remote sensing

MDPI

Article

A Multi-Sensor Fusion Framework Based on Coupled Residual Convolutional Neural Networks

Hao Li [1,*], Pedram Ghamisi [2], Behnood Rasti [2], Zhaoyan Wu [1,3], Aurelie Shapiro [4], Michael Schultz [1] and Alexander Zipf [1]

[1] GIScience Chair, Institute of Geography, Heidelberg University, 69120 Heidelberg, Germany; zhaoyan.wu@uni-heidelberg.de (Z.W.); michael.schultz@uni-heidelberg.de (M.S.); zipf@uni-heidelberg.de (A.Z.)

[2] Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Exploration, Chemnitzer Str. 40, D-09599 Freiberg, Germany; p.ghamisi@gmail.com (P.G.); b.rasti@hzdr.de (B.R.)

[3] The School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

[4] Here+There Mapping Solutions, 10115 Berlin, Germany; aurelie@here-there-mapping.com

[*] Correspondence: hao.li@uni-heidelberg.de; Tel.: +49-6221-54-5534

Received: 22 May 2020; Accepted: 23 June 2020; Published: 26 June 2020

Abstract: Multi-sensor remote sensing image classification has been considerably improved by deep learning feature extraction and classification networks. In this paper, we propose a novel multi-sensor fusion framework for the fusion of diverse remote sensing data sources. The novelty of this paper is grounded in three important design innovations: 1- a unique adaptation of the coupled residual networks to address multi-sensor data classification; 2- a smart auxiliary training via adjusting the loss function to address classifications with limited samples; and 3- a unique design of the residual blocks to reduce the computational complexity while preserving the discriminative characteristics of multi-sensor features. The proposed classification framework is evaluated using three different remote sensing datasets: the urban Houston university datasets (including Houston 2013 and the training portion of Houston 2018) and the rural Trento dataset. The proposed framework achieves high overall accuracies of 93.57%, 81.20%, and 98.81% on Houston 2013, the training portion of Houston 2018, and Trento datasets, respectively. Additionally, the experimental results demonstrate considerable improvements in classification accuracies compared with the existing state-of-the-art methods.

Keywords: deep learning; data fusion; hyperspectral image classification; residual learning; multi-sensor fusion; convolutional neural networks (CNNs); auxiliary loss function

1. Introduction

Multi-sensor image analysis of remotely sensed data has become a growing area of research in recent years. Space and airborne remote sensing data streams are providing increasingly abundant data suited for earth observation and environmental monitoring [1]. The spatial, temporal and spectral capabilities of optical remote sensing systems are also increasing over time. Besides the evolution of multispectral imaging (MSI), hyperspectral imaging (HSI) [2–4] and light detection and ranging (LiDAR) observation platforms have also gained relevance [5–7]. An increasing diversity of platforms of HSI and LiDAR acquisition systems are available for terrestrial, space and airborne-based data collection. While MSI and HSI rely on solar radiance as a passive illumination source, LiDAR devices emit their own source of active radiance for measurement. MSI and HSI systems produce pixels representing two-dimensional bands of their respective wavelengths while LiDAR systems measure structure via point clouds organized in a three-dimensional sphere for their respective wavelengths. Combining such data at image or feature level yields both opportunities and challenges. For instance, fusion of HSI and LiDAR data of the same event in space offers a rich feature space allowing distinct separation of observed objects based on their spectral signature and elevation characteristics [8,9]. Meanwhile, multi-sensor datasets can contain sophisticated heterogeneous data structures and different data formats or characteristics (e.g., asymptotic properties, spatial and spectral resolutions etc.). Given the increasing availability and complexity of multi-sensor data, fusion techniques are evolving to address meaningful data exploitation to cope with multi-source inputs. This paper is addressing the large potential volume of existing combined multi-sensor data on classification algorithms. Depending on the study site and classification scheme, multi-sensor feature spaces can possess unique hybrid properties introducing new challenges for the production and deployment of appropriate training data. Sources of accurate training data are often scarce, and the production is expensive, particularly for novel hybrid multi-sensor feature spaces. Therefore, conventional classification systems and networks often become less efficient for such diverse and complicated datasets. Hence, the effective fusion of heterogeneous multi-sensor data for classification applications is essential to our remote sensing research.

A wide variety of multi-sensor data fusion methods have been developed to leverage the use of heterogeneous data sources, most prominently for HSI and LiDAR data fusion [10–17]. In [10], morphological-level features, specifically attribute profiles (APs), were embedded with a subspace multinominal logistic regression model for the fusion of HSI and LiDAR data. The capability of APs in extracting discriminating spatial features was again confirmed in [11], where extended attribute profiles (EAPs) were used to extract features from HSI and LiDAR data, respectively. Moreover, morphological extinction profiles (EPs) have been proposed to overcome the threshold determination difficulties of APs and further boost the performance of feature extraction [12]. EPs have been successfully applied to fuse HSI and LiDAR data with a total variation subspace model in [13]. Regarding various supervised fusion algorithms, a high number of research works have been dedicated towards the development of more robust models, for instance, a generalized graph-based fusion model in [14]; a spare and low-rank component model in [15]; a multi-sensor composite kernel model in [16]; a decision-level fusion model based on a differential evolution method in [17]; semi-supervised graph-based fusion in [18]; and discriminant correlation analysis in [19]. One mutual objective of these fusion algorithms is to simultaneously determine the optimized classification decision boundary by considering heterogeneous feature spaces. Nevertheless, their success often requires a comprehensive understanding of sensor systems and individual domain expertise, and hand-crafted morphological features are naturally redundant and may still suffer from problems such as the *curse of dimensionality*, which is also termed as *Hughes phenomenon* [20].

More recently, the rapid development of deep learning techniques has led to an explosive growth in the field of remote sensing image processing, especially the classification of HSI [21]. Deep learning models, especially convolutional neural networks (CNNs), open up a new possibility for invariant feature learning of HSI data, from hand-crafted to end-to-end, from manual configurations to fully automatic, from shallow to deep [22].

At the same time, there are various research efforts developing novel multi-sensor fusion approaches based on deep learning [23–28]. Among the first studies, in [23], a deep fusion model was designed for the fusion of HSI and LiDAR data, where CNNs performed as both feature extractor and classifier. In [24], the joint use of HSI and LiDAR data was further explored by combining morphological EPs and high-level deep features via a composite kernel (CK) technique. In [25], a dual-branch CNN was proposed to learn spectral-spatial and elevation features from HSI and LiDAR, respectively, then all features were fused via a cascaded network. Besides the fusion of HSI and LiDAR data, the similar superior performance of deep learning models was also confirmed in [26], where Landsat-8 and Sentinel-2 satellite images were fed into a two branched residual convolutional neural networks (ResNet) for local climate zone classification. However, the training of such deep learning fusion models might be challenging, with problems arising from the fact that deep fusion models mostly require sophisticated network designs with more parameters to simultaneously handle multi-sensor inputs, while the network training will become more difficult when the network becomes deeper [29].

Fortunately, these issues can be mitigated using the residual learning technique, where low-level features are successively passed to deeper layers via identity mapping [30]. Based on this approach, we propose a novel multi-sensor fusion framework via designing multi-branched coupled residual convolutional neural networks, namely CResNet. Moreover, the proposed framework is designed to be a generalized deep fusion framework, where the inputs are not limited to specific sensor systems. To this end, the proposed framework is designed to automatically fuse different types of multi-sensor datasets.

The proposed CResNet mainly consists of three individual ResNet branches along with coupled fully connected layers for data fusion. Different to [24], which requires a separate training step of CK classifiers, the proposed CResNet is trained in an end-to-end manner which lowers the computational complexity during data fusion. To highlight the generalized fusion capability of CResNet, we test the proposed framework on three distinct multi-sensor datasets with inputs ranging from HSI, RGB to LiDAR feature spaces, and various land cover classes. The major contributions of this paper are summarized as threefold:

1. The proposed CResNet adopts novel residual blocks (RBs) with identity mapping to address the gradient vanishing phenomenon and promotes the discriminant feature learning from multi-sensor datasets.
2. The design of coupling individual ResNet with auxiliary loss enables the CResNet to simultaneously learn representative features from each dataset by considering an adjusted loss function, and fuse them in a fully automatic end-to-end manner.
3. Considering that CResNet is highly modularized and flexible, the proposed framework leads to competitive data fusion performance on three commonly used multi-sensor datasets, where the state-of-the-art classification accuracy are achieved using limited training samples.

Section 2 describes the concept of residual feature learning and introduces the detailed architecture of the CResNet. The data descriptions and experimental setups are reported in Section 3. Then, Section 4 is devoted to the discussion of experiment results on three multi-sensor datasets. The main conclusions are summarized in Section 5.

2. Methodology

We present the structure of the proposed CResNet as shown in Figure 1. The fusion framework can be divided into three main components: feature learning via residual blocks, multi-sensor data fusion via coupled ResNet, and auxiliary training via an adjusted loss function. Although there is no limit in the number of datasets being fused using the proposed method, we evaluate the framework by applying it on three co-registered datasets for multi-sensor data fusion and classification.

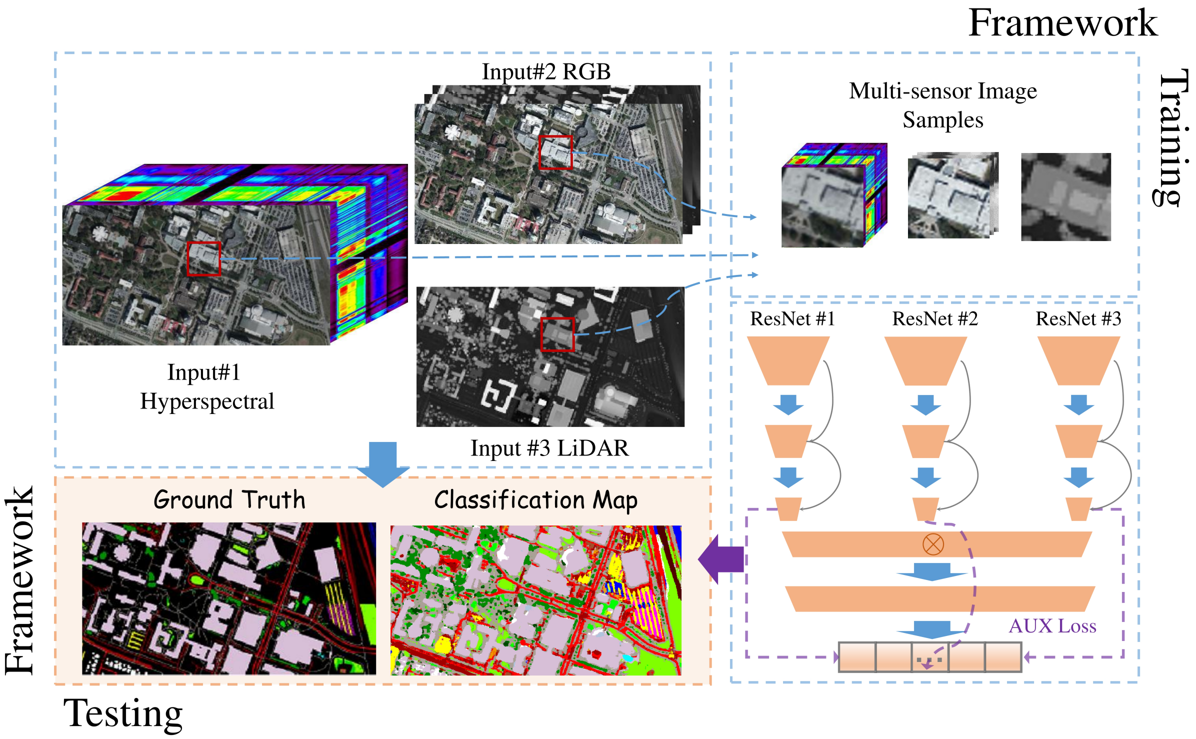

Figure 1. The illustration of the proposed framework in training and testing phases.

2.1. Feature Learning via Residual Blocks

Recently, ResNet has become a popular deep learning technique [29], and has achieved significant classification performance on heterogeneous remote sensing datasets [31,32], where multi-sensor data sources (e.g., HSI, MSI, LiDAR) have been intensively investigated. Residual blocks (RBs), as the characterized architecture of ResNet, are proposed to alleviate the gradient vanishing and explosion issues of CNNs during training [29]. By solving the optimization degradation issue, such blocks are found to be helpful in terms of training accuracy, which is a prerequisite for testing and validation accuracies. In this paper, ResNet with multiple RBs are selected as the base feature learning networks, which are lately aggregated together as a generalized multi-branched data fusion network. As shown in Figure 2, a residual block can be considered to be an extension of several convolutional layers, where gradients in the deeper layers could be intuitively propagated back to the lower layers via identity mapping. To be noticed, identity mapping was proposed in [30] to further improve the training and regularization of origin design of ResNet in [29].

Within each RB, we follow the design in [30] and have three successive convolutional layers with kernel sizes of $1 \times 1 \times m$, $5 \times 5 \times m$, and $1 \times 1 \times m$, respectively, where m refers to the number of feature maps. In addition, such successive layers are also named *bottleneck* designs consisting of a $1 \times 1 \times m$ layer for dimension reduction, a $5 \times 5 \times m$ convolution layer, and a $1 \times 1 \times m$ layer for restoring dimension, with which we can optimize the model complexity, thus lead to a more efficient model due to computational consideration [29]. X_k and X_{k+1} refer to the input and output feature spaces of RBs, respectively, and their feature sizes are kept unchanged via a valid padding strategy. More importantly, by applying the identity mapping with full pre-activation feature spaces into deeper layers [30], the functionality of RBs is further formulated as follows:

$$X_{k+1} = X_k + F\left(\mathbf{X}_k, \mathbf{W}_k\right) \quad (1)$$

where $\mathbf{X}_k$ refers to feature maps of (k)th layer, and the $\mathbf{W}_k$ are the weights and biases of (k)th layer in the RBs. The function (F) is the pre-activation function, which combines the batch normalization function (BN) [33] and the nonlinear activation function (ReLUs) [34] in order to improve the speed and stability of the proposed CResNet.

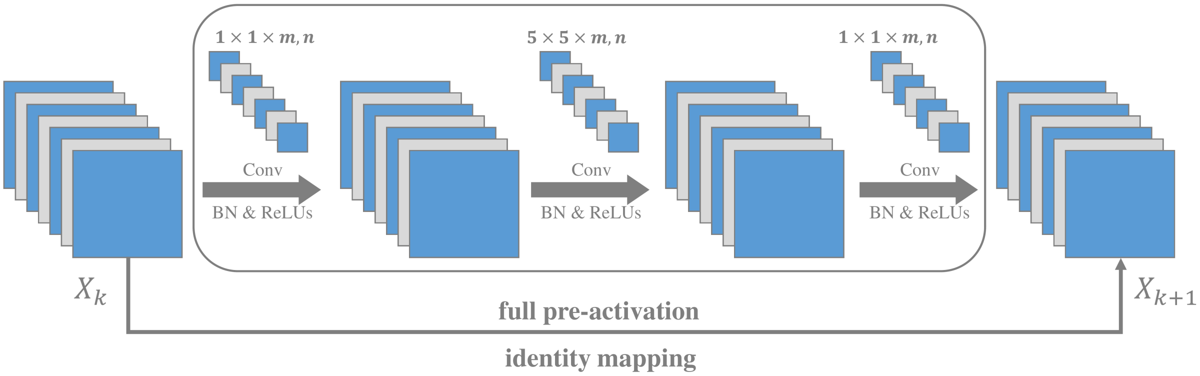

Figure 2. The network architecture of full pre-activation RBs.

Figure 2 shows how the full pre-activation shortcut connection is a direct channel for the gradient to propagate in both directions, forward and backward. Hence, the training process of such RBs is simplified and leads to improved generalization capabilities. One of the key characteristics of the full pre-activation shortcut would become more obvious, when multiple RBs are trained successively, thus we could recursively formulate the feature spaces as follows:

$$\begin{aligned} X_{k+2} &= X_{k+1} + F\left(\mathbf{X}_{k+1}, \mathbf{W}_{k+1}\right), \\ &= X_k + F\left(\mathbf{X}_k, \mathbf{W}_k\right) + F\left(\mathbf{X}_{k+1}, \mathbf{W}_{k+1}\right), \end{aligned} \tag{2}$$

where $\mathbf{W}_k$ are the weights and biases of (k)th layer in the RBs. Next, based on these recursive feature spaces, Equation (1) evolves as follows:

$$X_L = X_k + \sum_{l=k}^{L-1} F\left(\mathbf{X}_l, \mathbf{W}_l\right) \tag{3}$$

Hence, the feature space of any deeper layers (L) can be formulated as the feature space of any lower layers (k) plus a collection of convolutional functions $\sum_{l=k}^{L-1} F$. Moreover, this characteristic ensures the backward propagation of model gradients into lower layers as well, benefitting the overall feature learning with heterogeneous remote sensing datasets. For more detailed description of full pre-activation identity mapping, please refer to [30].

Here, the ResNet consisting of RBs with identity mapping is able to learn discriminative multi-sensor features from heterogeneous data sets due to their simplified training process, which further leads to better generalization capabilities. In this work, heterogeneous deep features are then fused with a coupled fully connected layer and a SoftMax layer (shown in Figure 3) for classification purpose. Regarding comprehensive investigations of deep learning feature extraction technique (i.e., HSI), one can further refer to [22,35].

2.2. Multi-Sensor Data Fusion via Coupled ResNets

In this paper, multi-sensor datasets are fused via coupled three-branched ResNets as shown in Figure 3. Given a set of heterogeneous input datasets $\mathbf{Y}_a \in \Re^{n \times m \times a}$, $\mathbf{Y}_b \in \Re^{n \times m \times b}$, and $\mathbf{Y}_c \in \Re^{n \times m \times c}$, for which various combination of HSI, RGB, (multispectral) LiDAR, and features generated by morphological methods (e.g., extinction profiles [12,36]) are considered in this paper in order to validate the performance of the proposed framework. More in detail, n and m refer to the spatial dimensions of image height and width, and a to c are the number of spectral bands of the input datasets.

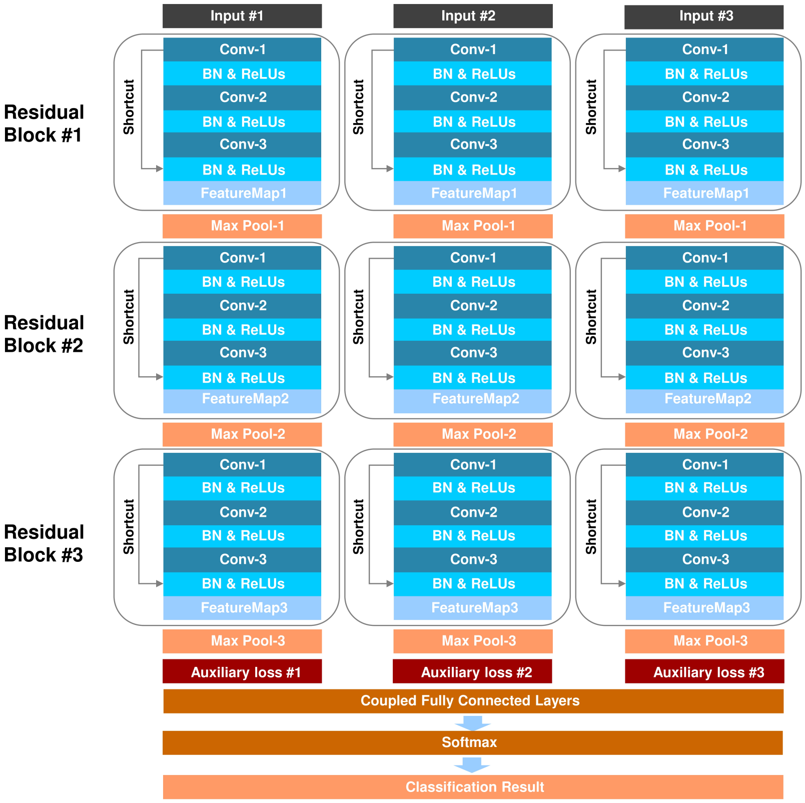

Figure 3. Network design of the proposed coupled residual convolutional neural networks.

As illustrated in Figure 1, for each pixel of inputs, a set of image patches $\mathbf{y}_a \in \Re^{s \times s \times a}$, $\mathbf{y}_b \in \Re^{s \times s \times b}$, and $\mathbf{y}_c \in \Re^{s \times s \times c}$ centered at the chosen pixel are extracted from $\mathbf{Y}_a$, $\mathbf{Y}_b$, and $\mathbf{Y}_c$, individually. Here, s refers to the neighboring window size, for which we empirically selected 24 according to [24,35]. Then the three multi-sensor image patches are fed into separate ResNets for residual feature learning, where each ResNet consists of three RBs. Regarding the classification tasks of HSI, two major challenges identified when applying supervised deep learning classification methods: the high heterogeneity and nonlinearity of spectral signatures and the few training samples against the high dimensionality of HSI [21]. In this context, the nonlinear spectral signature of corresponding ground surfaces can be better captured by coupling networks with multi-sensor inputs (e.g., LiDAR, HSI, and RGB) [1]. By connecting the lower features through the networks to the deeper layers, the design of such RBs provides an efficient way to train the deep learning classification networks even with limited training samples.

Between each of the RBs of ResNet, a 2D max-pooling layer is attached with a kernel size and a stride of 2 in order to reduce the feature variance as well as the computational complexity, with which the spatial dimension of deep feature from the previous layer is halved. In addition, since we empirically selected 24 as the neighboring window size, each individual ResNet consists of three RBs. With such a design, three RBs are trained successively to learn discriminative multi-sensor features. In addition, we increased the number of feature maps towards deeper blocks, which is doubled after each block. Here, the number of feature maps for all three RBs ranges from $\{32, 64, 128\}$. Next, a coupled fully connected layer with the SoftMax function is adopted to fuse the learned feature according to the total amount of classification categories. We use the element-wise maximization to keep the feature number unchanged even after data fusion.

2.3. Auxiliary Training via Adjusted Loss Function

Besides the coupled ResNets, an auxiliary training strategy is proposed to compensate the major loss function according to the training progress of each branch during the framework training stage. The auxiliary loss is a common technique used in other deep learning architecture (e.g., Inception network [37]). In our case, given a set of training samples $\{\mathbf{y}_a^i, \mathbf{y}_b^i, \mathbf{y}_c^i\}$ together with ground-truth labels $\mathbf{t}^i$ and predicted labels $\hat{\mathbf{t}}^i$, where $\{i = 1, 2, \ldots N\}$ and N is the number of training samples, the main model loss could be computed by the categorical cross-entropy loss function.

$$\mathcal{L} = (-1) \times \frac{1}{N} \sum_{i=1}^{N} \left[\mathbf{t}^i \log\left(\hat{\mathbf{t}}^i\right) + \left(1 - \mathbf{t}^i\right) \log\left(1 - \hat{\mathbf{t}}^i\right) \right] \tag{4}$$

Besides the main categorical cross-entropy loss, individual auxiliary loss functions specified for different input branches $\{\mathbf{y}_a^i, \mathbf{y}_b^i, \mathbf{y}_c^i\}$ are computed in a similar manner, where $\mathcal{L}_a$, $\mathcal{L}_b$, and $\mathcal{L}_c$ are designed to guide the training process of each input dataset respectively. Moreover, our auxiliary training strategy further adjusts the main loss using these auxiliary losses as follows:

$$\mathcal{L}_{AUX} = \mathcal{L} + \varepsilon_a \times \mathcal{L}_a + \varepsilon_b \times \mathcal{L}_b + \varepsilon_c \times \mathcal{L}_c \tag{5}$$

where $\{\varepsilon_a, \varepsilon_b, \varepsilon_c\}$ are the weights of auxiliary losses in the overall loss function. To set up the weights, there are two main considerations: first, the auxiliary losses should help in passing information through different branches and prevented from disturbing the overall training process; second, the main loss should be the most important, thus the weights of auxiliary losses should be smaller than the main loss. In this paper, we empirically set $\{\varepsilon_i = 10^{-4} \mid i = a, b, c\}$.

The auxiliary loss function $\mathcal{L}_{AUX}$ could be considered to be an intelligent regularization that helps to make features from individual branches more accurate. More importantly, $\mathcal{L}_{AUX}$ only provides complementary information during the training phase of our framework, not affecting the testing phase.

3. Experiment

3.1. Data Descriptions

3.1.1. Houston 2013

The Houston 2013 dataset is from an urban area of Houston, USA, which was originally distributed for the 2013 GRSS Data Fusion Contest [38]. The image size of the HSI and LiDAR-derived data are 349 × 1905 with a spatial resolution of 2.5 m. The HSI data includes 144 spectral bands, which range from 0.38 to 1.05 µm. Here, the HSI data are cloud-shadow removed. The Houston 2013 dataset has in total 15 classes in the scheme, which range from different vegetation types to highway features. Figure 4 shows the false color HSI, the LiDAR-derived DSM together with the corresponding training and testing samples. The detailed number of training and test samples are listed in Table 1.

Table 1. Houston University 2013: The number of training samples, testing samples, and the total number of samples per class.

Class No.	Class Name	Training	Testing	Samples
1	Healthy grass	198	1053	1251
2	Stressed grass	190	1064	1254
3	Synthetic grass	192	505	697
4	Tree	188	1056	1244
5	Soil	186	1056	1242
6	Water	182	143	325
7	Residential	196	1072	1268
8	Commercial	191	1053	1244
9	Road	193	1059	1252
10	Highway	191	1036	1227
11	Railway	181	1054	1235
12	Parking Lot 1	192	1041	1233
13	Parking Lot 2	184	285	469
14	Tennis court	181	247	428
15	Running track	187	473	660
	Total	2832	12,197	15,029

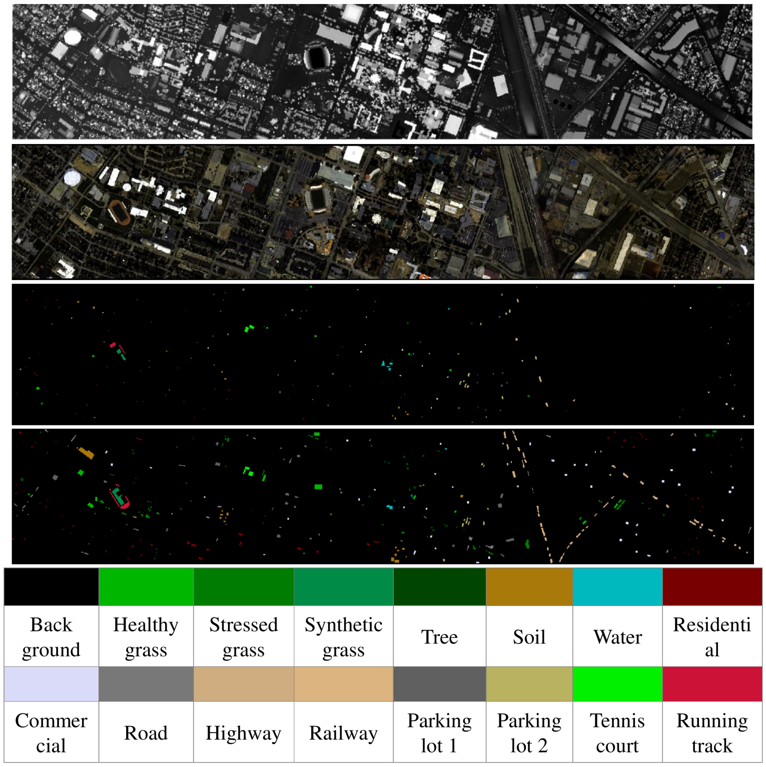

Figure 4. Houston 2013: From top to bottom, the LiDAR-derived DSM image, the false color HSI image, the training samples, and the testing samples.

3.1.2. Houston 2018

The Houston 2018 dataset (identified as GRSS_DFC_2018 dataset) captured over the area of the University of Houston, contains HSI, multispectral LiDAR, and very high resolution (VHR) RGB images. This dataset was originally distributed for the 2018 GRSS Data Fusion Contest [39]. In this paper, we used the training portion of the dataset. The HSI dataset was captured using an ITRES CASI 1500 in 48 bands with spectral range 380–1050 nm at a 1 m ground sampling distance (GSD). The multispectral LiDAR data were acquired using an Optech Titam MW (14SEN/CON340), which include point cloud data at 1550, 1064, and 532 nm, intensity raster, and DSMs at a 50 cm GSD. The RGB was acquired with a VHR RGB imager (DiMAC ULTRALIGHT) with a 70 mm focal length. The VHR color image includes Red, Green, and Blue bands at a 5 cm GSD. This co-registered dataset contains 601 × 2384 pixels. Twenty classes of interest were extracted for Houston data and corresponding training and test samples are given in Figure 5. Figure 5 also depicts the LiDAR-derived DSM and the VHR RGB image (downsampled). The number of training and testing samples used in this study are given in Table 2.

Table 2. Houston University 2018: The number of training samples, testing samples, and the total number of samples per class.

Class No.	Class Name	Training	Testing	Samples
1	Healthy grass	1458	8341	9799
2	Stressed grass	4316	28,186	32,502
3	Synthetic grass	331	353	684
4	Evergreen Trees	2005	11,583	13,588
5	Deciduous Trees	676	4372	5048
6	Soil	1757	2759	4516
7	Water	147	119	266
8	Residential	3809	35,953	39,762
9	Commercial	2789	220,895	223,684
10	Road	3188	42,622	45,810
11	Sidewalk	2699	31,303	34,002
12	Crosswalk	225	1291	1516
13	Major Thoroughfares	5193	41,165	46,358
14	Highway	700	9149	9849
15	Railway	1224	5713	6937
16	Paved Parking Lot	1179	10,296	11,475
17	Gravel Parking Lot	127	22	149
18	Cars	848	5730	6578
19	Trains	493	4872	5365
20	Seats	1313	5511	6824
	Total	34,477	470,235	504,712

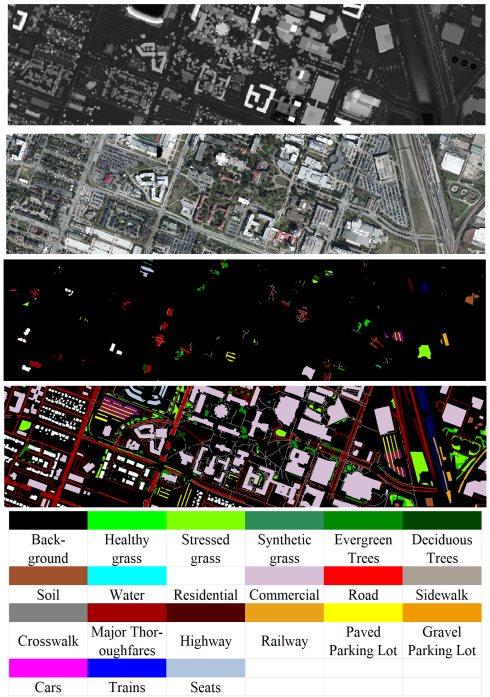

Figure 5. Houston 2018: From top to bottom, the LiDAR-derived DSM image, the VHR RGB Image (downsampled), the training samples, and the testing samples.

3.1.3. Trento

The Trento dataset was captured over a rural area in the south of the city of Trento, Italy. LiDAR and HSI data were acquired by the Optech ALTM 3100EA and the AISA Eagle sensor, respectively. This data has a spatial resolution of 1 m. The size of data is of 600 × 166 pixels in 63 bands ranging from 402.89 to 989.09 nm with the spectral resolution of 9.2 nm. Six classes of interest were extracted for this dataset, including Buildings, Wood, Apple trees, Roads, Vineyard, and Ground. A false color composite of the HSI data and the corresponding training and testing samples are shown in Figure 6. The number of training and testing samples for different classes of interest are given in Table 3.

Table 3. Trento: The number of training samples, testing samples, and the total number of samples per class.

Class No.	Class Name	Training	Testing	Samples
1	Apple trees	129	3905	4034
2	Buildings	125	2778	2903
3	Ground	105	374	479
4	Wood	154	8969	9123
5	Vineyard	184	10,317	10,501
6	Roads	122	3052	3174
	Total	819	29,395	30,214

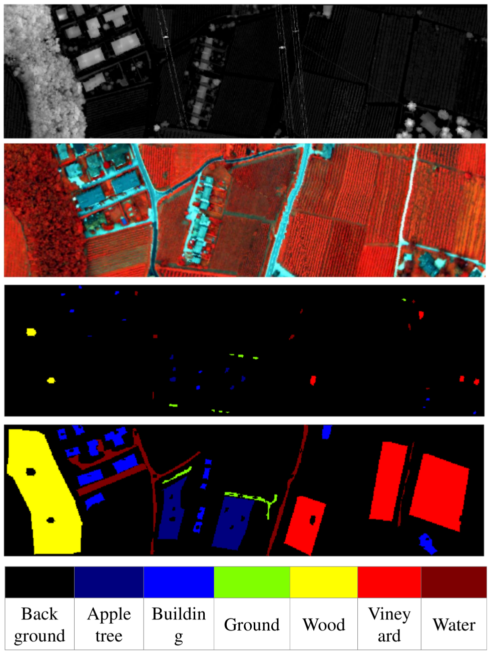

Figure 6. Trento: From top to bottom, the LiDAR-derived DSM image, the false color HSI image, the training samples, and the testing samples.

3.2. Experimental Setup

To evaluate generalized performance of the proposed data fusion framework, the aforementioned three datasets, consisting of two or three co-registered multi-sensor inputs are explored in different ways. In detail, as for the Houston 2013 and Trento datasets, the morphological EPs features of HSI and LiDAR are generated to extract the corresponding spatial and elevation information [12], then a single branch ResNet is used to classify HSI, LiDAR, EPs-HSI, and EPs-LiDAR, respectively. As for the Houston 2018 dataset, instead of using morphological features, HSI, LiDAR, and RGB are directly classified with a single branch ResNet, respectively. Next, the combinations of EPs features and HSI are fused with the proposed CResNet for the Houston 2013 and Trento datasets, while a distinct combination of RGB, LiDAR, and HSI are considered with the Houston 2018 dataset in order to validate the proposed framework's generalized capability in handling highly heterogeneous input datasets.

The implementation of CResNet is based on the Tensorflow framework together with the Keras functional API. The Nesterov Adam optimizer is selected as the optimization algorithm for our ResNet due to its faster convergence performance compared with the stand stochastic gradient descent algorithm [26], where default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ are used. The learning rate, training epochs and batch size are set to 0.001, 200, 64, respectively.

We evaluated the classification accuracy of our proposed framework with respect to the overall accuracy (OA), the average accuracy (AA), the Kappa coefficient, and individual class accuracy. Since the Houston 2013 dataset is intensively used in the state-of-the-art data fusion research, we thus compared the performance of our proposed framework with previous analyses on this dataset.

4. Discussion

4.1. Classification Results

4.1.1. Fusion Performance of Morphological EPs and HSI

Tables 4 and 5 give the results of the fusion of morphological EPs and HSI using CResNet for the Houston 2013 and Trento datasets, respectively. CResNet-AUX denotes to CResNet trained with adjusted auxiliary loss function. The results are compared with the results obtained from EPs-LiDAR-ResNets, EPs-HSI-ResNets, LiDAR-ResNets, and HSI-ResNets.

- First, it is observed that HSI-ResNet considerably outperforms LiDAR-ResNet for both datasets, which also supports that the redundant spectral-spatial information of HSI has higher discriminative capability than the elevation information of LiDAR data. However, we notice that such discriminative capability of morphological feature (EPs-HSI and EPs-LiDAR) may become relatively uniform, for which EPs-HSI outperforms by 1.24% in the Houston 2013 and EPs-LiDAR outperforms by 2.88% in the Trento dataset. The reason behind this could be that morphological features consist of low-level features based on hand-crafted feature engineering, which not only extracts informative features but also bring high redundancy into feature space, thus the integration of low-level hand-crafted features and high-level deep features can further boost the classification performance [24].
- Second, the fusion of EPs and HSI with CResNet+AUX achieves the best OA (93.57% and 98.81%), AA (93.44% and 94.50%) in both datasets, again confirming the capability and effectiveness of the proposed framework in invariant feature learning from both low-level morphological features and high-level deep features.
- Finally, we observe a common improvement of classification accuracy by training ResNet with adjusted auxiliary loss function. In the Houston 2013 dataset, CResNet-AUX outperforms the original CResNet by producing the highest OA (93.57%) and AA (93.44 %) as well as kappa value of 0.9302. Similar findings are also discovered in the Trento dataset. As explained in Section 2.3, the performance boosting can be attributed to the design of our auxiliary training strategy, where the overall loss function is regularized with the complementary losses from each individual dataset.

Table 4. Houston 2013: Classification accuracies for per class, OA, AA (in %), kappa coefficient (is of no unit). The bold refers to the best OA, AA, and Kappa performance.

#	Class	HSI-ResNet	LiDAR-ResNet	EPs-HSI-ResNet	EPs-LiDAR-ResNet	CResNet	CResNet-AUX
	Number of features	(144)	(1)	(225)	(71)	(144+225+71)	(144+225+71)
1	Healthy grass	77.68	51.76	74.83	54.13	83.00	86.51
2	Stressed grass	98.59	47.09	76.60	56.77	99.81	98.01
3	Synthetic grass	86.53	87.33	87.33	94.06	84.36	87.87
4	Tree	86.46	51.52	51.89	68.09	96.69	85.52
5	Soil	89.11	43.56	93.94	52.37	99.91	87.02
6	Water	81.12	78.32	91.61	79.02	95.80	99.81
7	Residential	93.75	67.07	74.07	75.93	90.11	100.00
8	Commercial	81.86	75.12	80.53	83.57	95.73	95.72
9	Road	88.67	58.55	55.71	59.87	90.65	96.68
10	Highway	74.52	73.84	54.05	72.78	70.46	100.00
11	Railway	95.64	90.32	68.98	98.29	94.68	85.54
12	Parking Lot 1	85.78	68.20	73.20	78.10	97.50	95.80
13	Parking Lot 2	82.81	75.44	68.07	72.28	79.30	94.05
14	Tennis court	100.00	90.28	93.12	88.66	100.00	95.10
15	Running track	68.92	39.32	41.23	15.43	89.85	93.87
	OA(%)	86.60	63.82	70.63	69.39	91.42	**93.57**
	AA(%)	86.10	66.51	72.34	69.96	91.19	**93.44**
	Kappa	0.8545	0.6074	0.6809	0.6676	0.9068	**0.9302**

Table 5. Trento: Classification accuracies for per class, OA, AA (in %), kappa coefficient (is of no unit). The bold refers to the best OA, AA, and Kappa performance.

#	Class Number of Features	HSI-ResNet (63)	LiDAR-ResNet (1)	EPs-HSI-ResNet (225)	EPs-LiDAR-ResNet (71)	CResNet (63+225+71)	CResNet-AUX (63+225+71)
1	Apple trees	98.21	0.00	96.67	98.39	98.10	99.74
2	Buildings	93.12	15.77	87.83	97.52	97.77	99.60
3	Ground	77.54	39.84	77.01	64.71	77.01	75.40
4	Wood	98.99	98.27	99.74	100.00	99.90	100.00
5	Vineyard	99.96	97.00	94.92	97.77	100.00	100.00
6	Roads	60.52	2.62	75.75	83.65	92.46	92.27
	OA (%)	94.40	66.30	93.74	96.62	98.43	**98.81**
	AA (%)	88.06	42.25	88.65	90.34	94.21	**94.50**
	Kappa	0.9250	0.5178	0.9166	0.9548	0.9790	**0.9841**

Figures 7 and 8 show classifications corresponding to the aforementioned methods for the Houston 2013 and Trento datasets, respectively. The Houston 2013 dataset is characterized as complex urban structures and mixed residential and commercial areas. From Figure 7a–d, it is shown that single input features are insufficient in distinguishing categories like Highway and Parking lot, for which the multi-sensor fusion methods (Figure 7e,f) are able to produce accurate classification results. In this context, the similar visualization patterns in a rural region of Trento can be obtained, where homogeneous Vineyard is successfully depicted.

It is suggested that deep learning methods need to go deeper in order to learn discriminative features [21], while the training of such methods can become even more challenging, especially with limited training samples. In this paper, we tackle this problem by construing a novel arrangement of RBs with identity mapping that successively pass the low-level features through the entire networks.

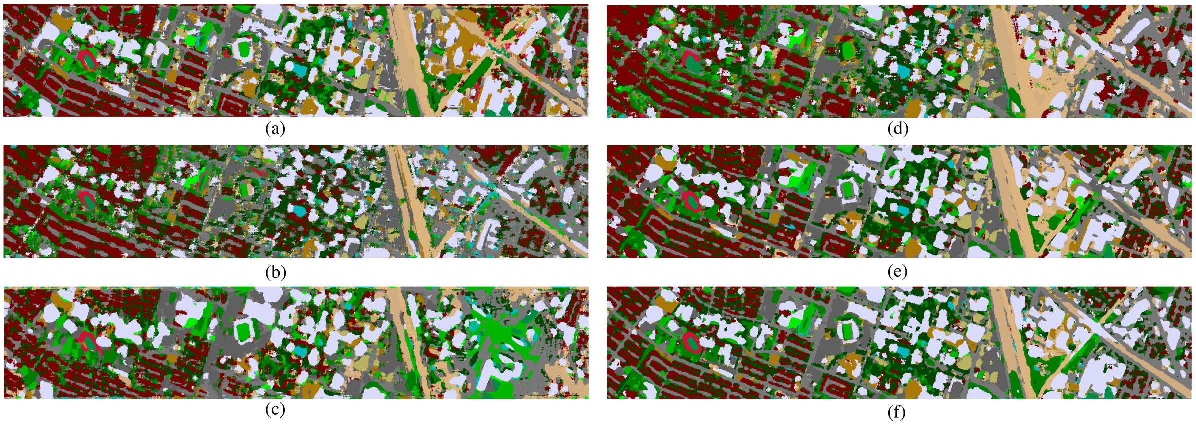

Figure 7. The Houston 2013 dataset: Classifications generated from different features and models. (**a**) HSI-ResNet, (**b**) LiDAR-ResNet, (**c**) EPs-HSI-ResNet, (**d**) EPs-LiDAR-ResNet, (**e**) CResNet, and (**f**) CResNet-AUX.

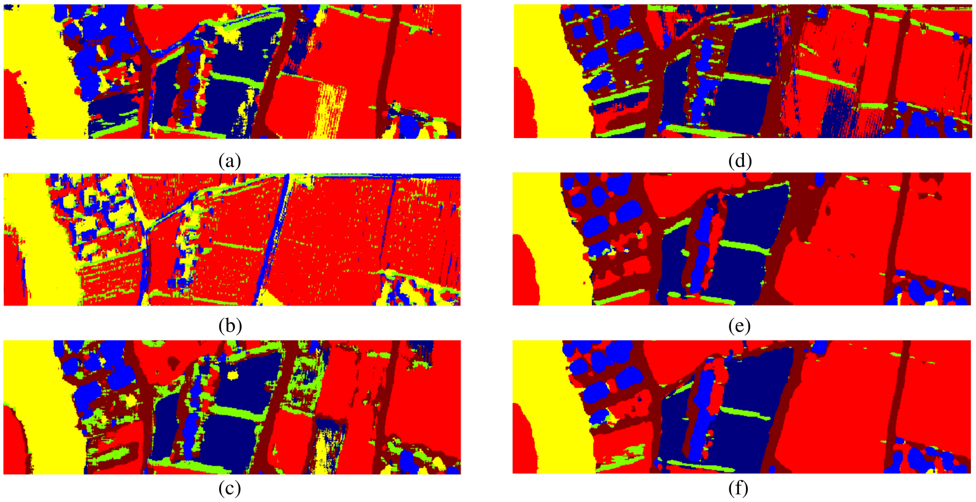

Figure 8. The Trento dataset: Classifications generated from different features and models. (**a**) HSI-ResNet, (**b**) LiDAR-ResNet, (**c**) EPs-HSI-ResNet, (**d**) EPs-LiDAR-ResNet, (**e**) CResNet, and (**f**) CResNet-AUX.

4.1.2. Fusion Performance of RGB, MS LiDAR, and HSI

In this scenario, we do not use EPs. However, we rely on the deep network developed to extracted the spatial, spectral, and elevation features from RGB, HSI, and multispectral LiDAR. Table 6 demonstrates the performance of CResNet for the fusion of HSI, multispectral LiDAR, and RGB. The proposed CResNet fusion framework leads to substantial improvements with respect to HSI (OA: 12.79%), LiDAR (OA: 10.36%), and RGB (OA: 11.09%). Additionally, the results show that the auxiliary training could further improve the OA by 0.58%. To be noticed here, the degradation of individual accuracy in Water class can be potentially attributed to the high imbalance of training sample numbers as listed in Table 2.

Table 6. Houston 2018: Classification accuracies for per class, OA, AA (in %), kappa coefficient (is of no unit). The bold refers to the best OA, AA, and Kappa performance.

#	**Class** **Number of features**	**HSI-ResNet** **(48)**	**LiDAR-ResNet** **(7)**	**RGB-ResNet** **(3)**	**CResNet** **(48+7+3)**	**CResNet-AUX** **(48+7+3)**
1	**Healthy grass**	46.35	24.25	41.54	18.77	75.90
2	**Stressed grass**	79.64	74.80	79.37	90.43	67.79
3	**Synthetic grass**	82.72	100.00	100.00	100.00	100.00
4	**Evergreen Trees**	93.59	90.02	93.05	94.74	95.24
5	**Deciduous Trees**	46.27	43.62	44.26	59.54	59.47
6	**Soil**	36.17	31.39	86.48	43.82	36.82
7	**Water**	42.02	0.00	22.69	30.25	1.68
8	**Residential**	89.86	87.51	91.08	90.79	88.00
9	**Commercial**	71.24	70.89	66.35	92.71	92.75
10	**Road**	54.44	61.35	65.97	64.14	72.77
11	**Sidewalk**	63.14	73.80	75.18	62.26	71.27
12	**Crosswalk**	3.95	2.40	2.87	3.02	3.95

Table 6. *Cont.*

#	Class Number of features	HSI-ResNet (48)	LiDAR-ResNet (7)	RGB-ResNet (3)	CResNet (48+7+3)	CResNet-AUX (48+7+3)
13	**Major Thoroughfares**	47.50	62.67	56.97	65.15	57.62
14	**Highway**	31.82	34.97	37.22	42.34	44.82
15	**Railway**	77.58	84.75	84.74	63.77	63.96
16	**Paved parking Lot**	85.60	97.31	94.80	83.64	89.48
17	**Gravel parking Lot**	100.00	100.00	100.00	100.00	100.00
18	**Cars**	32.24	37.24	50.89	29.91	34.57
19	**Trains**	93.49	99.36	98.75	92.44	97.74
20	**Seats**	63.49	99.84	98.42	61.13	73.42
	OA (%)	67.83	70.26	69.53	80.62	**81.20**
	AA (%)	62.16	63.81	69.53	64.47	**66.36**
	Kappa	0.5944	0.6287	0.6253	0.7416	**0.7506**

Figure 9 shows the classifications obtained by different techniques for the Houston 2018 dataset. There are relatively well-mapped ground-truth samples extracted from the original GRSS_DFC_2018 training dataset as shown in Figure 9a. By comparing Figure 9e,f with Figure 9b–d, the improved classifications using CResNet can be observed compared to the other techniques, especially for categories like healthy grass and commercial, where large commercial blocks and grassland are well delineated.

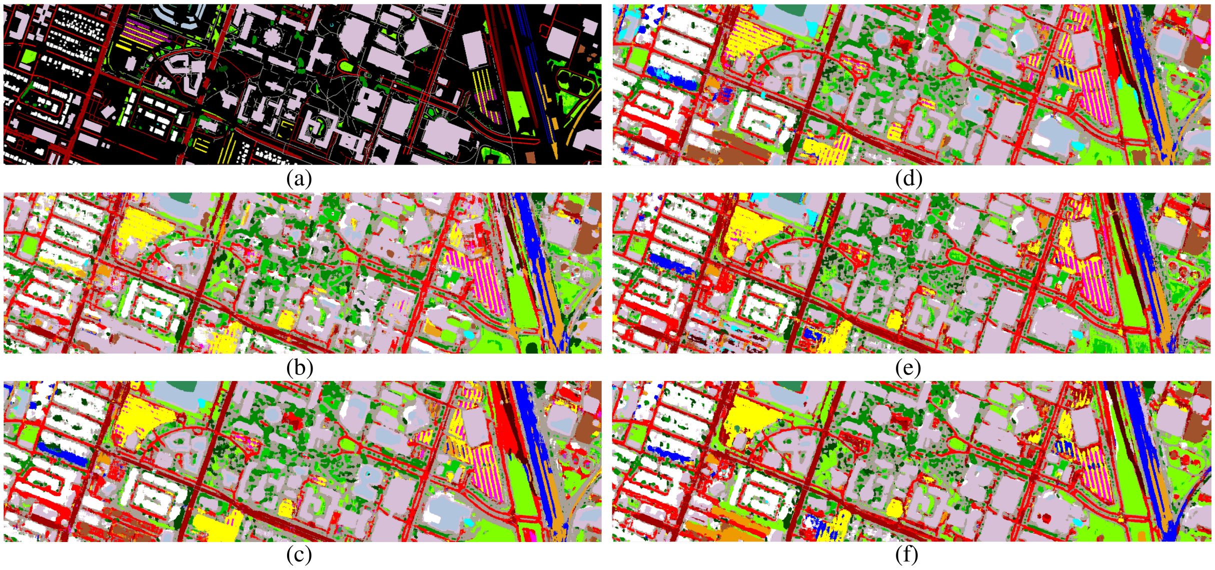

Figure 9. The Houston 2018 dataset: (**a**) Ground-truth label map; (**b**–**f**) classification maps generated on different features and models. (**b**) HSI-ResNet, (**c**) LiDAR-ResNet, (**d**) RGB-ResNet, (**e**) CResNet, and (**f**) CResNet-AUX.

To summarize, based on the results obtained on the Houston 2018 dataset, we can validate the generalized capability of the proposed multi-sensor fusion framework. Although we use a uniform network architecture, the CResNet-AUX can automatically extract informative features via RBs and simultaneously regularize the data fusion via auxiliary loss fusion. The reason could be due to the fact that our CResNet actually consists of much deeper CNNs layers as shown in Figure 3, which can be fitted to different datasets, and trained through residual learning. In this context, we believe that the proposed

CResNet presents a new possibility in developing flexible end-to-end fusion methods even with multiple inputs from different sensor systems.

4.2. Comparison to State-of-the-Art

The Houston 2013 dataset is one of the most widely used datasets, comprising a challenging mixture of urban structures. In this context, we compare the classification performance of our proposed framework with the following state-of-the-art methods listed in Table 7: The multiple subspace feature learning method (**MLR*sub***) in [10], the total variation component-based method (**OTVCA**) in [13], the sparse and low-rank component-based method (**SLRCA**) in [15], the deep fusion method (**DeepFusion**) in [23], the extinction profiles fusion via CNNs and graph-based feature fusion method (**EPs-CNN**) in [8], and the composite kernel-based three-stream CNNs method (**CK-CNN**) in [24]. All these methods including the proposed method in this paper use the benchmark sets of training and testing samples published with the dataset for the classification purpose and therefore, the classification results are fully comparable.

Table 7. Houston 2013: Performance comparison with the state-of-the-art models. The bold refers to the best OA, AA, and Kappa performance.

Methods	MLR*sub* [10]	OTVCA [13]	SLRCA [15]	DeepFusion [23]	EPs-CNN [8]	CK-CNN [24]	CResNet	CResNet-AUX
OA (%)	92.05	92.45	91.30	91.32	91.02	92.57	91.42	**93.57**
AA (%)	92.85	92.68	91.95	91.96	91.82	92.48	91.19	**93.44**
Kappa	0.9137	0.9181	0.9056	0.9057	0.9033	0.9193	0.9068	**0.9302**

In general, these methods can be classified into two main categories: conventional shallow methods and deep learning-based methods. The highest OA, AA, and Kappa for each of those categories are 92.45%, 92.68%, and 0.9181 obtained by **OTVCA** and 92.57%, 92.48%, and 0.9193 obtained by **CK-CNN**, for which the CResNet-AUX improves both methods by around 1% in terms of OA. This performance improvement over the state-of-the-art methods further validates the effectiveness of the proposed multi-sensor framework. In addition, the superior performance compared to existing deep learning-based methods confirmed the effectiveness of the proposed CResNet in mitigating the gradient vanishing phenomenon and discriminant feature learning from heterogeneous datasets. More importantly, with the proposed multi-sensor fusion framework, the data fusion results can be achieved automatically in an end-to-end manner.

4.3. The Performance with Respect to the Number of Training Samples

To evaluate the performance of the proposed framework with respect to the number of training samples, we randomly selected 10, 25, 50, or 100 training samples per class and repeat the experiment 10 times on the Houston 2018 dataset. In Figure 10, the means and standard deviations of OA are depicted with respect to different numbers of training samples using CResNet and CResNet+AUX, respectively. In the case of 10 samples, the OAs are less than 50%, which reveals the dependency of the deep learning techniques to the adequate amount of training samples. However, the high achievements of almost 20% in terms of OA for both techniques in the case of 25 samples per class demonstrates the efficacy of the proposed deep learning-based fusion framework in the case of a limited number of samples. Additionally, the steady increase in the slope of the CResNet+AUX's graph compared with the CResNet's graph confirm that the auxiliary training loss function provides robustness in the performance of the CResNet with respect to the number of samples. Moreover, CResNet+AUX outperforms CResNet for all four cases, which supports the advantage of the CResNet+AUX.

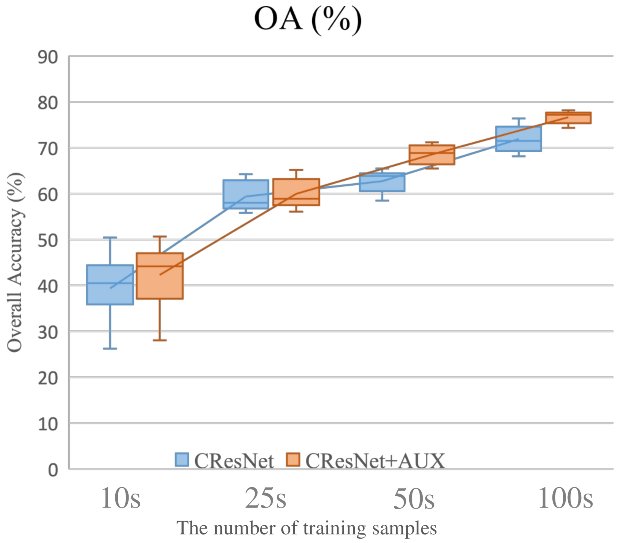

Figure 10. Analysis of the classification OA w.r.t the number of training samples on the Houston 2018 dataset. We select 10, 25, 50, or 100 training samples per each class.

4.4. Sensitivity Analysis of OA with Respect to the Weights of Auxiliary Losses

As mentioned in Section 2.3, the general network training can benefit from considering auxiliary losses from individual branches. Here, we analyzed the sensitivity of CResNet-AUX with respect to ε_i in terms of OA. To test the effect of different $\{\varepsilon_i \mid i = a, b, c\}$, we compared the classification OA for the Houston 2018 dataset by selecting ε_i in the range of $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. In addition, the weights of individual branches are set to be identical, since we assume that no prior knowledge of multi-sensor inputs is available. Figure 11 shows that $\varepsilon_i \geq 10^{-4}$ is a confident region for the selection of ε_i. To this end, we empirically used 10^{-4} in this paper.

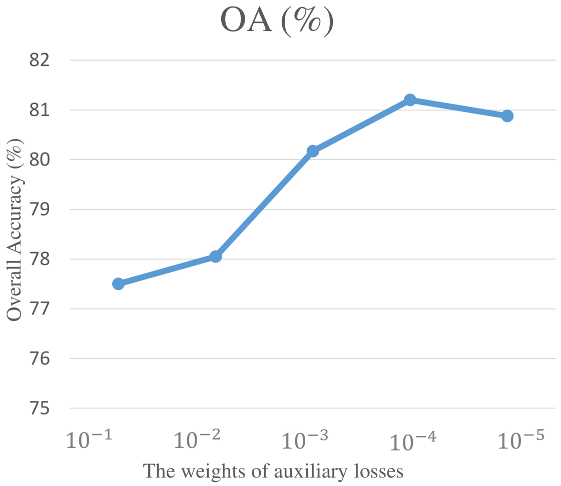

Figure 11. Analysis of classification OA w.r.t the weights of auxiliary losses on Houston 2018 dataset.

4.5. Computational Cost

In addition to the classification accuracy, Table 8 reports the computational cost for the proposed framework, where training and testing times were given in *minutes* and *seconds*, respectively. All experiments were implemented on a workstation with 2 GeForce RTX 2080Ti graphical processing units (GPUs), each with 12 GB memory.

Table 8. Computational time for three multi-sensor datasets. The bold refers to the best OA, AA, and Kappa performance.

Houston 2013	**HSI-ResNet**	**LiDAR-ResNet**	**EPs-HSI-ResNet**	**EPs-LiDAR-ResNet**	**CResNet**	**CResNet-AUX**
Train (min)	8.84	5.837	8.61	**5.67**	16.11	16.61
Test (s)	4.38	**3.04**	5.53	3.61	8.15	16.25
Trento	**HSI-ResNet**	**LiDAR-ResNet**	**EPs-HSI-ResNet**	**EPs-LiDAR-ResNet**	**CResNet**	**CResNet-AUX**
Train (min)	5.69	**5.04**	6.88	5.79	11.73	13.66
Test (s)	6.28	**5.62**	9.15	7.13	13.57	14.06
Houston 2018	**HSI-ResNet**	**LiDAR-ResNet**	**RGB-ResNet**	**CResNet**	**CResNet-AUX**	
Train (min)	82.50	**63.11**	58.13	159.9	168.33	
Test (s)	53.64	**35.84**	38.38	102.91	107.79	

As shown in Table 8, CResNet consumes up to three times more processing time than the individual branches since networks are simultaneously learning from multiple inputs. Compared to the sum of individual branches reveals that the training of CResNet is more efficient and faster, saving up to 35% of training time. However, this computational efficiency may slightly decrease through the application of the auxiliary training strategy because the adjusted loss function can lead to additional computation cost. As shown in Figures 10 and 12, by compromising the training time to some extent, the adjusted auxiliary loss function leads to further accuracy improvement for all three datasets. Therefore, the additional computational cost is justified for our proposed framework. More importantly, although the training time may take up to several hours for the feeding forward of testing samples (measured in seconds), the additional cost is negligible. To summarize, the auxiliary training design can improve general multi-sensor fusion accuracy by adjusting the training time within affordable ranges.

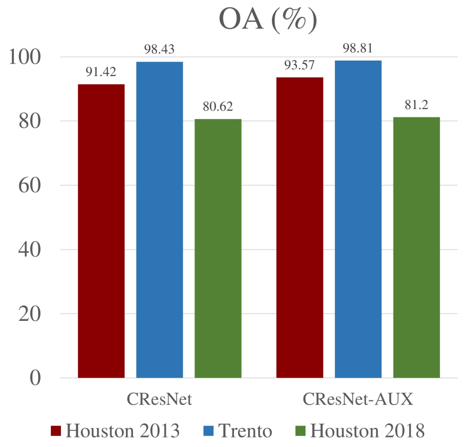

Figure 12. Comparison of classification accuracy with and without auxiliary loss functions for three datasets.

5. Conclusions

In this paper, we presented the development of a novel multi-sensor data fusion framework, which is capable of fusing heterogeneous data types either captured by different sensor systems (e.g., HSI, LiDAR, RGB) or generated by feature extraction algorithms (e.g., extinction profiles). The designed coupled residual neural networks with auxiliary training (i.e., CResNet-AUX) consists of highly modularized residual blocks with identity mapping and an intelligent regularization strategy with adjusted auxiliary loss functions. Extensive experiments were applied on three multi-sensor datasets (i.e., Houston 2013, Trento, and Houston 2018) and based on classification accuracies the following outcomes have been achieved:

- The proposed CResNet fusion framework outperforms all the single sensor-based scenarios in the experiments for all three datasets.
- Both CResNet and CResNet-AUX outperform the state-of-the-art methods for the Houston 2013 dataset.
- The auxiliary training function boosts the performance of CResNet for all the datasets even for the case of limited training samples.
- The proposed CResNet fusion framework shows effective performance when the number of training samples is limited, which is of great importance in the case of applying deep learning techniques for remote sensing datasets.
- The experiments regarding the computational cost justifies the efficiency of the proposed algorithm considering the achievements in the classification accuracies.

More importantly, the proposed CResNet-AUX is designed to be a fully automatic generalized multi-sensor fusion framework, where the network architecture is largely independent from the input data types and not limited to specific sensor systems. Our framework is applicable to a wide range of multi-sensor datasets in an end-to-end, wall-to-wall manner.

Future works in developing intelligent and robust multi-sensor fusion methods may benefit from the insights we have produced in this paper. In further research we propose to test the performance of our framework on a large-scale application (continental and/or planetary) and include additional types of remote sensing data.

Author Contributions: H.L. and P.G. conceived and designed the work. H.L. and Z.W. designed and performed the experiments. H.L. and B.R. analyzed and interpreted the results. H.L. and B.R. and A.S. wrote the manuscript. P.G. and M.S. and A.Z. supervised the work. All authors revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Lorenzo Bruzzone of the University of Trento for providing the Trento data set, and the same thanks to the National Center for Airborne Laser Mapping (NCALM) at the University of Houston for providing the Houston data set and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for distributing the Houston 2013 and 2018 dataset. The work of Behnood Rasti was funded by Alexander von Humboldt foundation. This research was partly funded by the Klaus Tschira Stiftung, Heidelberg.

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

References

1. Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.; Anders, K.; Gloaguen, R.; et al. Multisource and Multitemporal Data Fusion in Remote Sensing: A Comprehensive Review of the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 6–39. [CrossRef]
2. Goetz, A.F.H.; Vane, G.; Solomon, J.E.; Rock, B. Imaging Spectrometry for Earth Remote Sensing. *Science* **1985**, *228*, 1147–1153. Available online: https://science.sciencemag.org/content/228/4704/1147.full.pdf (accessed on 16 December 2019). [CrossRef]

3. Bioucas-Dias, J.M.; Plaza, A.; Camps-Valls, G.; Scheunders, P.; Nasrabadi, N.; Chanussot, J. Hyperspectral Remote Sensing Data Analysis and Future Challenges. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–36. [CrossRef]
4. Ghamisi, P.; Yokoya, N.; Li, J.; Liao, W.; Liu, S.; Plaza, J.; Rasti, B.; Plaza, A. Advances in Hyperspectral Image and Signal Processing: A Comprehensive Overview of the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 37–78. [CrossRef]
5. Eitel, J.U.H.; Höfle, B.; Vierling, L.A.; Abellán, A.; Asner, G.P.; Deems, J.S.; Glennie, C.L.; Joerg, P.C.; LeWinter, A.L.; Magney, T.S.; et al. Beyond 3-D: The new spectrum of lidar applications for earth and ecological sciences. *Remote. Sens. Environ.* **2016**, *186*, 372–392. [CrossRef]
6. Höfle, B.; Hollaus, M.; Hagenauer, J. Urban vegetation detection using radiometrically calibrated small-footprint full-waveform airborne LiDAR data. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 134–147. [CrossRef]
7. Anders, K.; Winiwarter, L.; Lindenbergh, R.; Williams, J.G.; Vos, S.E.; Höfle, B. 4D objects-by-change: Spatiotemporal segmentation of geomorphic surface change from LiDAR time series. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 352–363. [CrossRef]
8. Ghamisi, P.; Höfle, B.; Zhu, X.X. Hyperspectral and LiDAR Data Fusion Using Extinction Profiles and Deep Convolutional Neural Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3011–3024. [CrossRef]
9. Hänsch, R.; Hellwich, O. Fusion of Multispectral LiDAR, Hyperspectral, and RGB Data for Urban Land Cover Classification. *IEEE Geosci. Remote. Sens. Lett.* **2020**, 1–5. [CrossRef]
10. Khodadadzadeh, M.; Li, J.; Prasad, S.; Plaza, A. Fusion of Hyperspectral and LiDAR Remote Sensing Data Using Multiple Feature Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2015**, *8*, 2971–2983. [CrossRef]
11. Pedergnana, M.; Marpu, P.R.; Dalla Mura, M.; Benediktsson, J.A.; Bruzzone, L. Classification of Remote Sensing Optical and LiDAR Data Using Extended Attribute Profiles. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 856–865. [CrossRef]
12. Ghamisi, P.; Souza, R.; Benediktsson, J.A.; Zhu, X.X.; Rittner, L.; Lotufo, R.A. Extinction Profiles for the Classification of Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5631–5645. [CrossRef]
13. Rasti, B.; Ghamisi, P.; Gloaguen, R. Hyperspectral and LiDAR Fusion Using Extinction Profiles and Total Variation Component Analysis. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3997–4007. [CrossRef]
14. Liao, W.; Pizurica, A.; Bellens, R.; Gautama, S.; Philips, W. Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 552–556. [CrossRef]
15. Rasti, B.; Ghamisi, P.; Plaza, J.; Plaza, A. Fusion of Hyperspectral and LiDAR Data Using Sparse and Low-Rank Component Analysis. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6354–6365. [CrossRef]
16. Ghamisi, P.; Rasti, B.; Benediktsson, J.A. Multisensor Composite Kernels Based on Extreme Learning Machines. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 196–200. [CrossRef]
17. Zhong, Y.; Cao, Q.; Zhao, J.; Ma, A.; Zhao, B.; Zhang, L. Optimal Decision Fusion for Urban Land-Use/Land-Cover Classification Based on Adaptive Differential Evolution Using Hyperspectral and LiDAR Data. *Remote Sens.* **2017**, *9*, 868. [CrossRef]
18. Xia, J.; Liao, W.; Du, P. Hyperspectral and LiDAR Classification With Semisupervised Graph Fusion. *IEEE Geosci. Remote Sens. Lett.* **2019**, 1–5. [CrossRef]
19. Jahan, F.; Zhou, J.; Awrangjeb, M.; Gao, Y. Fusion of Hyperspectral and LiDAR Data Using Discriminant Correlation Analysis for Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3905–3917. [CrossRef]
20. Hughes, G.F. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory* **1968**, *14*, 55–63. [CrossRef]
21. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [CrossRef]
22. Rasti, B.; Hong, D.; Hang, R.; Ghamisi, P.; Kang, X.; Chanussot, J.; Benediktsson, J.A. Feature Extraction for Hyperspectral Imagery: The Evolution from Shallow to Deep (Overview and Toolbox). *IEEE Geosci. Remote Sens. Mag.* **2020**. [CrossRef]
23. Chen, Y.; Li, C.; Ghamisi, P.; Jia, X.; Gu, Y. Deep Fusion of Remote Sensing Data for Accurate Classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1253–1257. [CrossRef]

24. Li, H.; Ghamisi, P.; Soergel, U.; Zhu, X.X. Hyperspectral and LiDAR Fusion Using Deep Three-Stream Convolutional Neural Networks. *Remote Sens.* **2018**, *10*, 1649. [CrossRef]
25. Xu, X.; Li, W.; Ran, Q.; Du, Q.; Gao, L.; Zhang, B. Multisource Remote Sensing Data Classification Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 937–949. [CrossRef]
26. Qiu, C.; Schmitt, M.; Mou, L.; Ghamisi, P.; Zhu, X.X. Feature Importance Analysis for Local Climate Zone Classification Using a Residual Convolutional Neural Network with Multi-Source Datasets. *Remote Sens.* **2018**, *10*, 1572. [CrossRef]
27. Zhang, M.; Li, W.; Du, Q.; Gao, L.; Zhang, B. Feature Extraction for Classification of Hyperspectral and LiDAR Data Using Patch-to-Patch CNN. *IEEE Trans. Cybern.* **2020**, *50*, 100–111. [CrossRef]
28. Xu, S.; Amira, O.; Liu, J.; Zhang, C.; Zhang, J.; Li, G. HAM-MFN: Hyperspectral and Multispectral Image Multiscale Fusion Network With RAP Loss. *IEEE Trans. Geosci. Remote Sens.* **2020**, 1–11. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 630–645.
31. Zhong, Z.; Li, J.; Luo, Z.; Chapman, M. Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 847–858. [CrossRef]
32. Qiu, C.; Mou, L.; Schmitt, M.; Zhu, X.X. Fusing Multiseasonal Sentinel-2 Imagery for Urban Land Cover Classification With Multibranch Residual Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2020**, 1–5. [CrossRef]
33. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; Bach, F., Blei, D., Eds.; PMLR: Lille, France, 2015; Volume 37, pp. 448–456.
34. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning (ICML 2010), Haifa, Israel, 21–24 June 2010
35. Chen, Y.H. Jiang, C.L.; Jia, X.; Ghamisi, P. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6250. [CrossRef]
36. Ghamisi, P.; Souza, R.; Benediktsson, J.A.; Rittner, L.; Lotufo, R.; Zhu, X.X. Hyperspectral Data Classification Using Extended Extinction Profiles. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1641–1645. [CrossRef]
37. 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 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
38. Debes, C.; Merentitis, A.; Heremans, R.; Hahn, J.; Frangiadakis, N.; van Kasteren, T.; Liao, W.; Bellens, R.; Pizurica, A.; Gautama, S.; et al. Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2405–2418. [CrossRef]
39. 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]