remote sensing

MDPI

Article

AMM-FuseNet: Attention-Based Multi-Modal Image Fusion Network for Land Cover Mapping

Wanli Ma *, Oktay Karakuş and Paul L. Rosin

School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, UK; karakuso@cardiff.ac.uk (O.K.); rosinpl@cardiff.ac.uk (P.L.R.)
* Correspondence: maw13@cardiff.ac.uk

Abstract: Land cover mapping provides spatial information on the physical properties of the Earth's surface for various classes of wetlands, artificial surface and constructions, vineyards, water bodies, etc. Having reliable information on land cover is crucial to developing solutions to a variety of environmental problems, such as the destruction of important wetlands/forests, and loss of fish and wildlife habitats. This has made land cover mapping become one of the most widespread applications in remote sensing computational imaging. However, due to the differences between modalities in terms of resolutions, content, and sensors, integrating complementary information that multi-modal remote sensing imagery exhibits into a robust and accurate system still remains challenging, and classical segmentation approaches generally do not give satisfactory results for land cover mapping. In this paper, we propose a novel dynamic deep network architecture, *AMM-FuseNet* that promotes the use of multi-modal remote sensing images for the purpose of land cover mapping. The proposed network exploits the hybrid approach of the channel attention mechanism and densely connected atrous spatial pyramid pooling (DenseASPP). In the experimental analysis, in order to verify the validity of the proposed method, we test AMM-FuseNet with three datasets whilst comparing it to the six state-of-the-art models of DeepLabV3+, PSPNet, UNet, SegNet, DenseASPP, and DANet. In addition, we demonstrate the capability of AMM-FuseNet under minimal training supervision (reduced number of training samples) compared to the state of the art, achieving less accuracy loss, even for the case with 1/20 of the training samples.

Keywords: multi-modal fusion; channel attention; land cover mapping

Citation: Ma, W.; Karakuş, O.; Rosin, P.L. AMM-FuseNet: Attention-Based Multi-Modal Image Fusion Network for Land Cover Mapping. *Remote Sens.* **2022**, *14*, 4458. https://doi.org/10.3390/rs14184458

Academic Editors: Jukka Heikkonen, Fahimeh Farahnakian and Dimitrios Makris

Received: 20 July 2022
Accepted: 5 September 2022
Published: 7 September 2022

1. Introduction

During the past few decades, human activities have posed serious threats to the environment, such as over-logging, over mining, illegal hunting, plastic pollution [1], which makes it necessary to monitor the Earth for the purpose of preventing damages to the environment. With the rapid development of remote sensing technology in the last couple of decades, various space- or air-borne remote sensing sensors have been made available to provide useful and large-scope information about the Earth, such as forest cover, glacier conditions, ocean surface, urban construction. Thus, utilizing remote sensing images for the purpose of environmental applications has become a feasible solution, but there are still many challenges for using remote sensing images in various environmental applications. To name but a few, (1) it is expensive and challenging to obtain high-resolution images for all possible problematic areas; (2) passive remote sensors (e.g., optical) are at the mercy of the clouds and the amount of sunshine; and (3) the contents of remote sensing images are generally very complicated and difficult to analyze.

In order to help overcome the aforementioned challenges and beyond, taking the advantage of leveraging different remote sensing modalities is a potential solution. Various environmental applications can benefit from multi-modal image fusion by exploiting complementary features provided by different types of remote sensors. Specifically, for

active-passive sensor data fusion, passive-optical sensors play the role of feeding the system with high spectral resolution of the Earth surface, which are useful for image analysis. However, for optical remote sensing imagery, in order to provide multi-spectral information, they tend to reduce spatial resolution so as to maintain acceptable bandwidth [2,3]. Additionally, this type of sensor is subject to the weather conditions and only provide useful information during the daytime and under clear weather conditions. On the contrary, active remote sensing sensors are able to acquire images without being affected by inclement weather conditions, such as heavy fog, storm, sandstorm, snowstorm. Additionally, they usually provide sufficient textural and structural information of observed objects [4]; however, they are mostly not capable of collecting color/spectral information. Lastly, since synthetic aperture radar (SAR) images are obtained via wave reflection, they have an important problem that degrades statistical inference, which is the presence of multiplicative speckle noise. The received back-scattered signals sum up coherently and then undergo nonlinear transformations. This in turn gives the resulting images a granular appearance, which is referred to as speckle noise [5]. Considering all these advantages and disadvantages, exploring remote sensing data coming from different modalities becomes crucial for many environmental applications via making use of the complementary advantages of each type of sensors.

Land cover mapping is one of the most widespread and important remote sensing applications in the literature. This is because, nowadays, decisions that concern the environment made by governments, politicians or organizations highly depend on adequate information for many complex interrelated aspects, where land cover/use is one such aspect. Furthermore, an improved understanding of land cover can help act to solve environmental problems, such as disorganized construction, loss of prime agricultural lands, destruction of important wetlands or forests, and loss of fish and wildlife habitats [6].

Land cover classification or mapping is a long-established application area that has been developing since the 1970s. The earliest Landsat land cover classification approach was mostly based on visual and manual approaches. This was done by drawing boundaries of different land cover types and marking each of the land cover classes [7]. In the late 1970s, with the development of computer technology, digital image analysis has become more widespread, and some platforms such as geographic information systems (GIS) were developed to make the analysis of remote sensing data more convenient. Following this, utilizing computer-based approaches for the purpose of land cover classification has become the common practice by geographic analysis specialists. In addition, due to the development of early automatic image processing methods, such as smoothing, sharpening and feature extraction [8], geographic experts have been able to use various traditional image processing algorithms to help perform land cover classification. Although one can generate digital land cover maps by using computers, manual annotation is generally required, which is time consuming and labor intensive. Specifically, in cases when the target scene accommodates plenty of objects to be classified, and the scene covers huge areas, manual annotation becomes more challenging.

In recent years, with the great success of deep learning in computer vision, automated land cover classification approaches have been significantly improved, which assign a class for each pixel among many particular classes, such as artificial surfaces, cultivated areas, and water bodies, as shown in Figure 1. Due to the similarity of land cover mapping and semantic segmentation, researchers have started to use segmentation networks to perform land cover mapping. Furthermore, machine learning based end-to-end frameworks make use of remote sensing data (spatial and spectral information) to achieve better performance in land cover classification compared to the traditional pixel-based methods [9]. However, considering the diversity and complexity of remote sensing data and along with imbalanced training samples, it is still challenging to achieve high performance for land cover classification.

Figure 1. Land cover mapping example. Image taken from ESA (available online: https://www.esa.int/ESA_Multimedia/Images/2020/03/Europe_land-cover_mapped_in_10_m_resolution, accessed on 20 June 2022).

For the purpose of improving the performance of land cover classification, multi-modal data fusion is an important choice whilst exploiting complementary features of different modalities. Although, in the current circumstances, multi-modal remote sensing data are available with the development of remote sensing sensors and observation techniques (e.g., active and passive), the literature is still far from fully leveraging the advantages of using multiple modalities for land cover classification. In order to successfully implement multi-modal remote sensing image fusion for environmental applications, there generally are two types of fusion approaches: (1) machine learning based; and (2) traditional methods, such as component substitution (CS) and multi-scale decomposition (MSD).

Although classical/traditional image fusion methods have been well studied for a few decades, there still are various challenges; for example, (i) precise and complex registration processing is required before the fusion step; (ii) it is highly dependent on the correlation between images being fused; and (iii) it is likely to lose information during the fusion process while replacing a part of the component of the original data during the processing.

On the other hand, ML-based methods generally demonstrate more powerful outcomes for image fusion. Thus, utilizing machine learning approaches in remote sensing imagery related applications has become a hot topic in the literature. However, since the contents in remote sensing images appear very different to the classical natural images, widely used network structures for natural images are not capable of and optimal for

processing remote sensing imagery. Meanwhile, the machine/deep learning approaches of remote sensing data fusion, especially for the environmental applications, can still be seen in early stages. Therefore, more robust and generalizable machine learning based methods, specifically for remote sensing data fusion, need to be explored and developed in order to provide suitable and accurate solutions for land cover applications, and beyond.

In this paper, we propose a novel dynamic deep network architecture, AMM-FuseNet, for the purposes of the land cover mapping application. AMM-FuseNet promotes the use of multi-modal remote sensing images whilst exploiting the hybrid approach of the channel attention mechanism and densely connected atrous spatial pyramid pooling (DenseASPP). In order to verify the validity of the proposed method, we test AMM-FuseNet under four test cases from three datasets (Potsdam [10], DFC2020 [11] and Hunan [12]). A comparative study is implemented to test AMM-FuseNet performance against six state-of-the-art network architectures of DeepLab V3+ [13], PSPNet [14], UNet [15], SegNet [16], DenseASPP [17], and DANet [18]. The contributions of this paper are as follow:

1. We design a novel encoder module, which combines a channel-attention mechanism and densely connected atrous spatial pyramid pooling (DenseASPP) module. This proposed feature extraction module enhances the representational power of the network by successfully weighting the output features obtained by the atrous convolution. This module can be easily extended to any other networks with an encoder–decoder structure.
2. We propose a machine learning based land cover mapping method specifically suitable for multi-modal remote sensing image fusion. The proposed network extracts information from multiple modalities in a parallel fashion, but performs training with a single loss function to make use of their complementary features. Meanwhile, the encoders in a parallel fashion show a better ability to cope with minimal (small number of) training samples. This has been experimentally validated in a set of test cases (Section 5.3), where we gradually reduce the number of training samples and measure the model performance using the same test sample set.
3. The proposed hybrid network exploits and combines many advantages of existing networks for the purpose of improving the performance of land cover mappings. The encoder of the proposed network combines two feature extraction modules (ResNet and Dense ASPP) in a parallel fashion to improve the feature extraction capabilities for each modality. In order to make more efficient use of the extracted features, skip connections are used to benefit from the low-, middle-, and high-level features at the same time. The proposed multi-modal image fusion network shows competitive performance for land cover mapping and outperforms the state of the art.

The rest of the paper is organized as follows: a general background and literature review are presented in Section 2, whilst Section 3 presents the proposed method AMM-FuseNet. Section 4 gives details of the datasets we have used, and Section 5 covers the experimental analysis. Section 6 concludes the paper with a brief summary and future works.

2. Related Work

Following the development in big data research area and its effects on computer vision research, especially in recent years, multi-modal remote sensing data for various applications have been made available under open-access licences (e.g., ESA Sentinel-1/2, fusion contest datasets, including DEM, airborne/UAV-based optical data). Thanks to their complementary features, multi-modal remote sensing imagery provides much richer information compared to single modality, especially for land cover/use applications. However, in the literature, most of the land cover mapping papers still use single modality data [19–21]. Along with the technical developments in computational imaging and deep/machine learning research, the usage of multi-modal for land cover mapping data [22,23], despite being in early stages and insufficient, have started to appear in some works in the recent years.

Considering the increasing demand for multi-modal information for land cover mapping in the literature, obviously, the key challenge of this research lies within answering this research question: "*how to make efficient use of complementary features in multi-modal remote sensing data?*". One of the most common answer to this question in the literature is to implement an image fusion approach which directly concatenates multi-modal images and provide them as the input of land cover mapping networks. Furthermore, Land cover mapping can be basically described as a classification application, which classifies each pixel of remote sensing images to one of various categories (analogous to the semantic segmentation). Especially in the last decade, semantic segmentation networks, such as UNet [15], DeepLabv3+ [13], SegNet [16] and PSPNet [14], developed rapidly, and have achieved great success for some natural imaging datasets, such as COCO [24] and PASCAL VOC 2012 [25]. Thus, one can develop a machine learning approach built upon classical semantic segmentation networks mentioned above and try to develop some improved architectures for the purpose of land cover mapping application.

When it comes to pixel-level classification, either semantic segmentation in computer vision or land cover classification/mapping in remote sensing, fully convolutional networks (FCNs) [26] have made a considerable contribution, which have made these models and their variants become the state of the art in the literature. SegNet [16] and UNet [15] adopt a symmetrical encoder–decoder structure and skip connections, whilst making use of multi-stage features in the encoder. Alternatively, PSPNet [14] proposes to use a pyramid pooling structure, which provides a global contextual prior to pixel-level scene parsing. Instead of leveraging from the traditional convolution layer used in the aforementioned networks, atrous convolution and atrous spatial pyramid pooling (ASPP) are proposed in Deeplab architecture [27]. This fact helps the DeepLab architecture to exploit the ability to perceive multi-scale spatial information, even using fixed-size convolution kernels. Regardless of the fact that the ASPP can benefit from acquiring information from multi-scale features, DenseASPP [17] argues that the feature resolution in the scale-axis is not dense enough. Thus, DenseASPP combined dense networks [28] and an Atrous convolution network to generate densely scaled receptive fields. DeepLabv3+ [13] proposed an improved hybrid approach that combines an encoder–decoder structure and the ASPP, which can control the resolution of extracted encoder features, trade-off precision and runtime via setting different dilation rates. Specifically, in DeepLabv3+, appending the ASPP module after a backbone of ResNet makes the network exploit deeper levels and extract high-level features with an aim of improving the performance around the segmentation boundaries [29,30]. However, DeepLabv3+ just simply concatenates the two levels of features that come from the output of the first backbone layer and the output of the ASPP module in its decoder. In this case, the network misses the features of the intermediate process of extracting features, which basically reduces the classification performance.

Remote sensing imagery has more complex challenges compared to natural images, such as (1) hardly separable land cover classes; (2) imbalanced class distributions; and (3) imagery content under strong random noise, such as speckle in radar imagery. These might sometimes cause the aforementioned semantic segmentation networks to achieve unsatisfactory results. For the purpose of finding solutions for the challenges mentioned above, the literature includes some semantic segmentation networks for land cover classification applications. Fusion-FCN [31] improves the FCN network and uses it for multi-modal remote sensing for land cover classification, and this network is the winner of 2018 IEEE GRSS Data Fusion Contest. DKDFN [12] is also based on FCN and collaboratively fuses multimodal data and assimilates highly generalisable domain knowledge (e.g., remote sensing indices such as NDVI, NDBI, and NDWI) at the same time. The performance of DKDFN is better than that of some of these classical semantic segmentation networks such as UNet, SegNet, PSPNet, DeepLab. It is worth noting that both Fusion-FCN and DKDFN extract multi modal features with different encoders, and this use of multi-encoders also appears in RGB-D fusion for the semantic segmentation of natural images [32]. DISNet [33] is another network for land cover classification, which uses the DeepLabv3+ framework

and only adds an attention-mechanism-based module in both the encoder and decoder of the network. This change improves the performance of land cover classification compared to the original DeepLabv3+ [13]. Xia [29] also implemented DeepLabv3+ and similarly proposed the global attention based up-sample module in the networks. Xia's network also passes multi-level features to the decoder to obtain efficient segmentation with accurate results. Similarly, Lei [34] proposed a multi-scale fusion network based on a variety of attention mechanisms for land cover classification, which also shows competitive performance. In order to combine the advantages of UNet and DeepLabv3+, ASPP-U-Net [35] was proposed for land cover classification and showed better results compared to the UNet and DeepLabv3+.

3. The Proposed AMM-FuseNet Architecture

In order to make use of multi-modal data for the purposes of land cover classification application, we propose a deep neural network named *AMM-FuseNet*, which explores the effect of channel attention on multi-modal data fusion. Specifically, a couple of channel attention modules are applied in the main structure and we also propose a channel attention based feature extractor called CADenseASPP. In the sequel, we explain the stages of the proposed AMM-FuseNet.

3.1. Channel-Attention Module

Channel attention is a kind of squeeze-and-excitation block, which focuses on finding channels that are meaningful and encourage decoders to use the most relevant features. Generally, the outputs of various channel attention modules are a weighted combination of input features according to the importance of each channel. Thus, the use of channel attention modules is prone to enhance the representative power of networks, and channel attention is able to make features more informative. Specifically, this paper adopts the efficient channel attention (ECA-Net) [36] module, which spatially squeezes the feature map and excites along the channel, as shown in Figure 2. Assume the feature map is

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_C], \tag{1}$$

where $\mathbf{u}_i \in \mathbb{R}^{H \times W}$ refers to ith channel of the feature map.

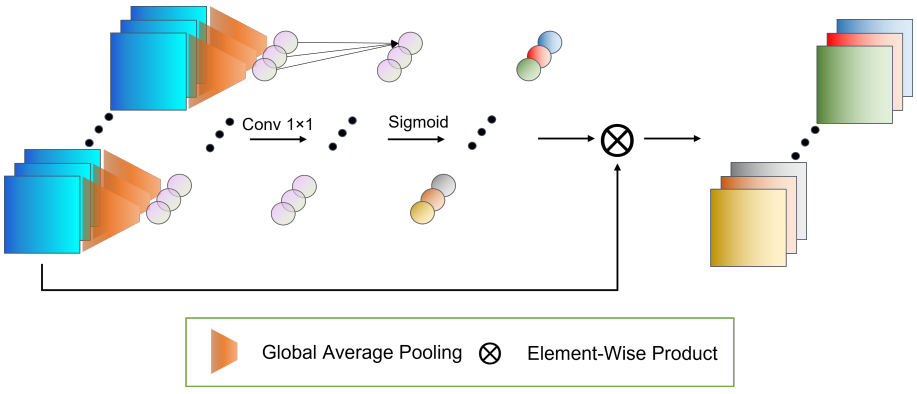

Figure 2. Visual representation of the channel attention module (CA-Module). Aggregated features are processed by average pooling to acquire the element for each channel. Channel weights are obtained after 1-D convolution and applying the sigmoid function. Corresponding channel weights and feature channels are shown in the figure by matching colors.

Spatially squeezing means applying a global average pooling operation to generate a vector $\mathbf{Z} \in \mathbb{R}^{1 \times 1 \times C}$, where the number of corresponding channels is C. In order to consider local cross-channel interaction, the vector will be input to a 1-dimensional convolution layer $\mathbf{W} \in \mathbb{R}^{k \times C}$, where k refers to the size of the convolutional kernel. Thus, the feature values $\mathbf{Z}$ for each channel become

$$\hat{\mathbf{Z}} = \mathbf{WZ}. \tag{2}$$

Here, we use convolution layers rather than fully connected (FC) ones since the convolution operation has much lower complexity compared to the FC, despite achieving similar performance (recommended in [36]). Following this, a sigmoid layer $\sigma(\cdot)$ is used to normalize the transformed vector $\hat{\mathbf{Z}}$ into [0, 1]. Finally, the channel attention will perform excitation of the original feature map along the channel as

$$\hat{\mathbf{U}} = F_c(\mathbf{U}) = [\sigma(\hat{z_1})\mathbf{u}_1, \sigma(\hat{z_2})\mathbf{u}_2, \cdots, \sigma(\hat{z_C})\mathbf{u}_C]. \tag{3}$$

It is worth noting that the value of $\sigma(\hat{z}_i)$ is the attention score of the ith channel, which represents the importance of the channel in the feature map.

3.2. CADenseASPP Module

This paper proposes a channel-attention-based dense atrous spatial pyramid pooling (CADenseASPP) module (named after the DenseASPP [17] module) which is an extension of DenseNet [28]. As is mentioned in Section 3.1, channel attention is able to produce a weighted combination of features and make these derived features more informative. This paper also explores the effect of the channel attention module for features obtained by atrous convolution to promote better representation of features in the proposed model. CADenseASPP combines DenseASPP and channel attention modules, and as shown in Figure 3, channel attention modules are appended just after each Atrous convolution layer to weight the output features by their channel importance scores. Following this, similar with the DenseNet [28] architecture, all features obtained from each atrous convolution branch are concatenated together with features from the additional pooling and identity mapping layers. Due to its dense characteristics, the proposed module also shares the advantages of DenseNet [28], including alleviating the gradient-vanishing problem and having substantially fewer parameters.

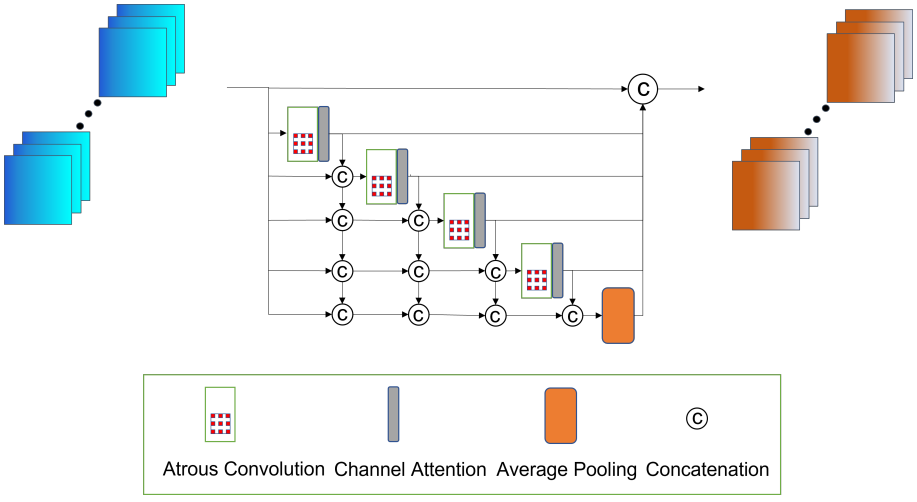

Figure 3. Channel-attention-based densely connected atrous spatial pyramid pooling (CADenseASPP). The atrous convolution layers and an average pooling layer are densely connected, and a channel attention module is added just after each atrous convolution layer.

3.3. AMM-FuseNet

The proposed multi-modal land cover mapping architecture, AMM-FuseNet, exploits the following:

- A dual structure for two modalities;
- Channel attention mechanism;
- CADenseASPP module;
- The use of low–mid–high-level features.

As shown in Figure 4, the proposed network exploits the use of multi-modal imagery via two encoders in a parallel fashion. AMM-FuseNet adopts ResNet-50 as its backbone and implements CADenseASPP to extract information from low level features that are obtained by the first convolution layer of the backbone. Additionally, in order to compensate for the low resolution of high-level features and make use of feature maps from middle or early layers, skip-connections are designed from encoder to decoder to make use of multi level features. Additionally, all branches from the backbone are optimized by corresponding channel attention modules. Then all features are up-sampled into the same spatial size and concatenated together with features obtained by the CADenseASPP module. In the AMM-FuseNet decoder, all features coming from both encoders are concatenated, and finally a segmentation head, consisting of two convolution layers and a ReLu activation function, carries out semantic segmentation.

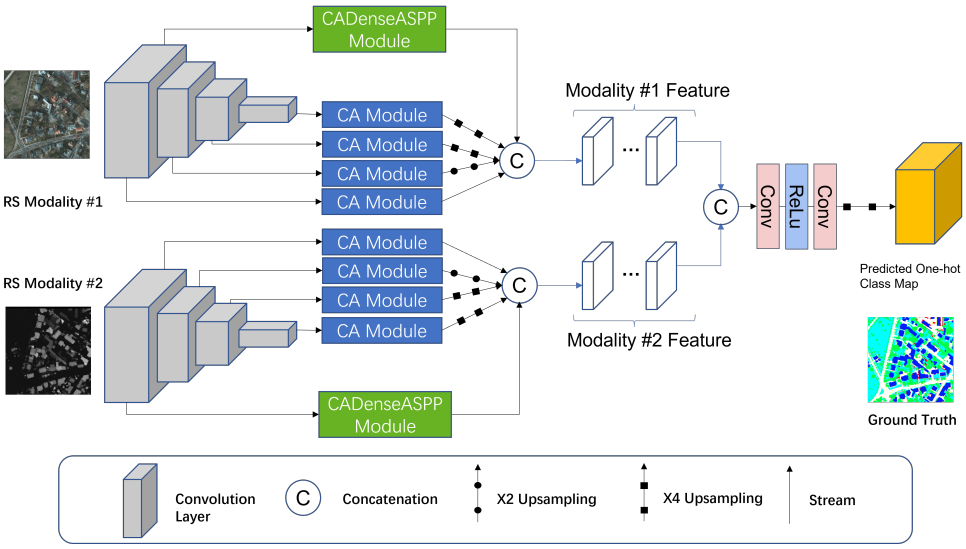

Figure 4. Attention-based multi-modal image fusion network (AMM-FuseNet). Two encoders are applied for two modalities, respectively. For each modality, there are two extractors used in parallel. Skip connections from the encoders to the decoder are used in the network.

The proposed network, AMM-FuseNet, is specifically designed for two-modality data fusion applications. The differences between multi-modal remote sensing data caused by different technologies for obtaining information (e.g., Sentinel-2 data mainly focuses on spectral information whilst DEM data mainly collects elevation of the objects) might cause various problems. To name but a few, (i) low correlation between multi modal images; (ii) registration error especially for large scale land cover data sets. Therefore, when two-modality data have low correlation and strong registration error, sharing parameters of the same encoder will not be optimal and cause severe performance degradation. In addition, using two encoders for multi-modal data fusion has been proved to be more efficient in [37] compared to various deep learning architectures. Thus, the proposed method splits each modality into different encoders to promote their useful features in different network levels, yet fusion operation of the multi-modal features has been performed in the later stages of the architecture.

The proposed AMM-FuseNet promotes using two feature extractors in parallel for each data modality to efficiently make use of low-, middle- and high-level features (grey and green boxes in Figure 4). ResNet-50 (grey) provides the information that has a deeper understanding of the semantic information. On the other hand, the CADenseASPP (green) module is used for acquiring low-level features, which are helpful for the extraction of textural features. Meanwhile, in order to avoid losing information in the process of the forward propagation of ResNet-50, the proposed network also exploits features in each stage of ResNet-50 to increase the semantic knowledge.

Since there are many channels in remote sensing data, typically Sentinel-2 (13 bands), the channel attention mechanism plays a key role in reducing the impact of redundant data and in making use of information coming from the most relevant channels. On the contrary, for natural images, e.g., RGB images only have 3 bands, exploiting all information from all spectral bands is general practice in the literature. However, when the number of channels of input data increases, channel attention becomes highly useful to make efficient use of 'important' channels and reduce the role of 'unimportant' channels in order to obtain accurate classification or segmentation results. In addition, the usage of multiple channel attention module makes the proposed AMM-FuseNet a kind of dynamic neural network. Since the channel attention module rescales the features with input-dependent soft attention, applying the attention mechanism on features is equivalent to performing convolution with dynamic weights, which has been proved to enhance the representation power of deep networks [38]. Thus, we choose to use channel attention in the skip-connection branches and feature extractors.

3.4. Methodological Framework

A methodological overview of the whole approach is given in Figure 5. Each dataset used in this paper provides non-overlapping training and testing data. Both training and testing data include multi-modal remote sensing images and corresponding labels. The whole methodological approach consists of training and testing stages:

- In the training stage, shown in the upper branch of Figure 5, multi-modal images are given as input to the deep learning networks followed by acquiring the land cover prediction. Then, the prediction and corresponding ground truth are used to calculate the loss by using some functions, such as cross entropy, mean square error and/or Dice loss functions. Then, this loss value is used to update the parameters in the networks by using optimizers, such as stochastic gradient descent (SGD) [39], which is illustrated in detail in Section 5.1.
- In the testing stage, shown in the lower branch of Figure 5, multi-modal images are given as input to the trained models, and the land cover mapping predictions are obtained. Different from the training stage, the network will not be updated, and the prediction and ground truth are used for calculating the performance of the utilized network by using related assessment metrics of overall accuracy (OA), user's accuracy (UA), producer's accuracy (PA), mean intersection over union (mIoU), and F1-score, which are also stated in detail in Section 5.1.

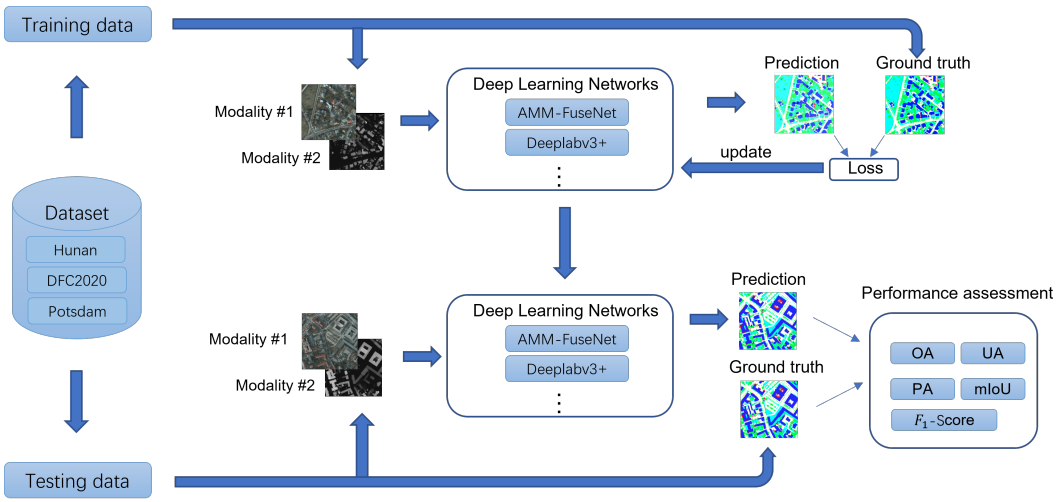

Figure 5. Methodological framework of the land cover mapping application via using deep network architectures. The related assessment metrics in the chart are overall accuracy (OA), user's accuracy (UA), producer's accuracy (PA), mean intersection over union (mIoU), and F1-score.

4. Data

In this paper, we use three open-access multi-modal data sets, which are constructed for land cover classification application, namely (1) Hunan [12], (2) Potsdam [10] and (3) DFC2020 [11] datasets. Although most of the open-access land cover mapping datasets in the literature focus on single modality, all of the three datasets used in this paper are multi-modal remote sensing image datasets for land cover mapping. In particular, the DFC2020 and Potsdam datasets are highly representative and commonly used in the literature, whilst the Hunan dataset is a new dataset published in 2022 [12]. It includes three different remote sensing modalities for land cover mapping, which makes it highly suitable for testing the proposed method's performance. Some details regarding all three utilized datasets are presented in Table 1, where we have the following:

- SRTM refers to Shuttle Radar Topography Mission Digital Elevation Model (DEM) data;
- TOP refers to the true orthophoto;
- DSM refers to a digital surface model.

Table 1. Dataset information.

Dataset	Modality	# of Bands	Spatial Resolution	# of Samples	# of Classes
Hunan [12]	Sentinel-1/2, SRTM	2/13/1	10 m to 30 m	500 (256 × 256)	7
DFC2020 [11]	Sentinel-1/2	2/13	10 m to 20 m	6114 (256 × 256)	10
Potsdam [10]	TOP/DSM	4/1	5 cm	38 (6000 × 6000)	6

4.1. Hunan

Hunan [12] is a multi-modal dataset for land cover mapping of Hunan province in China. This dataset consists of three remote sensing modalities of multi-spectral (Sentinel-2), SAR (Sentinel-1) and SRTM digital elevation model data (DEM). Specifically, the temporal resolutions of Sentinel-2 MSI and Sentinel-1 SAR imagery in Hunan are 5 and 6 days (combined constellation), respectively [40,41], which were captured in 2017. The SRTM (shuttle radar topography mission) is mounted on a space shuttle and obtains Earth surface data by utilizing a synthetic aperture radar. During its 11-day flight, from 11 February 2000 to 22 February 2000, it obtained data covering 80% of the Earth's surface [42]. The obtained data were converted into digital elevation model (DEM) data, which provide height information of the Earth surface. More details for this dataset are listed in Table 2. All 13 bands in Sentinel-2 are used in our experiments. Sentinel-1 data in Hunan dataset were pre-processed by thermal noise removal, radiometric calibration, terrain correction and logarithmic conversion. There are two bands in Sentinel-1 data, corresponding to dual-polarization of VV and VH, respectively. SRTM provides both elevation and slope data, which provide extra topographic information, but only elevation data are used in our experiments. The creators of the Hunan dataset resampled all data to a spatial resolution of 10 m by the default resampling strategy nearest neighbor in GEE [12]. Since the Hunan dataset contains three different modalities of remote sensing images as mentioned above, we have therefore divided them into two fusion pairs of (i) Sentinel-2 and Sentinel-1, and (ii) Sentinel-2 and DEM.

Table 2. Details of Hunan data modalities.

Data Type	Product	Bands	Spatial Resolution
Sentinel-2	Sentinel-2 MSI	B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B10, B11, and B12	10 m to 20 m
Sentinel-1	Sentinel-1 SAR	VV and VH	10 m
SRTM	SRTM Digital Elevation Data Version 4	Elevation and slope	30 m

The Hunan dataset consists of 500 image tiles for each modality, as well as their corresponding land cover labels. The size of all the images is 256 × 256. Geology experts manually labeled the data according to the Sentinel-2 mosaic. This dataset contains 7 imbalanced class labels, which are cropland (23.34%), forest (42.37%), grassland (7.35%), wetland (1.89%), water (13.35%), unused land (1.56%), and built-up area (10.14%). This distribution of land cover classes is based on the data collected in Hunan province, China in 2017.

4.2. DFC2020

DFC2020 is based on the SEN12MS dataset [43], which provides Sentinel-1 SAR imagery, Sentinel-2 multispectral imagery, and corresponding land cover maps on 7 areas (see Table 3) in the world between 2016 and 2017. The temporal resolution and collection time of modalities in DFC2020 is listed in Table 4. The size of all patches is 256 × 256 pixels. The fine-grained IGBP classification scheme in SEN12MS was aggregated to 10 coarser-grained classes, which are forest (11.3%), shrubland (6.9%), savanna (23.6%), grassland (16.8%), wetlands (1.1%), croplands (17.9%), urban/built-up (10.6%), snow/ice (0.0%), barren (5.2%), and water (6.5%). The class distributions are similar to the SEN12MS and DFC2020 datasets. On the other hand, since the DFC2020 dataset is a subset of SEN12MS, there is one class showing zero percentage. The comparison between the standard IGBP classes of SEN12MS and DFC2020 label classes can be found in detail in [11].

Table 3. Study area and data collection time on DFC2020.

Area	Collection Time
Mexico city, Mexico	Winter—1 December 2016 to 28 February 2017
Kippa-Ring, Australia	Winter—1 December 2016 to 28 February 2017
Khabarovsk, Russia	Spring—1 March 2017 to 30 May 2017
Black Forest, Germany	Summer—1 June 2017 to 31 August 2017
Mumbai, India	Fall—1 September 2017 to 30 November 2017
Cape Town, South Africa	Fall—1 September 2017 to 30 November 2017
Bandar Anzali, Iran	Fall—1 September 2017 to 30 November 2017

Table 4. Temporal resolution and collection time of the modalities in DFC2020.

	Temporal Resolution	Year
Sentinel-2 MSI	5 days (combined constellation) [40]	2016–2017
Sentinel-1 SAR	6 days (combined constellation) [41]	2016–2017

4.3. ISPRS Potsdam

The ISPRS Potsdam Semantic Labeling dataset is an open-access benchmark dataset provided by the International Society for Photogrammetry and Remote Sensing (ISPRS). This dataset provides 38 multi-source patches (all of size 6000 × 6000), which contains infrared (IR), red, green and blue orthorectified optical images with corresponding digital surface models (DSM). For calculation purposes, we sub-divided all these data tiles into 512 × 512 patches, which leads to 3456 and 2016 samples for the training and test,

respectively. The ground sampling distance of the two modalities of the true orthophoto (TOP) and the DSM is 5 cm. This dataset was classified manually into six land cover classes, which are impervious surfaces, buildings, low vegetation, trees, cars, clutter/background.

5. Experimental Results and Analysis

5.1. Implementation Details

We implemented our methods by using the Pytorch [44] environment. Following [13], we used a mini-batch SGD optimizer and adopted a poly learning rate policy, where the current learning rate equals the initial learning rate multiplied by $\left(1 - \frac{\text{iter}}{\text{max-iter}}\right)^{\text{power}}$. We set the initial learning rate and power to 0.01 and 0.9, respectively. The batch size of the input data was also set to 10. The objective function for training models is cross-entropy loss function. All the experiments were performed in the GW4 Supercomputer Isambard [45], the details of which are shown in Table 5.

Table 5. Experimental environment configuration.

Item	Detail
CPU	AMD EPYC 7543P
GPU	NVIDIA A100-sxm
Deep Learning Framework	Pytorch 1.10.2
Programming Language	Python 3.7.11

In the comparison analysis, we compared the proposed AMM-FuseNet to six state-of-the-art models of DeepLabv3+, Unet, SegNet, PSPNet, DenseASPP, and DANet. In order to make the reference models suitable for loading multi-band images, the number of input channels in the first convolution layer was set to the number of bands of each dataset. By following the original papers of the reference models, we initiated the backbones of DeepLabv3+, SegNet, PSPNet, DANet with pretrained weights on ImageNet. The AMM-FuseNet has two experimental versions, namely the initial backbone with and without pretrained weights on ImageNet.

We comprehensively analyzed all the models by quantifying performance via class-related measures, such as overall accuracy (OA), user's accuracy (UA), producer's accuracy (PA), mean intersection over union (mIoU), and F_1-score. It is worth noting that the UA represents correct positive predictions relative to the total positive predictions, whilst the PA represents correct positive predictions relative to total actual positives. Expressions of all five performance metrics are given as follows:

$$\text{OA} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{UA} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{PA} = \frac{TP}{TP + FN} \tag{6}$$

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

$$F_1 = \frac{2 \cdot PA \cdot UA}{PA + UA} \tag{8}$$

where TP, TN, FP, and FN refer to the numbers of pixels that are true positives, true negatives, false positives, and false negatives for each class, respectively.

5.2. Quantitative Results and Analysis

The proposed method was tested from two different perspectives:

1. We first used four test data coming from the three data sets discussed in Section 4 in order to promote the generalization capacity of the proposed method.
2. Subsequently, we tested the capabilities of AMM-FuseNet under minimal training supervision in comparison to the reference methods.

For the first set of experiments, we first demonstrated the performance of all methods for each land cover class from the Hunan data set (Sentinel 1/2). This analysis aims to demonstrate the capabilities of each model for different land cover classes and different multi-modal remote sensing data. Figure 6 depicts IoU values of each of the seven land cover classes of the Hunan data set. Examining the bar plots in Figure 6, AMM-FuseNet achieved the highest IoU values for five out of seven categories, with particular considerable improvements for cropland, wetland and bare land. For the remaining two classes (forest and grassland), the IoU results are closer for each model, whilst SegNet, UNet, DeepLabv3+ and AMM-FuseNet compete to become the best performing method.

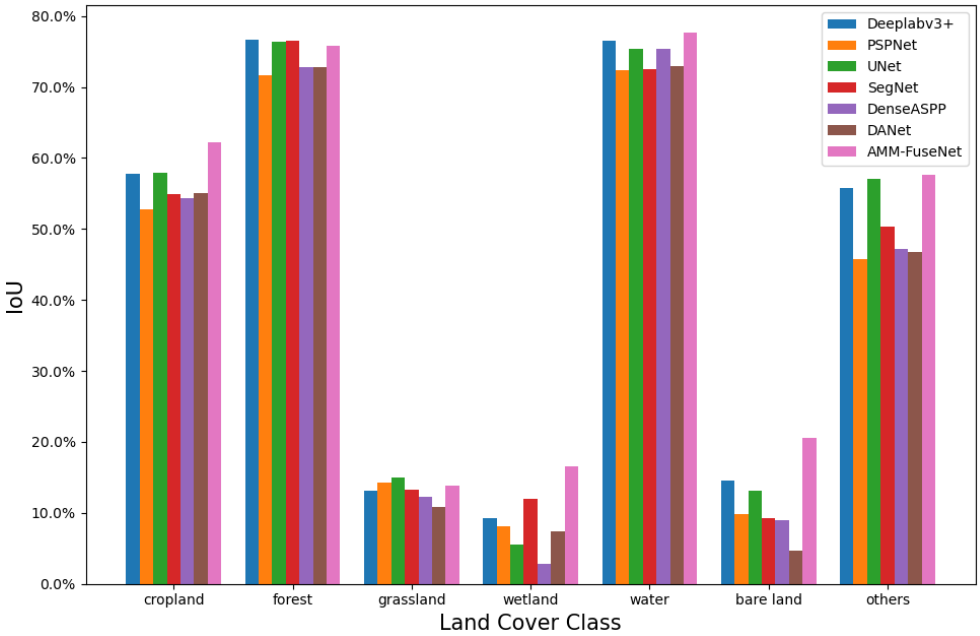

Figure 6. IoU of each land cover class on the Hunan dataset (Sentinel-2 and Sentinel-1).

In order to test the overall performance of each model, as mentioned above, we created four pairs of two-modality remote sensing data from three data sets. In order to reflect a better comprehensive performance of the models, we used *mIoU*, *average OA*, *UA*, *PA* and F_1 of all land cover classes for each model. We present these values in Tables 6–9. Please note that in our initial analysis, pretraining on ImageNet does not yield a consistent result due to the different characteristics of each dataset. Hence, we decided to test each model with both a pretrained backbone and a backbone that was not pretrained, and then only share the best performing result for each with an indicator in the "PT" columns in each table. Furthermore, in order to provide visual evidence for the performance metrics presented in Tables 6–9, we plotted land cover mapping outputs of each model for a randomly selected image from each of the two-modality data pairs in Figure 7–10.

Examining the performance metrics in Table 6, the proposed AMM-FuseNet showed the best performance on the first two-modality Hunan (Sentinel 1/2) data set for most of the cases. In particular, AMM-FuseNet obtained 4% UA gain compared to the second best model. Furthermore, the mIoU value of AMM-FuseNet is about 3 % higher than that of the second best model whilst PSPNet is slightly better than AMM-FuseNet in terms of the PA value.

Table 6. Performance comparison for Hunan (Sentinel-2 and Sentinel-1) dataset. (✓: pre-trained (PT) and ×: not PT backbone).

	PT	mIoU	OA	UA	PA	F_1
DeepLabv3+ [13]	✓	43.38%	78.34%	52.74%	62.64%	57.27%
PSPNet [14]	✓	39.26%	74.81%	48.61%	**62.75%**	54.78%
UNet [15]	×	42.91%	77.77%	52.97%	56.86%	54.84%
SegNet [16]	✓	41.24%	76.64%	50.67%	59.92%	54.91%
DenseASPP [17]	×	39.94%	75.79%	49.45%	58.27%	53.50%
DANet [18]	✓	38.65%	75.61%	47.64%	57.38%	52.06%
AMM-FuseNet	×	**46.31%**	**79.06%**	**57.35%**	61.04%	**59.13%**

The second two-modality data set consists of Sentinel-2 and DEM modalities, which also come from the Hunan data set. AMM-FuseNet shows the best performance in terms of all performance metrics presented in Table 7. The proposed network has a considerable mIoU improvement of more than 4% when compared with the second best model, UNet, whilst around 5% improvement is obtained in terms of F_1 score compared to the DeepLabV3+.

Table 7. Performance comparison for Hunan (Sentinel-2 and DEM) dataset. (✓: pre-trained (PT) and ×: not PT backbone).

	PT	mIoU	OA	UA	PA	F_1
DeepLabv3+ [13]	✓	40.83%	77.15%	50.02%	59.15%	54.20%
PSPNet [14]	✓	37.35%	74.15%	46.83%	56.96%	51.40%
UNet [15]	×	41.42%	76.92%	51.12%	56.64%	53.74%
SegNet [16]	✓	39.05%	75.56%	48.47%	55.04%	51.54%
DenseASPP [17]	×	38.72%	75.06%	48.29%	56.22%	51.95%
DANet [18]	✓	37.28%	73.99%	46.46%	59.07%	52.01%
AMM-FuseNet	×	**45.70%**	**78.64%**	**56.31%**	**61.09%**	**58.61%**

From the two experimental sets of Hunan dataset presented in Tables 6 and 7, the state-of-the-art models show instabilities when dealing with different combinations of multi-modal imagery. For example, apart from AMM-FuseNet, DeepLabv3+ achieved the second best performance for most of the metrics on Sentinel 1/2 fusion whereas, in terms of Sentinel-2 and DEM fusion, UNet appeared to beat DeeplabV3+ and became the best. Despite these inconsistencies of the state-of-the-art models, the proposed AMM-FuseNet has shown consistent performance with better generalisability. It showed competitive performance either for Sentinel 1/2 or Sentinel-2/DEM fusion, even there are no pretrained weights to initialize the network.

Similar to the first Hunan dataset, the DFC2020 dataset also provides Sentinel 1/2 imagery. Examining the performance results presented in Table 8, AMM-FuseNet showed a competitive performance on the DFC2020 dataset in terms of all performance metrics. AMM-FuseNet improved slightly for all metrics compared to the second best network, Unet. Furthermore, the performance difference between the AMM-FuseNet and the remaining networks is relatively high. For example, there are more than 6% and 2% improvements in terms of mIoU and OA, respectively, between the proposed network and PSPNet.

Table 8. Performance comparison for DFC2020 (Sentinel-2 and Sentinel-1) dataset. (✓: pre-trained (PT) and ×: not PT backbone).

	PT	mIoU	OA	UA	PA	F_1
DeepLabv3+ [13]	✓	78.67%	92.94%	88.97%	85.88%	87.39%
PSPNet [14]	✓	76.99%	92.29%	87.51%	84.92%	86.19%
UNet [15]	×	82.47%	94.31%	90.20%	89.40%	89.79%
SegNet [16]	✓	80.28%	93.53%	88.47%	88.15%	88.31%
DenseASPP [17]	×	76.20%	92.54%	85.09%	85.51%	85.30%
DANet [18]	✓	77.50%	92.50%	87.36%	85.62%	86.48%
AMM-FuseNet	✓	**83.14%**	**94.56%**	**91.15%**	**89.53%**	**90.33%**

As of the last analysis of the first set of experiments, we tested all models on the Potsdam dataset which consists of IRRGB and DEM imagery. Unlike previous datasets, the proposed AMM-FuseNet has not shown the best performance when compared to the other networks. AMM-FuseNet is about 1% lower than the best model as shown in Table 9. Our first thought on this result is based on the dramatic, significant differences between the Potsdam dataset and the three previously used test cases in terms of frequency bands, semantic details and spatial resolution. We are going to discuss this in the next section in more detail.

Table 9. Performance comparison for Potsdam dataset. (✓: pre-trained (PT) and ×: not PT backbone).

	PT	mIoU	OA	UA	PA	F_1
DeepLabv3+ [13]	✓	67.12%	84.60%	77.23%	79.84%	78.52%
PSPNet [14]	✓	69.10%	85.85%	78.79%	82.09%	80.41%
UNet [15]	×	67.01%	84.40%	77.06%	79.81%	78.41%
SegNet [16]	✓	67.79%	84.97%	77.78%	79.79%	78.77%
DenseASPP [17]	×	67.03%	84.48%	77.33%	79.98%	78.63%
DANet [18]	✓	**69.77%**	**86.12%**	**79.56%**	**82.15%**	**80.83%**
AMM-FuseNet	✓	68.40%	85.28%	78.29%	80.36%	79.31%

In order to demonstrate results visually and provide visual evidence for the performance metrics presented in the tables above, we chose to show four cases from each two-modality data sets. As shown in Figures 7–10, we depict the RGB imagery, ground truth land cover labels and results predicted by all the models. On each example of land cover mapping results, we draw a rectangular box to help readers to compare the differences between different models.

For the Hunan dataset, the prediction of AMM-FuseNet is visually closer to the ground truth compared with other predictions. Additionally, for DFC2020, the prediction of AMM-FuseNet is more accurate when compared to other models. Similarly, for Potsdam, AMM-FuseNet shows the ability to understand the data more deeply. Only AMM-FuseNet and UNet correctly recognize the single tree structure (green area in the rectangular box) in the Potsdam case, where all other networks incorrectly detect more tree related regions. Additionally, the UNet prediction shows a redundant and incorrect classification in the upper right part of the rectangular box (red pixels).

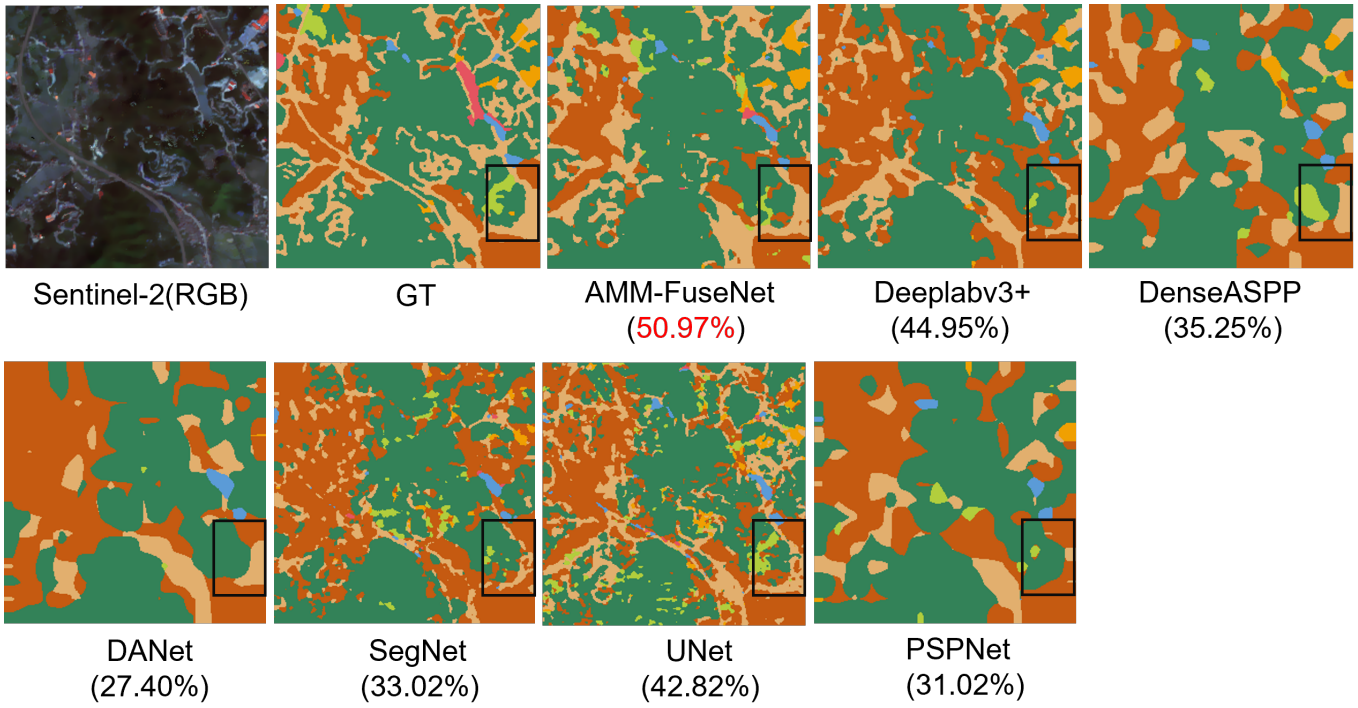

Figure 7. Results on the Hunan dataset (Sentinel-2 and Sentinel-1). Values between parentheses refer to mIoU in percentages for the example image for each model, where the best model is shown in red.

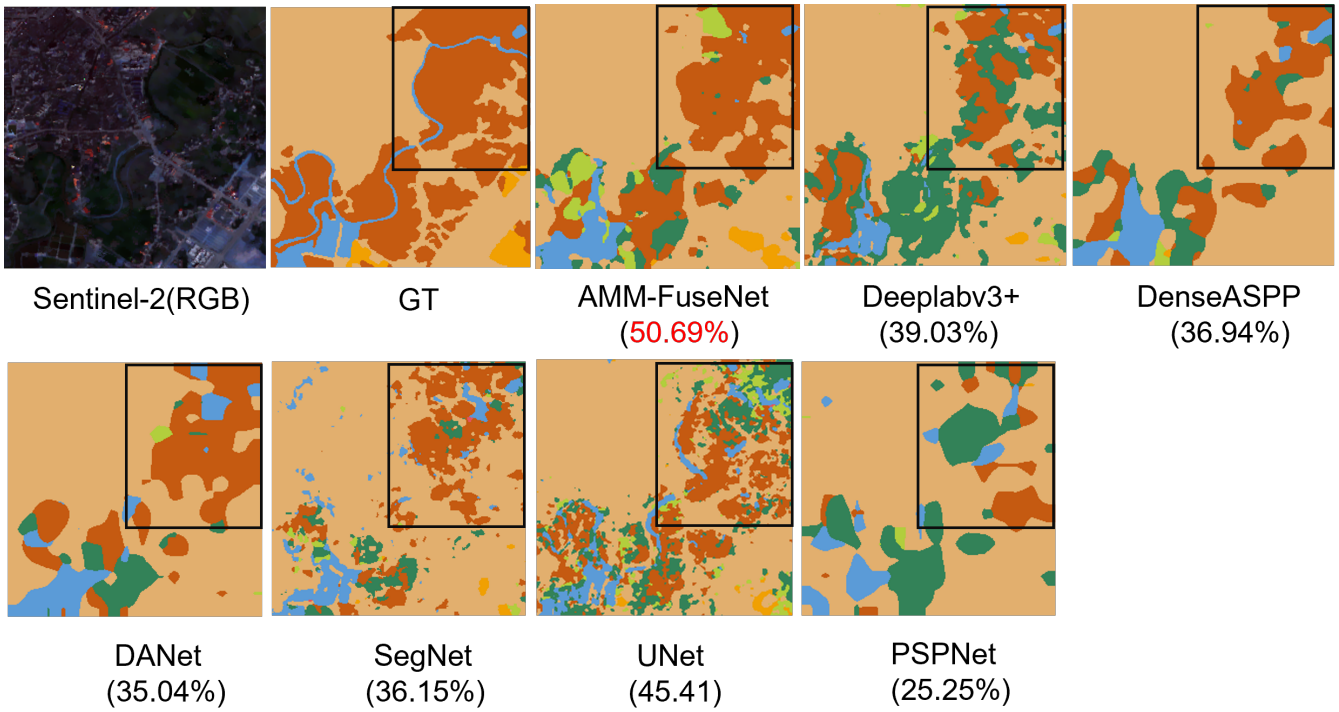

Figure 8. Results on the Hunan dataset (Sentinel-2 and DEM) values between parentheses refer to mIoU in percentages for the example image for each model, where the best model is shown in red.

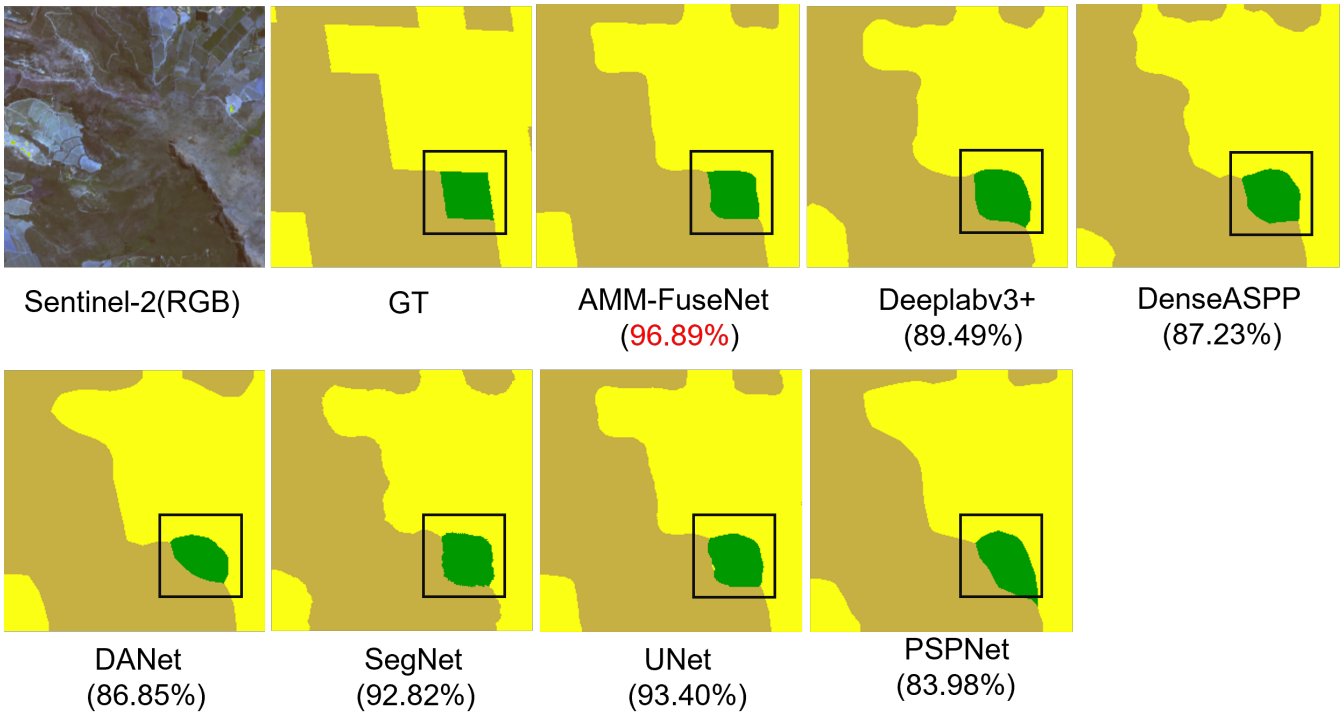

Figure 9. Results on the DFC2020 dataset. Values between parentheses refer to mIoU in percentages for the example image for each model, where the best model is shown in red.

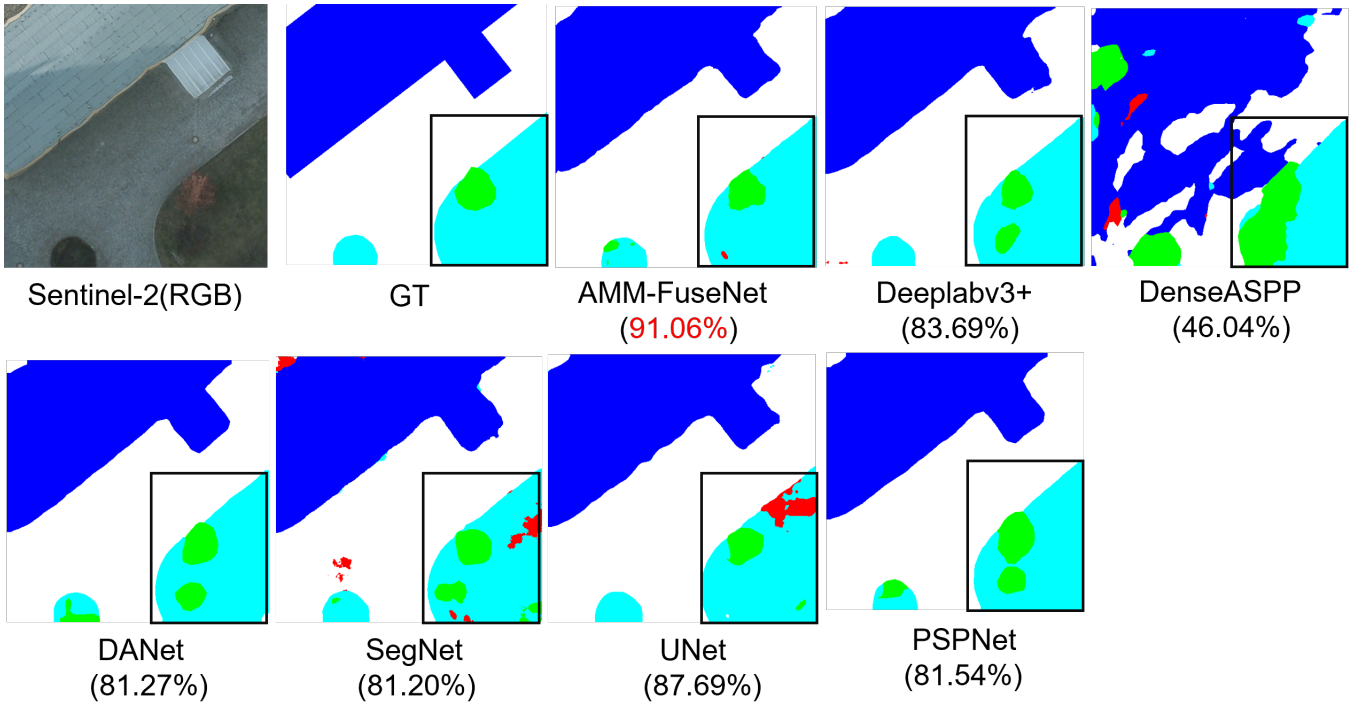

Figure 10. Results on the Potsdam dataset. Values between parentheses refer to mIoU in percentages for the example image for each model, where the best model is shown in red.

5.3. Minimal Supervision Analysis

As we mentioned in the previous section, all the models showed similar performance for the Potsdam dataset, where the proposed AMM-FuseNet could not manage to obtain the best performance outcome in terms of all five metrics. We consider three categories of potential reasons for these results.

1. Due to the lower number of spectral bands in the data set, AMM-FuseNet's ability to extract useful information was affected and became closer to the state of the art.
2. Since all imagery is collected in an urban area with very high spatial resolution (5 cm), discrimination between the objects becomes much easier compared to the other datasets, which leads to all methods performing at a similar level.
3. The Potsdam dataset has a relatively high number of training samples compared to the other datasets (around four times compared to DFC2020 considering the input image size of 512 × 512). This has given the state-of-the-art networks enough data samples to fully learn.

Considering remote sensing imagery applications mostly having highly limited number of labeled samples (as in the case of Hunan dataset), the Potsdam experimental analysis cannot directly lead to correct conclusions for the purposes of remote sensing imagery applications.

Our purpose whilst developing AMM-FuseNet (along with having good performance in land cover mapping applications) is to make it suitable for minimal training supervision cases. AMM-FuseNet separates each modality into different encoders via utilising the same ground truth information, for the purpose of inheriting the advantages of consistency regularization [46], which shows considerable success in minimal supervision cases in the literature [47,48]. The key idea of consistency regularization is to force perturbed models (or perturbed inputs) to have consistent outputs to supervise each other, even when there is a very small number of labeled data. In this case, one modality can be regarded as a perturbed version of another modality and they are expected to have the same output. Thus, the weights in two encoders are supervised by each other during the training.

In order to experimentally prove this point and provide numerical evidence, we gradually reduced the number of training samples of the Potsdam dataset. Although the number of training samples is reduced, the class distributions in each minimal supervision test case is kept relatively unchanged. In addition, we used the same fixed set of test data samples for all the minimal supervision test cases so as to enable a consistent comparison of the results. Details of each test case in this set of experiments are shown in Table 10.

Table 10. Minimal training supervision test cases and class distributions.

Fraction	Training Samples	Test Samples	Class Distributions for Training					
			Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
1	3456	2016	9.12%	27.11%	25.36%	22.54%	14.31%	1.56%
1/2	1728	2016	8.89%	27.41%	25.47%	22.41%	14.28%	1.55%
1/4	864	2016	9.41%	26.92%	25.07%	23.00%	13.93%	1.68%
1/6	576	2016	9.43%	27.09%	25.21%	21.91%	14.75%	1.61%
1/8	432	2016	8.80%	27.47%	25.45%	22.49%	14.26%	1.53%
1/10	346	2016	9.40%	26.47%	26.19%	21.81%	14.58%	1.55%
1/12	288	2016	8.75%	27.91%	24.97%	22.65%	14.16%	1.55%
1/15	230	2016	9.26%	27.37%	24.61%	22.94%	14.08%	1.73%
1/20	173	2016	8.92%	27.50%	25.30%	22.80%	13.88%	1.60%

The results of each minimal training supervision test cases for $\frac{1}{2} - \frac{1}{20}$ are shown in Tables 11 and 12. Whilst gradually reducing the number of training samples, AMM-FuseNet starts to show its capability to work with smaller number of training samples. In contrast, the state-of-the-art approaches are affected by the small number of training samples and their results deteriorated dramatically. In particular, from ratio $\frac{1}{6}$, namely 576 training samples, AMM-FuseNet starts to be the best model in terms of mIoU and overall accuracy on the Potsdam dataset by having lesser performance loss. Even for only 173 training samples for the $\frac{1}{20}$ case, the proposed AMM-FuseNet obtained more than 60% mIoU and 80% overall accuracy values.

Table 11. Performance comparison of different methods on the Potsdam dataset for reduced training samples in terms of mIoU and accuracy (✓: Initial backbone with pre-trained (PT) weights on ImageNet. ×: Initial backbone Randomly).

		1/2		1/4		1/6		1/8	
	PT	**mIoU**	**Accuracy**	**mIoU**	**Accuracy**	**mIoU**	**Accuracy**	**mIoU**	**Accuracy**
DeepLabv3+ [13]	✓	64.98%	83.46%	62.24%	82.02%	58.85%	79.40%	58.08%	79.57%
PSPNet [14]	✓	67.45%	84.95%	65.01%	83.53%	62.22%	82.21%	62.20%	82.17%
UNet [15]	×	64.69%	83.02%	61.98%	81.19%	58.97%	79.25%	57.96%	78.69%
SegNet [16]	✓	65.91%	83.93%	65.14%	83.45%	62.68%	82.13%	61.16%	81.33%
DenseASPP [17]	×	65.15%	83.66%	62.48%	82.11%	60.02%	80.68%	59.18%	80.29%
DANet [18]	✓	**68.45%**	**85.51%**	**66.03%**	**84.01%**	62.62%	82.52%	62.39%	81.97%
AMM-FuseNet	✓	66.49%	84.28%	65.06%	83.49%	**63.53%**	**82.55%**	**63.48%**	**82.63%**

Table 12. Performance comparison of different methods on the Potsdam data for reduced number of training samples in terms of mIoU and accuracy (✓: Initial backbone with pre-trained (PT) weights on ImageNet. ×: Initial backbone Randomly).

		1/10		1/12		1/15		1/20	
	PT	**mIoU**	**Accuracy**	**mIoU**	**Accuracy**	**mIoU**	**Accuracy**	**mIoU**	**Accuracy**
DeepLabv3+ [13]	✓	57.80%	78.54%	56.02%	77.84%	54.56%	76.81%	54.06%	76.86%
PSPNet [14]	✓	61.41%	81.13%	60.28%	80.84%	58.59%	79.99%	58.28%	79.71%
UNet [15]	×	58.75%	79.17%	57.49%	78.26%	55.84%	76.94%	56.10%	78.02%
SegNet [16]	✓	60.15%	80.72%	58.67%	80.00%	58.27%	79.26%	57.51%	79.20%
DenseASPP [17]	×	58.54%	79.67%	57.72%	79.29%	56.15%	78.01%	55.71%	77.92%
DANet [18]	✓	61.41%	81.65%	60.52%	80.93%	58.99%	79.93%	57.81%	79.17%
AMM-FuseNet	✓	**62.06%**	**81.73%**	**61.60%**	**81.30%**	**61.38%**	**81.06%**	**60.05%**	**80.62%**

In order to visually draw a conclusion on the values in the tables, we depict two graphs (Figures 11 and 12) of percentage decreases on mIoU and accuracy for different partitions of the Potsdam training samples for each network. It is clear to see that performance loss of AMM-FuseNet is 8% in terms of mIoU and 4% of accuracy even though the training set is 20 times smaller than the original dataset. This provides numerical/experimental evidence to our remark that the proposed AMM-FuseNet architecture is better able to cope with cases where the data set has few training samples, and its performance improvement can be seen better under test cases with a small number of training samples. It is also evident that single-encoder architectures show drastic performance drops when the number of training samples is reduced. It is also important to note that methods such as DeepLabV3+ and Unet, which are two of the best approaches on the Hunan and DFC2020 datasets, have the highest performance drops in terms of both mIoU and accuracy, showing they are highly sensitive to the number of training samples.

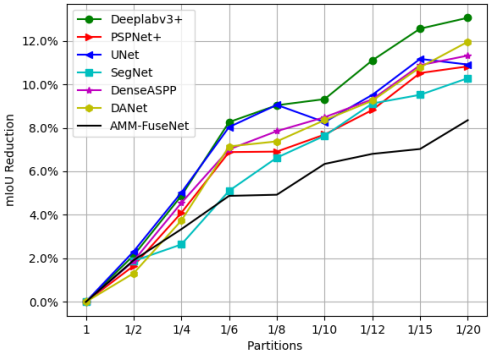

Figure 11. Percentage decrease on mIoU for different partitions of the Potsdam dataset for each network.

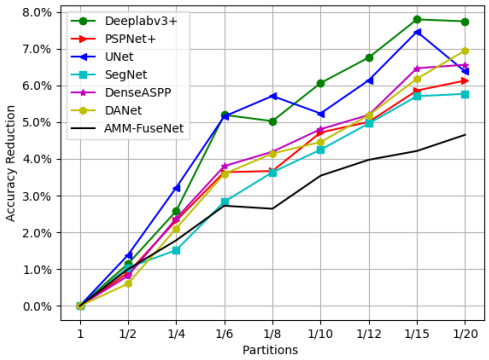

Figure 12. Percentage decrease on accuracy for different partitions of the Potsdam dataset for each network.

6. Conclusions

In this work, we proposed a channel-attention based multi modal image fusion network, *AMM-FuseNet*, constructed with a proposed novel feature extraction module, CA-DenseASPP. The proposed network showed competitive performance when compared to the state-of-the-art segmentation networks, such as DeepLabv3+, SegNet and Unet, and appeared to be more robust and generalisable when applied to various multi-modal remote sensing data sets. For most of the cases with different remote sensing modalities of RGB, multi-spectral, SAR and DEM, the AMM-FuseNet showed a consistent performance by being the best model in terms of various performance metrics.

In presenting the proposed approach AMM-FuseNet in this paper,

- We contributed to the literature with a multi-modal attention-based deep network architecture with improved land cover mapping/classification performance compared to the state of the art.
- The parallel feed of multi-modal remote sensing information into the hybrid proposed encoder module of CADenseASPP improved the segmentation performance dramatically via weighting features in a dense atrous convolution operation.
- It was experimentally proven by using the Potsdam data that the proposed network showed more powerful performance under small number of training sample (minimal training supervision) despite its relatively complex structure due to having two parallel encoders. This issue proved our contribution to the literature that the proposed approach could be a great choice for the segmentation applications that only have a small amount of labeled information.
- As it stands, AMM-FuseNet appears as a candidate model to be a high-performing approach for other segmentation tasks in remote sensing and computer vision beyond the land cover mapping application.

The indistinguishable performance of all models under the full-Potsdam dataset has shown us future research directions in terms of the AMM-FuseNet architecture. Ongoing work includes performing a detailed complexity analysis, and exploring AMM-FuseNet's capabilities (i) to deal with minimal supervision, as well as (ii) extracting useful information from higher spatial resolution imagery.

Author Contributions: Conceptualization, W.M. and O.K.; methodology, W.M., O.K. and P.L.R.; software, W.M.; validation, W.M, O.K. and P.L.R.; writing—original draft preparation, W.M. and O.K.; writing—review and editing, O.K. and P.L.R.; supervision, O.K. and P.L.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The AMM-FuseNet Python code can be downloaded at: AMM-FuseNet Repo accessed on 4 September 2022.

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

References

1. Chae, Y.; An, Y.J. Current research trends on plastic pollution and ecological impacts on the soil ecosystem: A review. *Environ. Pollut.* **2018**, *240*, 387–395. [CrossRef]
2. Azarang, A.; Ghassemian, H. A New Pansharpening Method Using Multi Resolution Analysis Framework and Deep Neural Networks. In Proceedings of the 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), Shahrekord, Iran, 19–20 April 2017; pp. 1–6.
3. Lai, Z.; Chen, L.; Jeon, G.; Liu, Z.; Zhong, R.; Yang, X. Real-time and effective pan-sharpening for remote sensing using multi-scale fusion network. *J. Real-Time Image Proc.* **2021**, *18*, 1635–1651. [CrossRef]
4. Zhang, H.; Shen, H.; Yuan, Q.; Guan, X. Multispectral and SAR Image Fusion Based on Laplacian Pyramid and Sparse Representation. *Remote Sens.* **2022**, *14*, 870. [CrossRef]
5. Karakuş, O.; Kuruoğlu, E.E.; Altınkaya, M.A. Generalized Bayesian model selection for speckle on remote sensing images. *IEEE Trans. Image Proc.* **2018**, *28*, 1748–1758. [CrossRef]
6. Anderson, J.R. *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*; US Government Printing Office: Washington, DC, USA, 1976, Volume 964.
7. Phiri, D.; Morgenroth, J. Developments in Landsat land cover classification methods: A review. *Remote Sens.* **2017**, *9*, 967. [CrossRef]
8. Steiner, D. Automation in photo interpretation. *Geoforum* **1970**, *1*, 75–88. [CrossRef]
9. Zhang, X.; Han, L.; Han, L.; Zhu, L. How well do deep learning-based methods for land cover classification and object detection perform on high resolution remote sensing imagery? *Remote Sens.* **2020**, *12*, 417. [CrossRef]
10. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The ISPRS Benchmark on Urban Object Classification and 3D Building Reconstruction. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1, Melbourne, Australia, 25 August–1 September 2012; pp. 293–298.
11. Robinson, C.; Malkin, K.; Jojic, N.; Chen, H.; Qin, R.; Xiao, C.; Schmitt, M.; Ghamisi, P.; Hänsch, R.; Yokoya, N. Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest. *IEEE J. Sel. Top. App. Earth Obs. Remote Sens.* **2021**, *14*, 3185–3199. [CrossRef]
12. Li, Y.; Zhou, Y.; Zhang, Y.; Zhong, L.; Wang, J.; Chen, J. DKDFN: Domain Knowledge-Guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. *ISPRS J. Photogram. Remote Sens.* **2022**, *186*, 170–189. [CrossRef]
13. 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.
14. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
15. 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.
16. 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] [PubMed]
17. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. DenseASPP for Semantic Segmentation in Street Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
18. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
19. Wambugu, N.; Chen, Y.; Xiao, Z.; Wei, M.; Bello, S.A.; Junior, J.M.; Li, J. A hybrid deep convolutional neural network for accurate land cover classification. *Int. J. Appl. Earth Obs. Geoinform.* **2021**, *103*, 102515. [CrossRef]
20. Zhang, T.; Su, J.; Xu, Z.; Luo, Y.; Li, J. Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier. *Appl. Sci.* **2021**, *11*, 543. [CrossRef]
21. Rousset, G.; Despinoy, M.; Schindler, K.; Mangeas, M. Assessment of deep learning techniques for land use land cover classification in southern new Caledonia. *Remote Sens.* **2021**, *13*, 2257. [CrossRef]
22. Hong, D.; Hu, J.; Yao, J.; Chanussot, J.; Zhu, X.X. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS J. Photogram. Remote Sens.* **2021**, *178*, 68–80. [CrossRef]
23. Solórzano, J.V.; Mas, J.F.; Gao, Y.; Gallardo-Cruz, J.A. Land use land cover classification with U-net: Advantages of combining sentinel-1 and sentinel-2 imagery. *Remote Sens.* **2021**, *13*, 3600. [CrossRef]
24. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
25. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. Available online: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html (accessed on 20 June 2022).
26. 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.

27. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
28. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
29. Xia, M.; Wang, T.; Zhang, Y.; Liu, J.; Xu, Y. Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery. *Int. J. Remote Sens.* **2021**, *42*, 2022–2045. [CrossRef]
30. Zhang, C.; Li, G.; Du, S. Multi-scale dense networks for hyperspectral remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9201–9222. [CrossRef]
31. Xu, Y.; Du, B.; Zhang, L.; Cerra, D.; Pato, M.; Carmona, E.; Prasad, S.; Yokoya, N.; Hänsch, R.; Le Saux, B. Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 1709–1724. [CrossRef]
32. Zhou, F.; Lai, Y.K.; Rosin, P.L.; Zhang, F.; Hu, Y. Scale-aware network with modality-awareness for RGB-D indoor semantic segmentation. *Neurocomputing* **2022**, *492*, 464–473. [CrossRef]
33. Zhang, X.; Wang, Z.; Cao, L.; Wang, M. A Remote Sensing Land Cover Classification Algorithm Based on Attention Mechanism. *Can. J. Remote Sens.* **2021**, *47*, 835–845. [CrossRef]
34. Lei, T.; Li, L.; Lv, Z.; Zhu, M.; Du, X.; Nandi, A.K. Multi-modality and multi-scale attention fusion network for land cover classification from VHR remote sensing images. *Remote Sens.* **2021**, *13*, 3771. [CrossRef]
35. Zhang, W.; Tang, P.; Zhao, L. Fast and accurate land-cover classification on medium-resolution remote-sensing images using segmentation models. *Int. J. Remote Sens.* **2021**, *42*, 3277–3301. [CrossRef]
36. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
37. Farahnakian, F.; Heikkonen, J. Deep learning based multi-modal fusion architectures for maritime vessel detection. *Remote Sens.* **2020**, *12*, 2509. [CrossRef]
38. Han, Y.; Huang, G.; Song, S.; Yang, L.; Wang, H.; Wang, Y. Dynamic neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, arXiv:2102.04906.
39. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
40. Mandanici, E.; Bitelli, G. Preliminary comparison of sentinel-2 and landsat 8 imagery for a combined use. *Remote Sens.* **2016**, *8*, 1014. [CrossRef]
41. Amitrano, D.; Martino, G.D.; Iodice, A.; Mitidieri, F.; Papa, M.N.; Riccio, D.; Ruello, G. Sentinel-1 for monitoring reservoirs: A performance analysis. *Remote Sens.* **2014**, *6*, 10676–10693. [CrossRef]
42. Van Zyl, J.J. The Shuttle Radar Topography Mission (SRTM): A breakthrough in remote sensing of topography. *Acta Astronaut.* **2001**, *48*, 559–565. [CrossRef]
43. Schmitt, M.; Hughes, L.H.; Qiu, C.; Zhu, X.X. SEN12MS–A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion. *arXiv* **2019**, arXiv:1906.07789.
44. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
45. GW4 Isambard. 2014. Available online: https://gw4.ac.uk/ (accessed on 1 July 2022).
46. French, G.; Laine, S.; Aila, T.; Mackiewicz, M.; Finlayson, G. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv* **2019**, arXiv:1906.01916.
47. Chen, X.; Yuan, Y.; Zeng, G.; Wang, J. Semi-supervised semantic segmentation with cross pseudo supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2613–2622.
48. Filipiak, D.; Tempczyk, P.; Cygan, M. *n*-CPS: Generalising Cross Pseudo Supervision to *n* networks for Semi-Supervised Semantic Segmentation. *arXiv* **2021**, arXiv:2112.07528.