) [66] proposed a bispace alignment network (BSANet) that uses two branches of modified U-Nets for unsupervised DA, automated labelling, and minimising the domain-shift problem. They used BSANet as a generator in a generative adversarial network (GAN) of two discriminators to minimise the discrepancy between the source and target domains.

The supervised DA in this research is similar to our previous study on Neupane et al. (2022) [67]. In this work, the knowledge from the MSA-UNET and MSA-ResUNET trained on larger and higher-resolution aerial images of 0.3 m is transferred to smaller and lower-resolution aerial images of 1.2 m. The DA involves a target domain of large and higher-resolution samples and a source domain of small and lower-resolution complex samples. DA from a large to a small dataset is useful to obtain higher precision with a low number of training samples in the target dataset [68]. Similarly, DA from higher to lower-resolution data is useful to make the model trained on the target low-resolution data apply the knowledge from the high-resolution data. A similar method demonstrated an improved extraction of building footprints in EO imagery [9,56] and in the studies of other domains [69]. The overall data preparation and method design is explained in the next section.

3. Method

3.1. Data Preparation

Two datasets are used for the experiments: (i) the WHU Building dataset and (ii) a new dataset that was developed: the Melbourne Building dataset. The WHU Building dataset (abbr. TR1) includes the satellite images of Christchurch, New Zealand, with a spatial resolution of 0.3 m as generated by Ji et al. in [9]. To maintain the uniform image size between the two datasets that were experimented on, the 512 × 512 sized-image samples of the WHU dataset are tiled into 256 × 256. The training and validation samples of 23,088 and 9664 tiles are prepared for TR1. Figure 1 shows the study area and the sample training images of TR1.

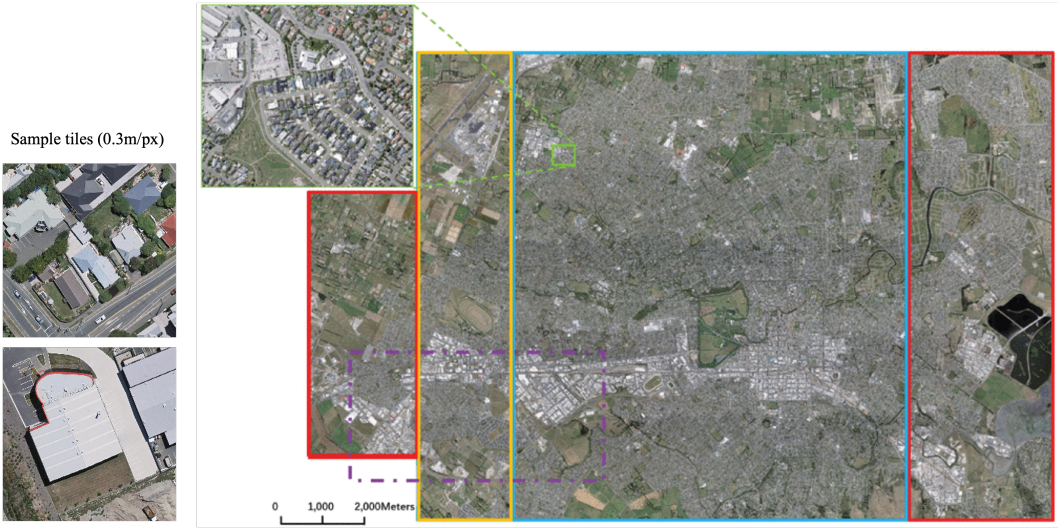

Figure 1. Aerial imagery of the WHU Building dataset set in Christchurch, New Zealand, and sample training images of 0.3 m spatial resolution with low-rise building structures (adopted from [9]).

The second dataset that is developed (label, image) was named the Melbourne Building dataset (abbr. TR2). The labels were developed by masking and tiling the building roof samples collected from the publicly available "2018 Building Footprints" data provided

by the City of Melbourne. The corresponding image tiles of 1.2 m spatial resolution of the labels were collected from Nearmap's API service. The training and validation samples were then separated using the boundary of Census of Land Use and Employment (CLUE) areas. Carlton suburb with 16.5% roof samples was taken as the validation area, and the remaining as the training area. A total of 4889 and 435 tiles were prepared for training and validation image tiles with an overlap of 50% between the adjacent tiles. Figure 2 shows the study area and the sample training images of TR2.

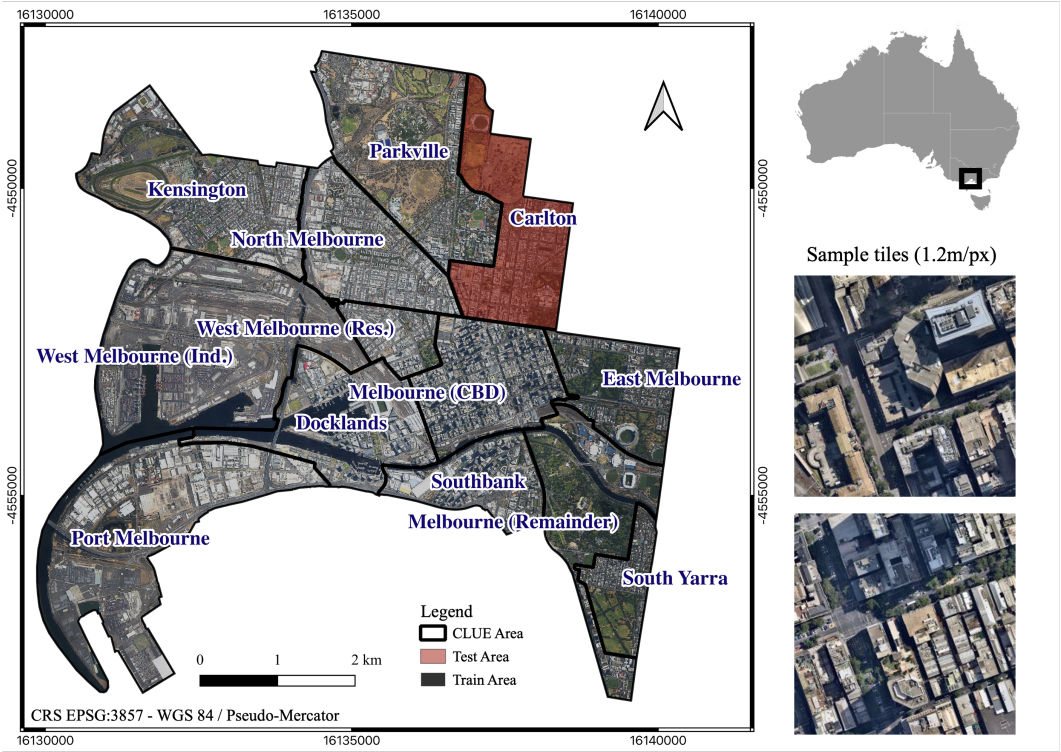

Figure 2. Study area of the City of Melbourne and sample training images of 1.2 m spatial resolution with high-rise building structures.

TR2 is five times smaller in size and has four times the lower spatial resolution compared to TR1 and includes a high number of high-rise buildings that are tedious to distinguish. TR2 includes 753 high-rise buildings and 67 skyscrapers above 150 m in height from the City of Melbourne. On the other hand, TR1 includes the building footprints of Christchurch, New Zealand, which contains just three high-rise buildings with the highest one of 86.5 m, and mostly the short commercial and residential buildings. The high-rises and skyscrapers increase the complexity of TR2 due to the shadows and the angle of inclination of the camera sensors. With these differences, TR2 is a smaller and more complex dataset and there exists a major shift in the domain between TR1 and TR2, resulting in poor accuracy from the DL models trained on TR1 when validated against TR2. Supervised DA is performed to tackle this problem.

3.2. MSA-UNET

The development of MSA-UNET was inspired by U-Net and FPN network architectures. A specific focus is on an aggregation of multi-scale features that are derived from a symmetrical FCN that preserves both contextual and spatial information. MSA-UNET first starts with a U-Net architecture with encoder blocks X^n_{En}, bottleneck X^5_{En}, and the corresponding decoder blocks X^n_{De}, where $n \in [1, 4]$. The network architecture is illustrated in the left part of Figure 3. Providing a brief explanation to U-Net, the encoder (contracting path) consists of recurring unpadded 3×3 convolutions, rectified linear unit (ReLU), and a 2×2 max pooling with a stride of 2 for downsampling. The number of feature channels is

doubled in each downsampling step of the encoder. The decoder (expansive path) upsamples the low-resolution feature maps of the encoder using 2 × 2 "up-convolution" layers that halve the number of feature channels. In addition, the feature map of the corresponding encoder layer is concatenated to the output of up-convolution using skip connections. This is followed by two 3 × 3 convolutions, each followed by a ReLU. A 1 × 1 convolution in the final layer maps the 64-component feature vectors to the number of classes.

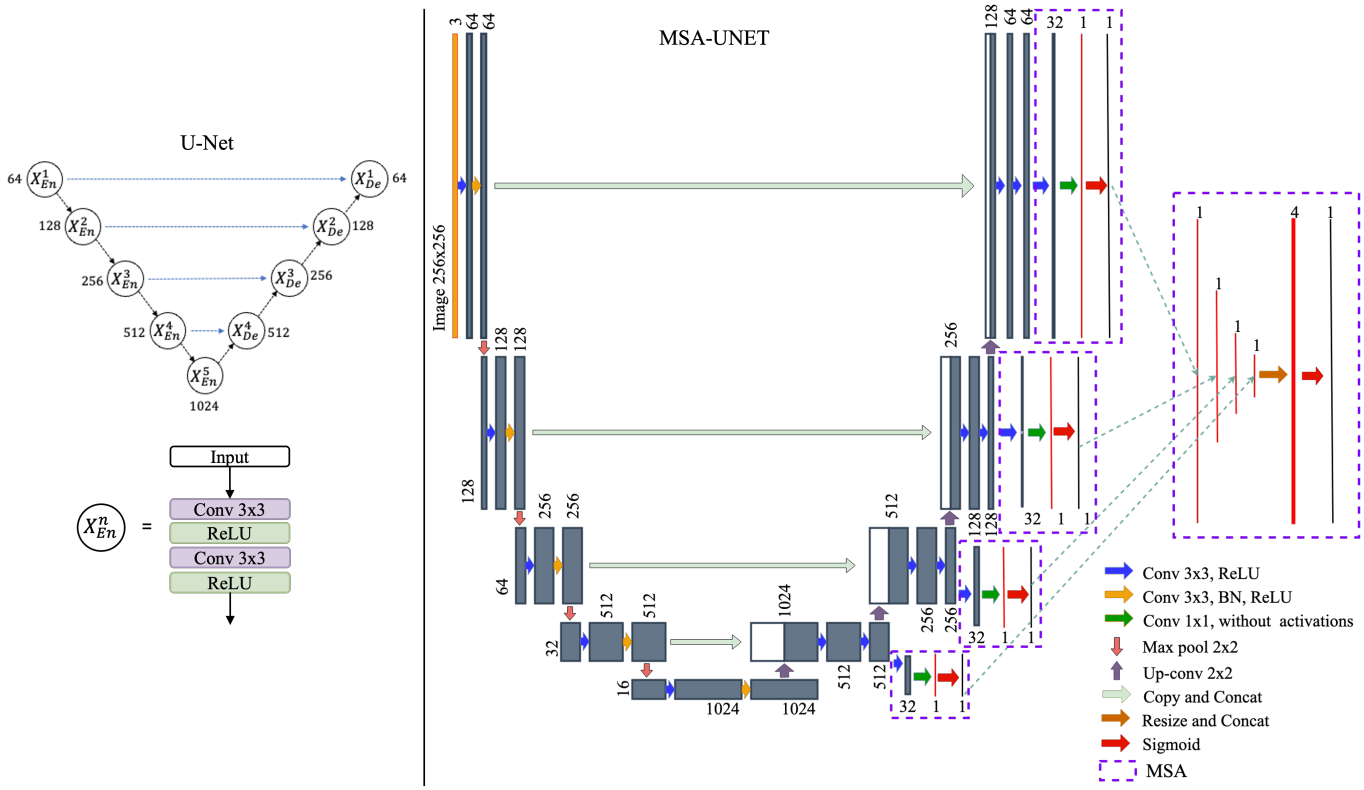

Figure 3. Network architecture of U-Net (left) and the proposed MSA-UNET (right). The MSA-block of the proposed MSA-UNET inherits the multi-scale features from the last layer of each (among four) decoder block of U-Net, and applies 3 × 3 convolutions with ReLU activation function and a 1 × 1 convolution with sigmoid function to produce a segmented map. The maps of all four scales are then aggregated by (i) re-sizing them to input image size, (ii) concatenating together, and (iii) applying a 1 × 1 convolution with a sigmoid activation function to obtain the final prediction.

From the explanation of U-Net, its encoder is essentially a CNN. The CNN extracts the features of an image without losing its characteristics. This extraction increases feature resolution while decreasing image resolution. This process converts an image into a vector while learning from the feature maps. In the decoder of the U-Net, the same feature maps are used to convert the vectors back to segmented image output by concatenating the high-resolution features coming from the encoder through skip connections to the spatially high-resolution image in the decoder. The symmetrical nature of the architecture, therefore, helps preserve the structural integrity and spatial relationship of the features in the image. To preserve and exploit the multi-scale feature information, MSA-UNET aggregates them using a multi-scale feature aggregation (abbr. MSA) inspired by FPN. The FPN proposed by Lin et al. (2017) [7] only uses the feature map of the highest spatial resolution and disregards the low-resolution feature maps. Unlike FPN, the added MSA collects the features maps of four scales from the last layer of each decoder blocks X^n_{De} of U-Net and applies 3 × 3 convolutions with 32 filters and a 1 × 1 convolution activated by a *sigmoid* function. This is illustrated in the right part of Figure 3, where the layers of U-Net are also illustrated in detail. The collected four sets of feature maps from U-Net are then aggregated by first re-sizing into the same scale of 256 × 256 (image size) and concatenating them

together. A 1×1 convolution is finally applied to the concatenated feature map with a *sigmoid* activation function to obtain the final prediction. This aggregation of low-resolution and high-resolution feature information is crucial to extracting urban features such as the building footprint from EO images. MA-FCN [6] follows a similar combination of an FCN and FPN, where they rely on an FCN with a VGG-16 encoder and its corresponding decoder.

3.3. MSA-ResUNET

MSA-ResUNET follows a similar architecture as the MSA-UNET, except this time, the MSA block is added to a ResUNET. ResUNET is a network first proposed by Zhang et al. (2018) [4] that uses residual blocks to build a deeper network. The residual block and the rich skip connections in the network minimise the vanishing and exploding gradient problem faced by the U-Net, providing an easier flow of gradients during backpropagation. ResUNET replaces the two 3×3 convolutions activated by the ReLU activation function with a pre-activated residual block in the encoder and the bottleneck. The decoder blocks then consist of a 2×2 upsampling layer that doubles the spatial dimension of the feature maps. These feature maps are concatenated to the features that are brought by the skip connection from the corresponding encoder block. These concatenated feature maps are then fed to the pre-activated residual block. In MSA-ResUNET, the feature maps of four scales from the last layer of each decoder block of ResUet are collected for multi-scale feature aggregation using the concept of MSA proposed in the previous section. The collected feature maps from each decoder block are then applied a 3×3 convolution with 32 filters, a ReLU, and a 1×1 convolution activated by a *sigmoid* function. The output of this process results in four feature maps of different scales. These maps are concatenated and apply a 1×1 convolution with a sigmoid activation function to obtain the final segmented output. The network architecture of MSA-ResUNET is illustrated in Figure 4.

3.4. Training and Supervised Domain Adaptation of MSA-UNET

MSA-UNET and MSA-ResUNET are trained in two steps, as shown in Figure 5. First, they are trained on TR1 to prepare a base model (M1) with a larger knowledge base. Secondly, supervised DA is performed on TR2 to produce M2. During the DA, the weights are initialised from M1 using transfer learning. The overall architecture, hyper-parameters, and augmentation steps are kept the same as before, but this time the encoder is frozen without affecting the knowledge gained from TR1. This allows a surgical fine-tuning of M1 such that the domain-adapted model M2 can leverage the knowledge from both TR1 and TR2. By leveraging the knowledge learnt from an existing large dataset into M2 trained on a smaller dataset, the overall method tackles the problem of domain-shift and also reduces the requirement of a large training dataset on a case study. Moreover, the DA from the higher-resolution source domain (TR1) to the lower-resolution target domain (TR2) is performed to make the model trained on TR2 use the knowledge from TR1 to make predictions on TR2.

To maximise the dice coefficient (also known as F1-score) and minimise the possible imbalance between the number of "background" and "building pixels", the dice loss [70] is used as the loss function for binary classification. This loss calculates the measure of overlap to assess the performance of segmentation when a ground truth (GT) is available. Dice loss is denoted by Equation (1).

$$DL(y, \hat{p}) = 1 - \frac{2y\hat{p} + 1}{y + \hat{p} + 1} \tag{1}$$

where y and $\hat{p}$ represent the GT and prediction, respectively. The smooth value of 1 is added in the numerator and denominator to make sure that the function is defined in the case of $y = \hat{p} = 0$, which is called an edge case scenario. In the dice loss, the product of y and $\hat{p}$ represents the intersection between GT and the prediction. In other words, dice loss is the negative of the dice coefficient that is used as one of the accuracy measures in the experiments.

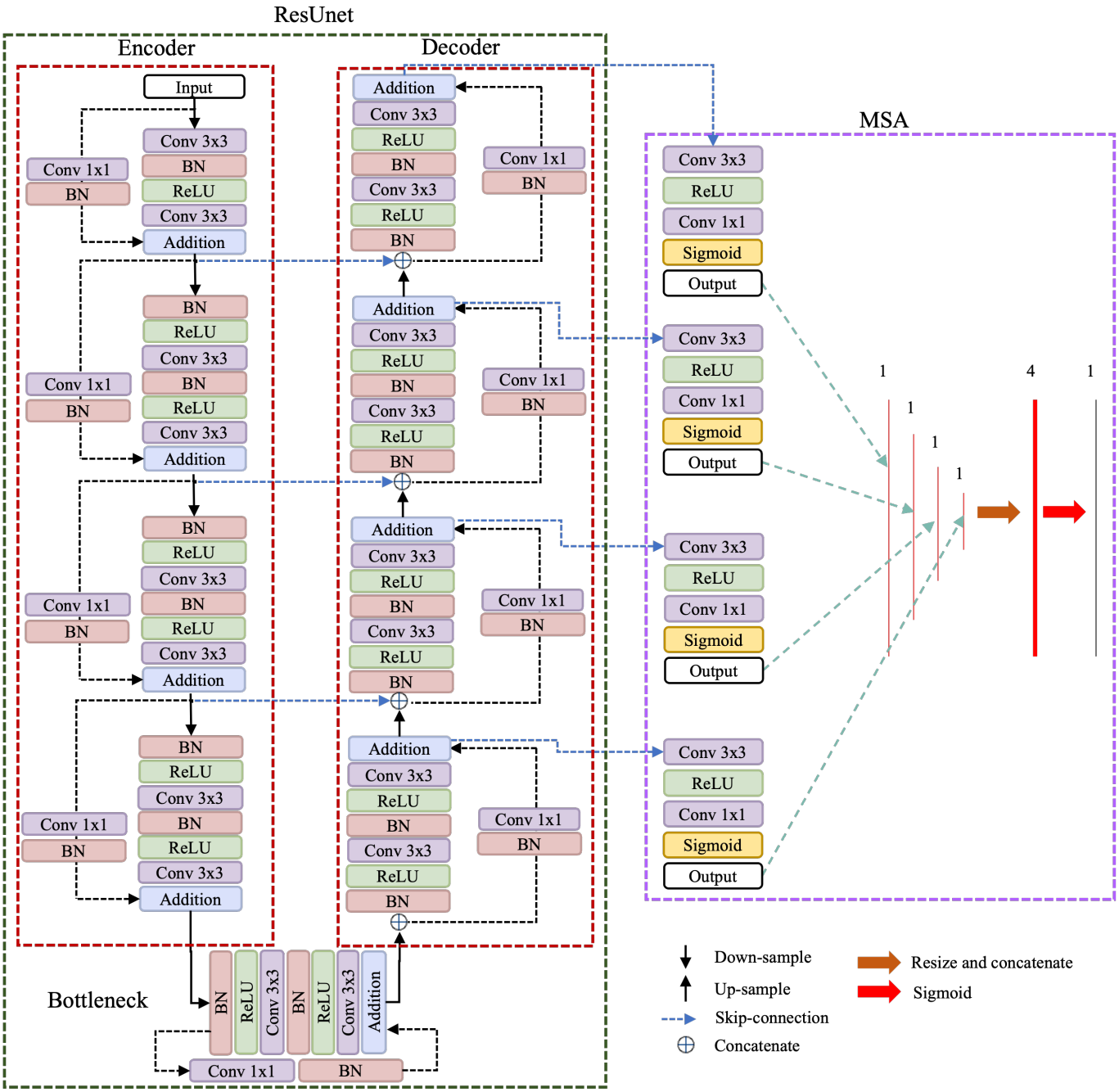

Figure 4. Network architecture of the proposed MSA-ResUNET. The encoder, bottleneck, and skip connections bring the knowledge derived from rich multi-scale feature maps to the decoder of the ResUNET [4]. The MSA-block of the proposed MSA-ResUNET inherits these multi-scale features from all four decoder blocks and applies 3×3 convolutions with ReLU activation function followed by a 1×1 convolution with sigmoid function to produce a segmented map. The maps of all four scales are then aggregated by (i) re-sizing them to input image size; (ii) concatenating together; and (iii) applying a 1×1 convolution with a sigmoid activation function to obtain the final prediction.

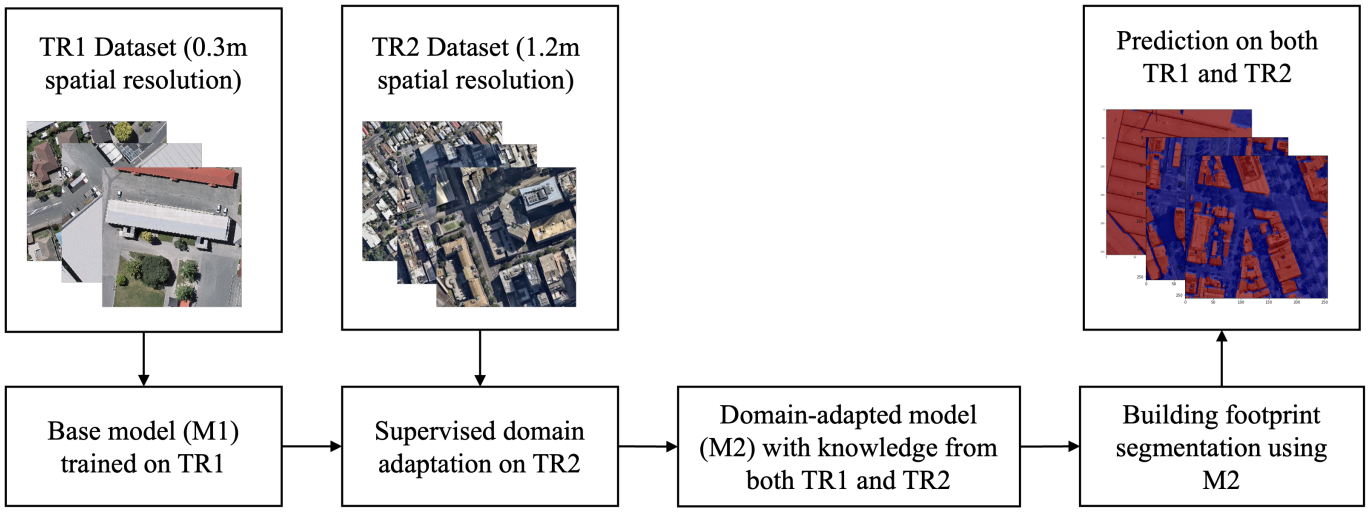

Figure 5. Overall workflow of training and supervised DA of MSA-UNET and MSA-ResUNET.

4. Experiments and Results

In the experimental design, MSA-UNET and MSA-ResUNET are evaluated against the U-Net-like symmetrical encoder–decoder networks: MA-FCN [6], U-Net [3], ResUNET [4], SegNet [5], U-Net++ [37], and U-Net3+ [38] using the two datasets TR1 and TR2. Furthermore, a cross-domain validation is performed to see the effects of domain-shift between the two experimental datasets.

4.1. Evaluation Metrics

The evaluation of the proposed network and method is performed using (i) pixel accuracy; (ii) adjusted accuracy; (iii) F1 score; (iv) IoU; and (v) MCC. Pixel accuracy (Equation (2)) measures how often the predictions and the binary labels match. Adjusted accuracy (Equation (3)) takes the average of the sensitivity (Equation (4)) and specificity (Equation (5)), which measures the proportion of correctly identified actual positives and actual negatives, respectively. F1 score (Equation (6)) and IoU (Equation (7)) are measured from the 'area of overlap' between the prediction and binary labels and 'area of union' (all of the predictions + binary labels - the overlap). Lastly, a measure of the difference between the binary labels and the prediction with a consideration of the ratio between positive and negative elements is calculated by MCC (Equation (8)) [71]. The symbolic representation of the metrics are:

$$Pixel\ accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Adjusted\ accuracy = \frac{Sensitivity + Specificity}{2} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = 1 - \frac{TN}{TN + FP} \tag{5}$$

$$F1score = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{6}$$

$$IoU = \frac{TP}{TP + FN + FP} \tag{7}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (8)$$

where TP is true positive (i.e., prediction = 1, label = 1); FP is false positive (prediction = 1, label = 0); FN is false negative (prediction = 0, label = 1); and TN is true negative (prediction = 0, label = 0).

4.2. Training Details

All the models in the experiments are wrapped in the Keras framework with a mini-batch size of 2. To keep the total number of steps the same or similar, it is set as the ratio of the number of training images to the batch size. The number of steps for TR1 and TR2 is therefore approximately 58K and 61K. A learning rate of 1×10^{-4} is used to train the base model M1, which is changed to the optimal value by lowering it or keeping it the same during DA, such that the minimum validation loss is obtained. *Adam*, *He Normal*, and *ReLU* are the optimiser, initialiser, and activation functions, respectively. A *sigmoid* function is used to obtain the final output maps, and a *dropout* of 50% is used to avoid over-fitting. A dice-coefficient loss is used to monitor the models. All the hyper-parameters are kept the same among all models used for comparison, except in the SegNet, an *argmax* function is used for pooling, unlike the max-pooling in other models. A Macbook M1 Pro with 8-core CPU, 14-core GPU, 14-core NPU, and 16GB RAM is used as the computing machine to account for the computational expense.

4.3. MSA-UNET and MSA-ResUNET

Table 1 presents the performance of the proposed MSA-UNET and MSA-ResUNET in terms of five accuracy measures in TR1 and TR2. The proposed networks perform better in the TR1 dataset than in TR2. This is due to the added complexity in TR2. The complexity comes from 5× smaller samples, 4× less spatial resolution, and the complex high-rise and skyscrapers. MSA-ResUNET outperforms MSA-UNET in TR1 in terms of pixel accuracy, IoU, and MCC. Similarly, MSA-UNET outperforms MSA-ResUNET in TR2 in terms of four measures except in adjusted accuracy.

Table 1. Same-domain validation results of MSA-UNET and MSA-ResUNET.

Dataset	Base Models	Pixel Acc.	Adjusted Acc.	F1 Score	IoU	MCC
TR1	MSA-UNET	0.979	0.863	0.735	0.672	0.718
	MSA-ResUNET	0.981	0.861	0.734	0.673	0.725
TR2	MSA-UNET	0.951	0.722	0.477	0.383	0.453
	MSA-ResUNET	0.928	0.764	0.414	0.321	0.430

4.4. Comparison to the SOTA

The base model M1 of MSA-UNET and MSA-ResUNET is compared against six existing networks from the literature. The comparison is categorised into three variants of FCN: U-Net variant (MSA-UNET and U-Net), ResUNET variant (MSA-ResUNET and ResUNET), and VGG encoder variant (MA-FCN and SegNet). Two other SOTA networks for comparison are U-Net++ and U-Net3+. The experiment is first carried out on both TR1 and TR2 datasets to examine the same-domain validation from the first step of training (Table 2). Same-domain validation is the case of having the training and validation data from the same dataset.

Table 2. Same-domain validation results of the base model (M1) of MSA-UNET on TR1 and TR2 compared to MA-FCN, U-Net, and SegNet. The highest values are highlighted in bold.

Dataset	FCN Variant	Base Models	Pixel Acc.	Adjusted Acc.	F1 Score	IoU	MCC
TR1	U-Net variant	MSA-UNET	0.979	**0.863**	0.735	0.672	0.718
		U-Net	0.973	0.826	**0.844**	**0.770**	0.681
	ResUNET variant	MSA-ResUNET	**0.981**	0.861	0.734	0.673	**0.725**
		ResUNET	0.969	0.833	0.822	0.741	0.665
	VGG-16 encoder	MA-FCN	0.981	0.862	0.729	0.667	**0.725**
		SegNet	0.969	0.831	0.666	0.584	0.673
	U-Net++		0.974	0.828	0.823	0.749	0.685
	U-Net3+		0.981	0.855	0.688	0.615	0.723
TR2	U-Net variant	MSA-UNET	**0.951**	0.722	0.477	0.383	0.453
		U-Net	0.928	0.760	0.437	0.342	**0.479**
	ResUNET variant	MSA-ResUNET	0.928	**0.764**	0.414	0.321	0.430
		ResUNET	0.872	0.745	**0.520**	**0.423**	0.351
	VGG-16 encoder	MA-FCN	0.945	0.729	0.453	0.360	0.436
		SegNet	0.854	0.735	0.376	0.277	0.376
	U-Net++		0.930	0.728	0.602	0.510	0.389
	U-Net3+		0.953	0.778	0.471	0.380	0.496

Compared among the U-Net and ResUNET variants on TR1, MSA-UNET and MSA-ResUNET perform better in terms of pixel accuracy, adjusted accuracy, and MCC when compared to U-Net and ResUNET. Compared among all base networks, MSA-UNET and MSA-ResUNET show better performance in the same three metrics. All models are affected by the complexity of TR2. Despite the complexity, MSA-UNET produces the highest pixel accuracy, F1 score, and IoU when compared among the U-Net variants. MSA-ResUNET produces the highest pixel accuracy, adjusted accuracy, and MCC among the ResUNET variants. The VGG-encoder variants fall behind in both datasets. The other measures except the pixel accuracy are lower because of their bias towards reporting only the positive case. The increment in accuracy measures from the MSA-UNET, MSA-ResUNET, and MA-FCN in TR2 comes from the efficiency of multi-scale aggregation of features in an FCN while training on a smaller and complex dataset. U-Net++ is the most stable in all accuracy measures with a higher F1 score and IoU. U-Net3+ also seems to have a lower F1 score and IoU compared to the other three accuracy measures. Figures 6 and 7 show the sample results from all base models on TR1 and TR2, respectively. As the models are trained using the dice loss, the models with the highest F1 score are saved. The samples, therefore, show a better performance for ResUNET than MSA-ResUNET.

4.5. Effects of Domain Shift

The previous section demonstrated the same-domain validation of the base model (M1) of different networks. M1, however, suffers from the problem of domain-shift and fails during the cross-domain validation, which is the case of having the training and validation data from a different dataset that does not follow the principle of independent and identical properties. To tabulate these effects, Table 3 presents the performance of M1 trained on TR1 (abbr. M1(TR1)) but validated against TR2, and M1 trained on TR2 (abbr. M1(TR2)) validated on TR1. The comparison is only made among the networks that used the concept of multi-scale feature aggregation. Compared to the same-domain validation of MSA-UNET, the cross-domain validation of M1(TR2) on TR1 shows lower pixel accuracy (0.951 vs. 0.909), adjusted accuracy (0.722 vs. 0.576), F1 score (0.477 vs. 0.217), IoU (0.383 vs. 0.138), and MCC (0.453 vs. 0.206). Similarly, M1(TR1) validated on TR2 shows a significantly lower pixel accuracy (0.979 vs. 0.927), adjusted accuracy (0.863 vs. 0.761), F1 score (0.735 vs. 0.522), IoU (0.672 vs. 0.401), and MCC (0.718 vs. 0.499). The

MA-FCN and MSA-ResUNET seem to suffer more from the domain-shift compared to MSA-UNET.

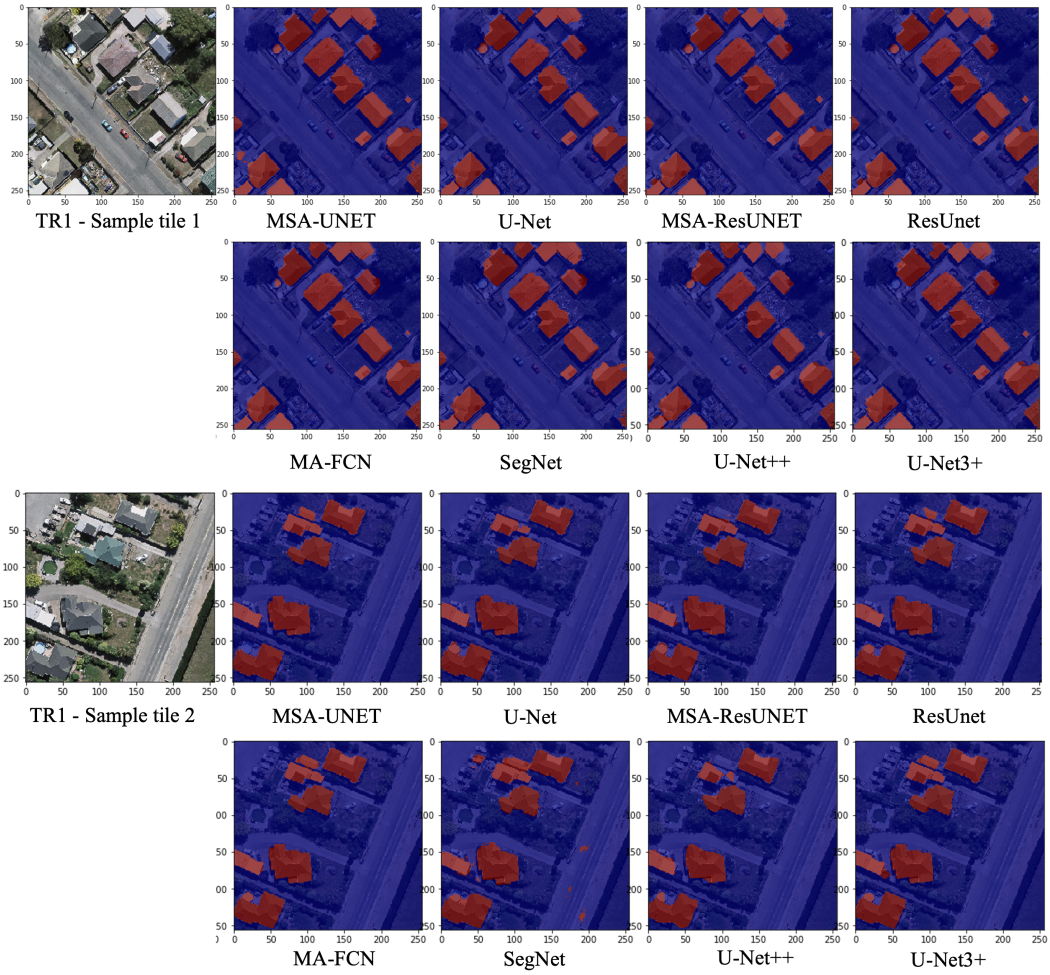

Figure 6. Sample results from the MSA-UNET, U-Net, MSA-ResUNET, ResUNET, MA-FCN, SegNet, U-Net++, and U-Net3+ on the WHU Building dataset (TR1). The left-most images are two different samples from TR1 followed by the corresponding outputs from the eight base models.

Table 3. Cross-domain validation results on TR1 and TR2. M1(TR1) and M1(TR2) refer to the base models trained on TR1 and TR2, respectively. To show the effects of the domain-shift, M1(TR1), and M1(TR2) are validated against TR2 and TR1, respectively.

Evaluation Metrics	M1(TR1) Validated on TR2			M1(TR2) Validated on TR1		
	MA-FCN	MSA-UNET	MSA-ResUNET	MA-FCN	MSA-UNET	MSA-ResUNET
Pixel Acc.	0.907	0.909	0.909	0.927	0.928	0.878
Adjusted Acc.	0.560	0.576	0.538	0.761	0.757	0.772
F1 score	0.154	0.217	0.120	0.522	0.529	0.451
IoU	0.092	0.138	0.072	0.401	0.412	0.327
MCC	0.165	0.206	0.141	0.499	0.500	0.450

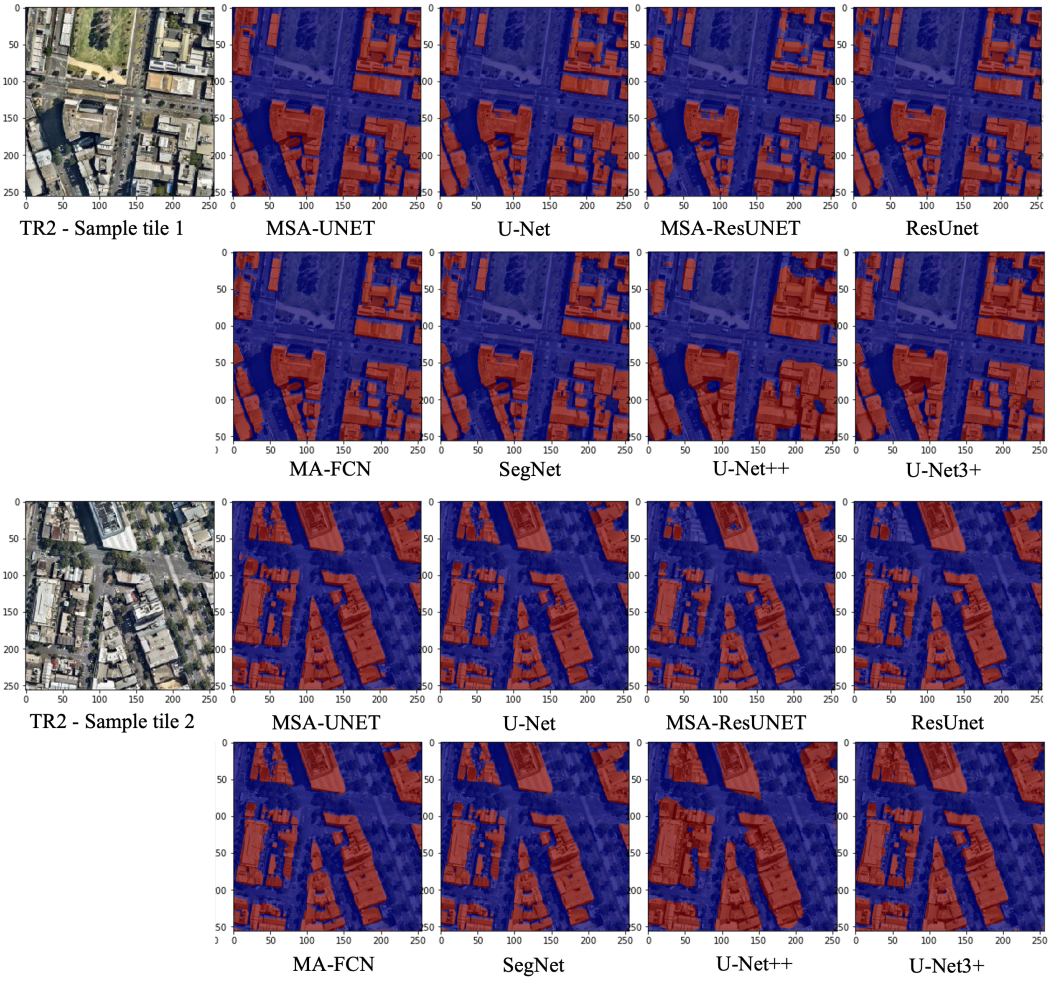

Figure 7. Sample results from the MSA-UNET, U-Net, MSA-ResUNET, ResUNET, MA-FCN, and SegNet, U-Net++, and U-Net3+ on the Melbourne Building dataset (TR2). The left-most images are two different samples from TR2 followed by the corresponding outputs from the eight base models.

4.6. Supervised Domain Adaptation

The domain-shift in the cross-domain validation is minimised by the supervised DA of the base model M1(TR1) on TR2, producing a model M2 that performs well on both datasets. For the DA, the weights and biases are frozen on the encoder side to keep the existing knowledge from TR1, and the learning is only updated on the remaining layers of the decoder and the multi-scale aggregation (MSA) block. Four settings are experimented for updating the learning (DA1, DA2, DA3, and DA4) on MSA-UNET. The experiments are only performed for the MSA-UNET because of its relatively higher performance during the cross-domain validation. DA1 represents the case where the weight and biases of the *MSA* and the decoder blocks of the U-Net ($UNet_{dec}$) in MSA-UNET are updated during the backpropagation of the error. In DA2, only the $UNet_{dec}$ layers are updated with the MSA frozen. Similarly, DA3 is a case where the last block of the U-Net ($UNet_{last}$) and the MSA are updated. In DA4, only the $UNet_{last}$ is updated without the MSA. A schematic representation of the four settings is shown in Figure 8. The learning rate is kept the same during this experiment as changing it decreased the accuracy measures in the experiments.

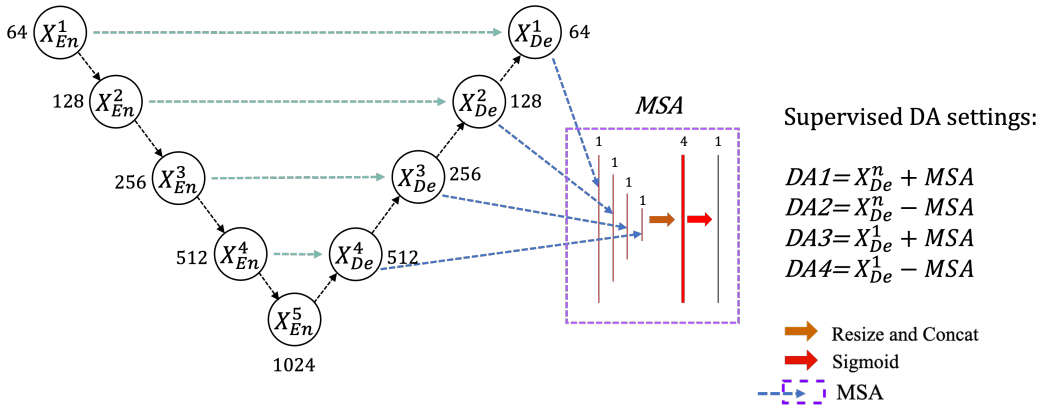

Figure 8. A schematic representation of the four settings of supervised DA. X^n_{En} where $n \in [1,4]$ represents the encoder blocks, X^5_{En} represents the bottleneck, X^n_{De} where $n \in [1,4]$ represents the decoder blocks, and MSA represents the multi-scale feature aggregation block of our MSA-UNET.

The accuracy measures obtained on the TR2 dataset along with the time taken for the four experimental settings to run one step of training during the four settings of DA are shown in Table 4. The results show that setting DA1 produces the highest F1 score and IoU, DA2 produces the highest pixel accuracy and MCC, and DA4 produces the highest adjusted accuracy. The pixel accuracy is biased toward only reporting the positive case, and DA1 is the setting with two leading accuracy measures. Therefore, DA1 is used as the setting for carrying out the supervised DA on MSA-UNET and MSA-ResUNET for the rest of the experiments.

Table 4. Experiments carried out in different settings of supervised DA. The highest values are highlighted in bold.

Evaluation Metric	DA1	DA2	DA3	DA4
Time/step (ms)	256	242	152	137
Pixel Acc.	0.942	**0.943**	0.937	0.934
Adjusted Acc.	0.736	0.740	0.712	**0.741**
F1 score	**0.649**	0.486	0.586	0.449
IoU	**0.558**	0.388	0.485	0.346
MCC	0.431	**0.453**	0.402	0.427

Table 5 compares the performance of both M1 (before DA) and M2 (after DA) on both TR1 and TR2. Starting with MSA-UNET, M2 shows a significantly higher performance against M1(TR1) when validated against TR2: pixel accuracy (0.942 vs. 0.909), adjusted accuracy (0.736 vs. 0.576), F1 score (0.649 vs. 0.217), IoU (0.558 vs. 0.138), and MCC (0.431 vs. 0.206). Similarly, the domain-adapted MSA-UNET (M2) also shows a significantly higher performance over M1(TR2) when validated on TR1 in terms of pixel accuracy (0.949 vs. 0.928), adjusted accuracy (0.828 vs. 0.757), F1 score (0.710 vs. 0.529), IoU (0.602 vs. 0.412), and MCC (0.606 vs. 0.5). MSA-ResUNET also shows a similar improvement in all accuracy measures on both TR1 and TR2. The domain-adapted MSA-UNET performs better on both datasets. Some sample results from the cross-domain validation of M1 and M2 using MSA-UNET are shown in Figure 9.

Table 5. Cross-domain validation results of MSA-UNET and MSA-ResUNET on TR1 and TR2. M1(TR1) and M1(TR2) refer to the base models trained on TR1 and TR2, respectively. M2 refers to the model previously trained on TR1 and supervised DA on TR2. The M1 and M2, therefore, provide before and after comparisons of the DA. The highest values are highlighted in bold.

Network	Evaluation Metrics	M1(TR1) Validated on TR2 (before DA)	M2 on TR2 (after DA)	M1(TR2) Validated on TR1 (before DA)	M2 on TR1 (after DA)
MSA-UNET	Pixel Acc.	0.909	0.942	0.928	0.949
	Adjusted Acc.	0.576	0.736	0.757	0.828
	F1 score	0.217	0.649	0.529	0.710
	IoU	0.138	0.558	0.412	0.602
	MCC	0.206	0.431	0.500	0.606
MSA-ResUNET	Pixel Acc.	0.909	0.945	0.878	0.926
	Adjusted Acc.	0.538	0.767	0.772	0.809
	F1 score	0.120	0.627	0.451	0.599
	IoU	0.072	0.531	0.327	0.486
	MCC	0.141	0.452	0.450	0.552

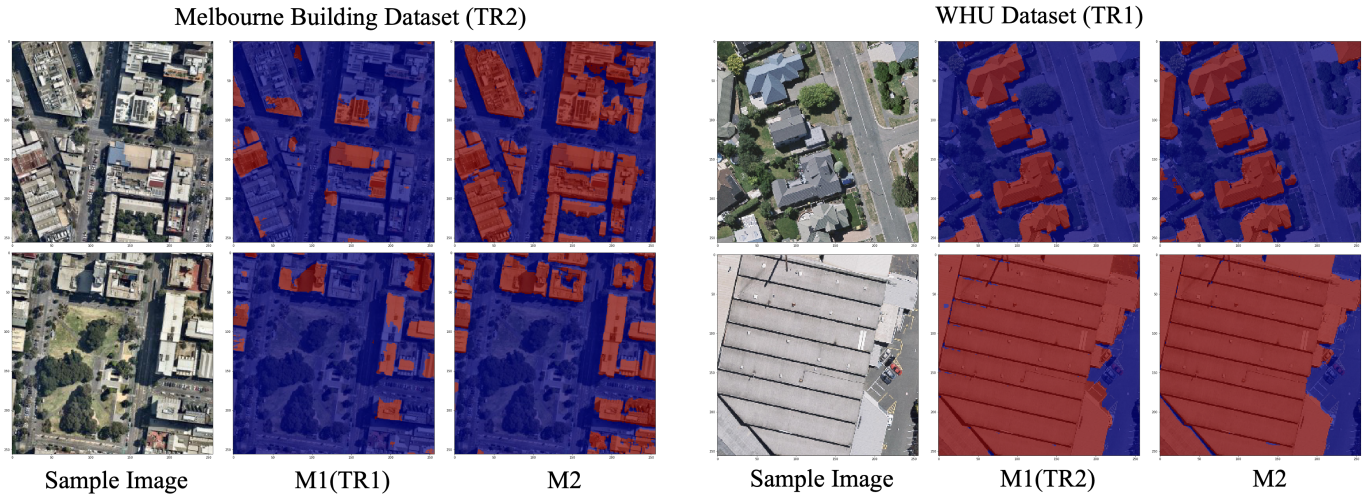

Figure 9. Sample results from cross-domain validation of the base model (M1) and domain-adapted model (M2) experimented with MSA-UNET. The two samples on the left compare M1(TR1) to M2 to show before and after supervised DA on the TR2 dataset. Similarly, the two samples on the right compare M1(TR2) to M2 to show before and after supervised DA on the TR1 dataset.

5. Discussion

5.1. Domain-Shift and Supervised Domain Adaptation

The experiments have shown that the performance of the base models is significantly reduced during cross-domain validation because of the domain-shift between the TR1 and TR2 datasets. Transfer learning methods such as supervised DA are useful in tackling this problem. Supervised DA from a higher resolution domain to a lower resolution domain is not new in VHR EO image classification. Wurm et al. (2019) [63] transferred the knowledge from an FCN trained on Quickbird's 0.5 m multi-spectral imagery (resampled from 2 m multi-spectral and 0.5 panchromatic image channels) to Sentinel-2 10 m imagery. They also experiment with the transfer learning from Quickbird to 6 m SAR imagery from TerraSAR-X. In their experiment, they conclude that transfer learning from the Quickbird domain to the Sentinel-2 domain improves the performance of semantic segmentation. It was observed that the method performs well between the source and target domain of optical imagery, and fails when the source is optical and the target is SAR imagery. This work has investigated a similar method of supervised DA on MSA-UNET to transfer the knowledge gained from VHR imagery of 0.3 m (TR1) to that of 1.2 m (TR2). The supervised DA

demonstrates an increase in accuracy measures by up to four times, which is a significant increase in robustness in the cross-domain setting.

5.2. Melbourne Building Dataset in Cross-Domain Validation

The complexity of the TR2 dataset demonstrates increased robustness in the cross-domain setting from the proposed networks when compared to the networks trained on TR1. Table 5 validates the increase in robustness in terms of all five accuracy measures. From the experiment, the domain-shift effect on the base model M1(TR2) is smaller than the M1(TR1). Looking at MSA-UNET, M1(TR2) produced a 29.4% higher MCC (0.500 vs. 0.206) when compared to M1(TR1). This shows that the developed Melbourne Building dataset (TR2) is 29.4% more robust in terms of MCC during the cross-validation of M1 among TR1 and TR2. Similarly, TR2 is approximately 18%, 31%, and 27% more robust in terms of adjusted accuracy, F1 score, and IoU, respectively. The complexity of TR2 has improved the robustness of both the dataset and the two proposed networks; however, there exist some limitations to the dataset, as discussed in the next section.

5.3. Limitations and Future Direction

Multi-scale feature aggregation increases the precision of MSA-UNET in complex urban building footprint segmentation, as shown by the experiments. The increased precision comes from the significant minimisation in over-segmentation that occurs when the "background" pixels are predicted as "buildings". The comparison of over-segmentation in the TR2 dataset using SegNet, U-Net, MA-FCN, and MSA-UNET is shown by their prediction samples in Figure 10. As seen, the MSA-UNET is less affected by over-segmentation. However, the problem of under-segmentation, where the buildings are recognised as background, persists in the TR2 dataset. This comes from the existence of multiple roof types of different shapes and heights within a single building, the fuzzy boundary between adjacent buildings on 1.2 m spatial resolution images, and the shadows from high-rise buildings and skyscrapers in the TR2 dataset. The complexity of TR2 has improved its robustness; however, the dataset can be further improved by adding more samples of multiple spatial resolutions from other cities.

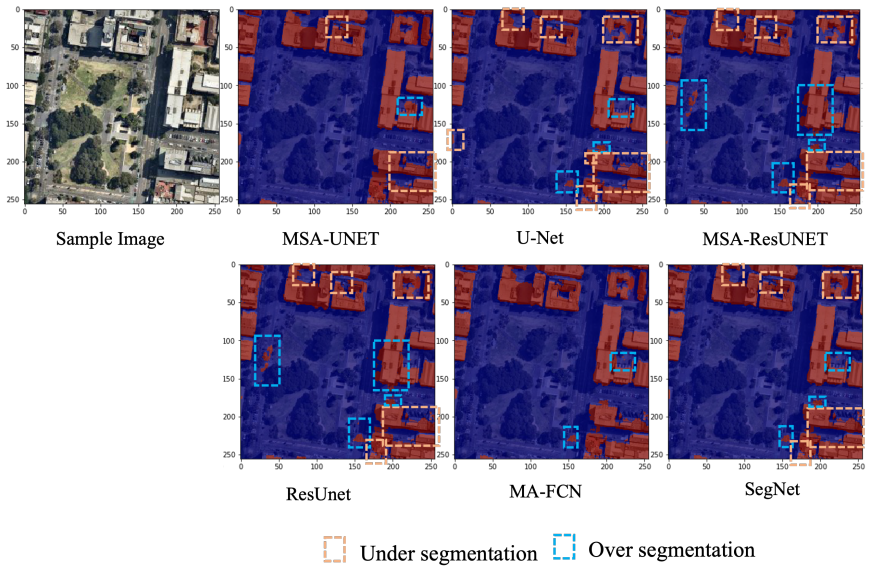

Figure 10. Sample results from the MSA-UNET, U-Net, MSA-ResUNET, ResUNET, MA-FCN, and SegNet (left-right) on the Melbourne Building dataset (TR2). The orange and blue region highlights the pixels affected by under- and over-segmentation, respectively.

6. Conclusions

DL methods are the current practices in extracting building footprints from EO images. U-Net and ResUNET are two commonly used DL networks of an FCN family. A CNN is the fundamental backbone of an FCN. Despite setting a positive advancement in feature extraction, CNN-based DL methods have several limitations. This study provides potential solutions to two major limitations associated with these DL methods. The first limitation comes from the loss of context information inside the layers of CNNs and the second comes from the inability of CNNs to robustly extract features in cross-domain settings due to the domain-shift problem. For the first limitation, this study integrates the partial concepts of FPN to aggregate the multi-scale feature maps of U-Net and ResUNET. The integration results in two novel DL networks called MSA-UNET and MSA-ResUNET. The second limitation is investigated by preparing an experimental setup of two building footprint datasets and performing the supervised domain adaptation of the developed network between the datasets. The first dataset is the benchmark WHU Building dataset and the second is a newly developed dataset from the complex urban setting of the City of Melbourne. The proposed solutions to the two limitations of CNN-based methods demonstrate significant efficiency gains. The developed networks outperform other SOTA networks such as MA-FCN, U-Net, SegNet, U-Net++, and U-Net3+ in terms of several accuracy measures. The MSA-UNET and MSA-ResUNET, respectively, outperform U-Net and ResUNET on the two building footprint datasets in terms of pixel accuracy, adjusted accuracy, and MCC. MSA-UNET also outperforms U-Net on the latter dataset in terms of pixel accuracy, F1 score, and IoU. The proposed networks also outperform MA-FCN, which implements the multi-feature aggregation of a similar nature. The experiments therefore conclude that the proposed multi-scale feature aggregation technique results in efficiency gains of FCNs such as U-Net and ResUNET. Unlike most experiments from the literature, this study performs a cross-domain validation between the two datasets. This validation allows a comprehensive study of the effects of domain shift that come from several differences between the datasets. The second dataset developed for this study has $5\times$ fewer samples with $4\times$ lower spatial resolution and includes complex high-rise buildings and skyscrapers when compared to the benchmark dataset. The cross-domain validation of the proposed networks demonstrates the reduction in accuracy measures by two to three times in percentage, concluding that the FCNs fail in cross-domain settings due to the domain-shift problem. Four supervised DA settings are studied to minimise the effects of domain-shift, resulting in significant improvement in the robustness of MSA-UNET and MSA-ResUNET in terms of all accuracy measures. The experiments also demonstrate that the developed small and complex datasets result in more robust FCNs in cross-domain settings when compared to the FCNs trained on the benchmark dataset. The developed high-resolution building footprint dataset can be further used by the geospatial industries as well as researchers in the domain of city planning and urban feature modelling. In future works, the use of robust post-processing techniques and the development of more robust datasets is recommended to increase accuracy in building footprint extraction in complex city environments.

Author Contributions: Conceptualisation, methodology, investigation, validation, visualisation, and writing—review and editing: B.N. and J.A.; data curation, software, writing—original draft preparation and formal analysis: B.N.; supervision, resources, project administration, and funding acquisition: J.A. All the authors read, edited, and critiqued the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research work is funded by the internal grant to Jagannath Aryal from the University of Melbourne.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Nearmap for providing the API service to collect the image data for the experiments. The authors would like to acknowledge the University of Melbourne for providing an internal research grant to the first author to carry out this work.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Neupane, B.; Horanont, T.; Aryal, J. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sens.* **2021**, *13*, 808. [CrossRef]
2. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
3. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
4. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
5. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
6. Wei, S.; Ji, S.; Lu, M. Toward automatic building footprint delineation from aerial images using CNN and regularization. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2178–2189. [CrossRef]
7. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
8. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
9. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
10. Stein, A.; Aryal, J.; Gort, G. Use of the Bradley-Terry model to quantify association in remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 852–856. [CrossRef]
11. Shadman Roodposhti, M.; Aryal, J.; Lucieer, A.; Bryan, B.A. Uncertainty assessment of hyperspectral image classification: Deep learning vs. random forest. *Entropy* **2019**, *21*, 78. [CrossRef]
12. Neupane, B.; Horanont, T.; Duy, H.N.; Suebvong, S.; Mahattanawutakorn, T. An Open-Source UAV Image Processing Web Service for Crop Health Monitoring. In Proceedings of the 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), Toyama, Japan, 7–11 July 2019; pp. 11–16.
13. Neupane, B.; Horanont, T.; Hung, N.D. Deep learning based banana plant detection and counting using high-resolution red-green-blue (RGB) images collected from unmanned aerial vehicle (UAV). *PLoS ONE* **2019**, *14*, e0223906. [CrossRef]
14. Neupane, B.; Horanont, T.; Aryal, J. Real-Time Vehicle Classification and Tracking Using a Transfer Learning-Improved Deep Learning Network. *Sensors* **2022**, *22*, 3813. [CrossRef]
15. Xie, Y.; Cai, J.; Bhojwani, R.; Shekhar, S.; Knight, J. A locally-constrained yolo framework for detecting small and densely-distributed building footprints. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 777–801. [CrossRef]
16. Aryal, J.; Dutta, R. Smart city and geospatiality: Hobart deeply learned. In Proceedings of the 2015 31st IEEE International Conference on Data Engineering Workshops, Seoul, Republic of Korea, 13–17 April 2015; pp. 108–109.
17. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
20. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June – 1 July 2016 ; pp. 770–778.
22. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
23. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
24. Mnih, V.; Hinton, G.E. Learning to detect roads in high-resolution aerial images. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 210–223.
25. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.

26. Saito, S.; Aoki, Y. Building and road detection from large aerial imagery. In Proceedings of the Image Processing: Machine Vision Applications VIII. International Society for Optics and Photonics, San Francisco, CA, USA, 8–12 February 2015 ; Volume 9405, p. 94050K.
27. Saito, S.; Yamashita, T.; Aoki, Y. Multiple object extraction from aerial imagery with convolutional neural networks. *Electron. Imaging* **2016**, *2016*, 1–9. [CrossRef]
28. Vakalopoulou, M.; Karantzalos, K.; Komodakis, N.; Paragios, N. Building detection in very high resolution multispectral data with deep learning features. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 1873–1876.
29. Zhong, Z.; Li, J.; Cui, W.; Jiang, H. Fully convolutional networks for building and road extraction: Preliminary results. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1591–1594.
30. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645–657. [CrossRef]
31. Marcu, A.E.; Leordeanu, M. Object contra context: Dual local-global semantic segmentation in aerial images. In Proceedings of the Workshops at the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
32. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building extraction from satellite images using mask R-CNN with building boundary regularization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 247–251.
33. Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B. Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. [CrossRef]
34. Griffiths, D.; Boehm, J. Improving public data for building segmentation from Convolutional Neural Networks (CNNs) for fused airborne lidar and image data using active contours. *ISPRS J. Photogramm. Remote Sens.* **2019**, *154*, 70–83. [CrossRef]
35. Zhang, M.; Gao, F.; Dong, J.; Qi, L. Multi-Scale Feature Fusion for Hyperspectral and Lidar Data Joint Classification. In Proceedings of the IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022 ; pp. 2856–2859.
36. Huang, L.; Chen, C.; Yun, J.; Sun, Y.; Tian, J.; Hao, Z.; Yu, H.; Ma, H. Multi-Scale Feature Fusion Convolutional Neural Network for Indoor Small Target Detection. *Front. Neurorobot.* **2022**, *16*, 881021. [CrossRef]
37. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef]
38. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
39. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
40. Li, W.; He, C.; Fang, J.; Zheng, J.; Fu, H.; Yu, L. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sens.* **2019**, *11*, 403. [CrossRef]
41. Yi, Y.; Zhang, Z.; Zhang, W.; Zhang, C.; Li, W.; Zhao, T. Semantic segmentation of urban buildings from vhr remote sensing imagery using a deep convolutional neural network. *Remote Sens.* **2019**, *11*, 1774. [CrossRef]
42. Bischke, B.; Helber, P.; Folz, J.; Borth, D.; Dengel, A. Multi-task learning for segmentation of building footprints with deep neural networks. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1480–1484.
43. Qin, Y.; Wu, Y.; Li, B.; Gao, S.; Liu, M.; Zhan, Y. Semantic segmentation of building roof in dense urban environment with deep convolutional neural network: A case study using GF2 VHR imagery in China. *Sensors* **2019**, *19*, 1164. [CrossRef] [PubMed]
44. Abdollahi, A.; Pradhan, B.; Alamri, A.M. An Ensemble Architecture of Deep Convolutional Segnet and Unet Networks for Building Semantic Segmentation from High-resolution Aerial Images. *Geocarto Int.* **2020**, 1–13.. [CrossRef]
45. Pan, Z.; Xu, J.; Guo, Y.; Hu, Y.; Wang, G. Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net. *Remote Sens.* **2020**, *12*, 1574. [CrossRef]
46. Sariturk, B.; Bayram, B.; Duran, Z.; Seker, D.Z. Feature Extraction from Satellite Images Using Segnet and Fully Convolutional Networks (FCN). *Int. J. Eng. Geosci.* **2020**, *5*, 138–143. [CrossRef]
47. Ayala, C.; Sesma, R.; Aranda, C.; Galar, M. A Deep Learning Approach to an Enhanced Building Footprint and Road Detection in High-Resolution Satellite Imagery. *Remote Sens.* **2021**, *13*, 3135. [CrossRef]
48. Jian, M.; Wang, J.; Yu, H.; Wang, G.; Meng, X.; Yang, L.; Dong, J.; Yin, Y. Visual saliency detection by integrating spatial position prior of object with background cues. *Expert Syst. Appl.* **2021**, *168*, 114219. [CrossRef]
49. Yang, D.; Liu, G.; Ren, M.; Xu, B.; Wang, J. A multi-scale feature fusion method based on U-Net for retinal vessel segmentation. *Entropy* **2020**, *22*, 811. [CrossRef]
50. Su, R.; Zhang, D.; Liu, J.; Cheng, C. MSU-net: Multi-scale U-net for 2D medical image segmentation. *Front. Genet.* **2021**, *12*, 639930. [CrossRef]
51. Lu, D.; Cheng, S.; Wang, L.; Song, S. Multi-scale feature progressive fusion network for remote sensing image change detection. *Sci. Rep.* **2022**, *12*, 11968. [CrossRef] [PubMed]

52. Liu, Y.; Fan, B.; Wang, L.; Bai, J.; Xiang, S.; Pan, C. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 78–95. [CrossRef]
53. Yang, H.; Wu, P.; Yao, X.; Wu, Y.; Wang, B.; Xu, Y. Building extraction in very high resolution imagery by dense-attention networks. *Remote Sens.* **2018**, *10*, 1768. [CrossRef]
54. Wu, G.; Shao, X.; Guo, Z.; Chen, Q.; Yuan, W.; Shi, X.; Xu, Y.; Shibasaki, R. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sens.* **2018**, *10*, 407. [CrossRef]
55. Chen, Q.; Wang, L.; Wu, Y.; Wu, G.; Guo, Z.; Waslander, S.L. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *arXiv* **2018**, arXiv:1807.09532.
56. Ji, S.; Wei, S.; Lu, M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int. J. Remote Sens.* **2019**, *40*, 3308–3322. [CrossRef]
57. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
58. Bozinovski, S.; Fulgosi, A. The influence of pattern similarity and transfer learning upon training of a base perceptron B2. In Proceedings of the Symposium Informatica, 1976; Volume 3, pp. 121–126. Available online: https://www.informatica.si/index.php/informatica/article/view/2828 (accessed on 25 November 2022).
59. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 9. [CrossRef]
60. Saha, A.; Rai, P.; Daumé, H.; Venkatasubramanian, S.; DuVall, S.L. Active supervised domain adaptation. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Athens, Greece, 5–9 September 2011; pp. 97–112.
61. Motiian, S.; Piccirilli, M.; Adjeroh, D.A.; Doretto, G. Unified deep supervised domain adaptation and generalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5715–5725.
62. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning. *Remote Sens.* **2019**, *11*, 83. [CrossRef]
63. Wurm, M.; Stark, T.; Zhu, X.X.; Weigand, M.; Taubenböck, H. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 59–69. [CrossRef]
64. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
65. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
66. Liu, W.; Su, F.; Jin, X.; Li, H.; Qin, R. Bispace Domain Adaptation Network for Remotely Sensed Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2020** . [CrossRef]
67. Neupane, B.; Aryal, J.; Rajabifard, A. Building Footprint Segmentation using Transfer Learning: A case study of the City of Melbourne. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2022**, *10*, 173–179. [CrossRef]
68. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A survey on deep transfer learning. In Proceedings of the International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; pp. 270–279.
69. Taormina, V.; Cascio, D.; Abbene, L.; Raso, G. Performance of fine-tuning convolutional neural networks for HEP-2 image classification. *Appl. Sci.* **2020**, *10*, 6940. [CrossRef]
70. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin, Germany, 2017; pp. 240–248.
71. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]