


## Article

# Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks

Diogo Duarte , Francesco Nex , Norman Kerle and George Vosselman 

Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede 7500 AE, The Netherlands; f.nex@utwente.nl (F.N.); n.kerle@utwente.nl (N.K.); george.vosselman@utwente.nl (G.V.)

\* Correspondence: d.duarte@utwente.nl; Tel.: +31-54-34896662

Received: 27 July 2018; Accepted: 9 October 2018; Published: 14 October 2018



**Abstract:** Remote sensing images have long been preferred to perform building damage assessments. The recently proposed methods to extract damaged regions from remote sensing imagery rely on convolutional neural networks (CNN). The common approach is to train a CNN independently considering each of the different resolution levels (satellite, aerial, and terrestrial) in a binary classification approach. In this regard, an ever-growing amount of multi-resolution imagery are being collected, but the current approaches use one single resolution as their input. The use of up/down-sampled images for training has been reported as beneficial for the image classification accuracy both in the computer vision and remote sensing domains. However, it is still unclear if such multi-resolution information can also be captured from images with different spatial resolutions such as imagery of the satellite and airborne (from both manned and unmanned platforms) resolutions. In this paper, three multi-resolution CNN feature fusion approaches are proposed and tested against two baseline (mono-resolution) methods to perform the image classification of building damages. Overall, the results show better accuracy and localization capabilities when fusing multi-resolution feature maps, specifically when these feature maps are merged and consider feature information from the intermediate layers of each of the resolution level networks. Nonetheless, these multi-resolution feature fusion approaches behaved differently considering each level of resolution. In the satellite and aerial (unmanned) cases, the improvements in the accuracy reached 2% while the accuracy improvements for the airborne (manned) case was marginal. The results were further confirmed by testing the approach for geographical transferability, in which the improvements between the baseline and multi-resolution experiments were overall maintained.

**Keywords:** earthquake; deep learning; UAV; satellite; aerial; dilated convolutions; residual connections

## 1. Introduction

The location of damaged buildings after a disastrous event is of utmost importance for several stages of the disaster management cycle [1,2]. Manual inspection is not efficient since it takes a considerable amount of resources and time. Preventing the use of such inspections results in the early response phase of the disaster management cycle [3]. Over the last decade, remote sensing platforms have been increasingly used for the mapping of building damages. These platforms usually have a wide coverage, fast deployment, and high temporal frequency. Space, air, and ground platforms mounted with optical [4–6], radar [7,8], and laser [9,10] sensors have been used to collect data to perform automatic building damage assessment. Regardless of the platform and sensor used, several central difficulties persist, such as the subjectivity in the manual identification of hazard-induced damages

from the remote sensing data, and the fact that the damage evidenced by the exterior of a building might not be enough to infer the building's structural health. For this reason, most scientific contributions aim towards the extraction of damage evidence such as piles of rubble, debris, spalling, and cracks from remote sensing data in a reliable and automated manner.

Optical remote sensing images have been preferred to perform building damage assessments since these data are easier to understand when compared with other remote sensing data [1]. Moreover, these images may allow for the generation of 3D models if captured with enough overlap. The 3D information can then be used to infer the geometrical deformations of the buildings. However, the time needed for the generation of such 3D information through dense image matching might hinder its use in the search and rescue phase because fast processing is mandatory in this phase.

Synoptic satellite imagery can cover regional to national extents and can be readily available after a disaster. The International Charter (IC) and the Copernicus Emergency Management Service (EMS) use synoptic optical data to assess building damage after a disastrous event. However, many signs of damage may not be identifiable using such data. Pancake collapses and damages along the façades might not be detectable due to the limited viewpoint of such platforms. Furthermore, its low resolution may introduce uncertainty in the satellite imagery damage mapping [11], even when performed manually [12,13].

To overcome these satellite imagery drawbacks, airborne images collected from manned aerial platforms have been considered in many events [14–17]. These images may not be as readily available as satellite data, but they can be captured at a higher resolution and such aerial platforms may also perform multi-view image captures. While the increase in the resolution aids in the disambiguation between damaged and non-damaged buildings, the oblique views enable the damage assessment of the façades [14]. These advantages were also realized by the EMS, which recently started signing contracts with private companies to survey regions with aerial oblique imagery after a disaster [18], as it happened in the 2016 earthquakes in central Italy.

Unmanned aerial vehicles have been used to perform a more thorough damage assessment of a given scene. The high portability and higher resolution, when compared to manned platforms, have several benefits: they allow for a more detailed damage assessment [17], which allows lower levels of damage such as cracks and smaller signs of spalling to be detected [19], and they allow the UAV flights to focus only on specific areas of interest [20].

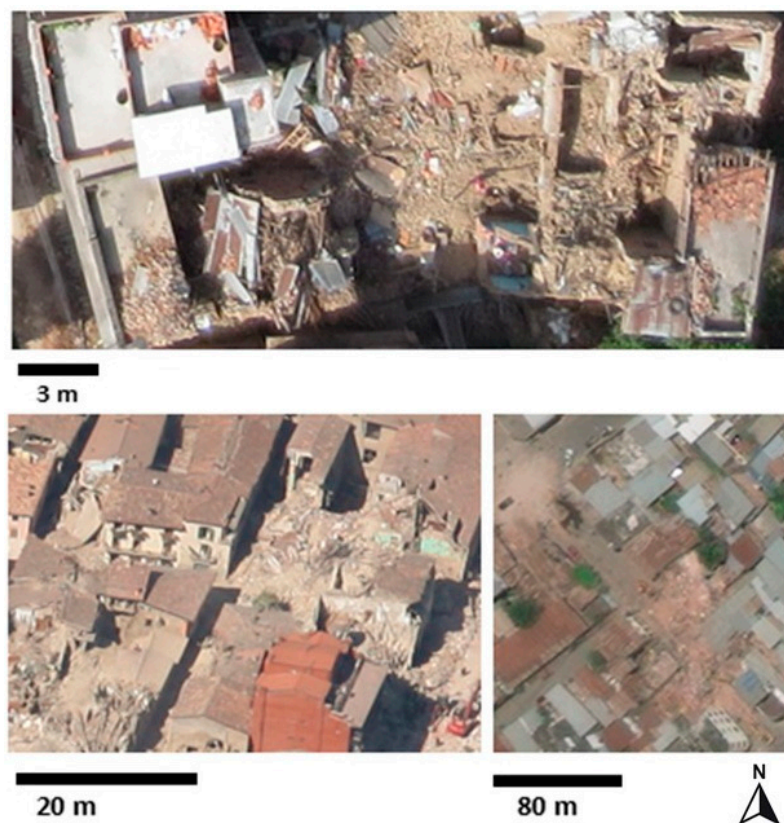
Recent advances in the computer vision domain, namely, the use of convolutional neural networks (CNN) for image classification and segmentation [21–23], have also shown their potential in the remote sensing domain [24–26] and, more specifically, for the image classification of building damages such as debris or rubble piles [17,27]. All these contributions use data with similar resolutions that are specifically acquired to train and test the developed networks. The use of multi-resolution data has improved the overall image classification and segmentation in many computer vision applications [24,28,29] and in remote sensing [25]. However, multi-resolution images are generated artificially when the input images are up-sampled and down-sampled at several scales and then fused to obtain a final stronger classifier. While in computer vision, the resolution of a given image is considered as another inherent difficulty in the image classification task, in remote sensing, there are several resolution levels defined by the used platform and sensor, and these are usually considered independently for any image classification task.

A growing amount of image data have been collected by map producers using different sensors and with different resolutions, and their optimal use and integration would, therefore, represent an opportunity to positively impact scene classification. More specifically, a successful multi-resolution approach would make the image classification of building damages more flexible and not rely only on a given set of images from a given platform or sensor. This would be optimal since there often are not enough image samples of a given resolution level available to generate a strong CNN based classifier. The first preliminary attempt in this direction, using image data from different platforms and optical sensors, has only been addressed recently [30]. This work focused on the satellite image

classification of building damages (debris and rubble piles) whilst also considering image data from other (aerial) resolutions in its training. The authors reported an improvement of nearly 4% in the satellite image classification of building damages by fusing the feature maps obtained from satellite and aerial resolutions. However, the paper limited its investigation to satellite images, not considering the impact of the multi-resolution approach in the case of aerial (manned and unmanned) images.

The present paper extends the previously reported work by thoroughly assessing the combined use of satellite and airborne (manned and unmanned) imagery for the image classification of the building damages (debris and rubble piles, as in Figure 1) of these same resolutions. This work focuses on the fusion of the feature maps coming from each of the resolutions. Specifically, the aim of the paper is twofold:

- Assess the behavior of several feature fusion approaches by considering satellite and airborne (manned and unmanned) (Figure 1) feature information, and compare them against two baseline experiments for the image classification of building damages;
- Assess the impact of multi-resolution fusion approaches in the model transferability for each of the considered resolution levels, where an image dataset from a different geographical region is only considered in the validation step.



**Figure 1.** Examples of damaged and undamaged regions in remote sensing imagery. Nepal (**top**), aerial (unmanned). Italy (**bottom left**), aerial (manned). Ecuador (**bottom right**), satellite. These image examples also contain the type of damaged considered in this study: debris and rubble piles.

The next section focuses on the related work of both image-based damage mapping and CNN feature map fusion. Section 3 presents the methodology followed to assess the use of multi-resolution imagery, where the used network is defined and the fusion approaches formalized. Section 4 deals with the experiments and results, followed by a discussion of the results (Section 5) and conclusions (Section 6).

## 2. Related Work

### 2.1. Image-Based Damage Mapping

Various methods have been reported for the automatic image classification of building damages. These aim to relate the features extracted from the imagery with damage evidences. Such methods are usually closely related to the platform used for their acquisition, exploiting their intrinsic characteristics such as the viewing angle and resolution, among others. Regarding satellite imagery, texture features have been mostly used to map collapsed and partially collapsed buildings due to the coarse resolution and limited viewing angle of the platform. Features derived from the co-occurrence matrix have enabled the detection of partial and totally collapsed buildings from the IKONOS and QuickBird imagery [6]. Multi-spectral image data from QuickBird, along with spatial relations formulated through a morphological scale-space approach have also been used to detect damaged buildings [31,32]. Another approach separated the satellite images into several classes; bricks and roof tiles were among them [33]. The authors assumed that areas classified as bricks are most likely damaged areas.

The improvement of the image sensors coupled with the aerial platforms have not only increased the amount of detail present in aerial images but have also increased the complexity of the automation of damage detection procedures [34]. Due to the high-resolution of the aerial imagery, object-based image analysis (OBIA) has started to be used to map damage [35–37] since objects in the scene are composed of a higher number of pixels. Instead of using the pixels directly, these approaches worked on the object level of an image composed of a set of pixels. In this way, the texture features were related not to a given pixel but to a set of pixels [38]. Specifically, OBIA was used, among other techniques, to assess façades for damage [14,19].

Overlapping aerial images can be used to generate 3D models through dense image matching, where 3D information can then be used to detect partial and totally collapsed buildings [14]. Additionally, the use of fitted planes allows us to assess the geometrical homogeneity of such features and distinguish intact roofs from rubble piles. The 3D point cloud also allows for the direct extraction of the geometric deformations of building elements [19], for the extraction of 3D features such as the histogram of the Z component of a normal vector [17], and for the use of the aforementioned features alongside the CNN image features in a multiple-kernel learning approach [14,17].

Videos recorded from aerial platforms can also be used to map damage. Features such as hue, saturation, brightness, edge intensity, predominant direction, variance, statistical features from the co-occurrence matrix, and 3D features have been derived from such video frames to distinguish damaged from non-damaged areas [39–41].

Focusing on the learning approach from the texture features to build a robust classifier, Vetrivel et al. [42] used a bag-of-words approach and assumed that the damage evidence related to debris, spalling, and rubble piles shared the same local image features. The popularity of the CNN for image recognition tasks has successfully led to approaches that consider such networks for the image classification of building damage (satellite and aerial) [27,30,42].

Despite the recent advancements in computer vision, particularly in CNN, these works normally follow the traditional approach of having a completely separate CNN for each of the resolution levels for the image classification of building damages from remote sensing imagery [17,27]. In this work, the use of a multi-resolution feature fusion approach is assessed.

### 2.2. CNN Feature Fusion Approaches in Remote Sensing

The increase in the amount of remote sensing data collected, be it from space, aerial, or terrestrial platforms, has allowed for the development of new methodologies which take advantage of the fusion of the different types of remote sensing data [43]. The combination of several streams of data in CNN architectures has also shown to improve the classification and segmentation results since each of the data modalities (3D, multi-spectral, RGB) contribute differently towards the recognition of a given



object in the scene [43,44]. While the presented overview focusses on CNN feature fusion approaches, there are also other approaches which do not rely on CNNs to perform data fusion [45–47].

The fusion of 3D data from laser sensors or generated through dense image matching using images has been already addressed [17,44,48,49]. Liu et al. [50] extracted handcrafted features from Lidar data alongside CNN features from the aerial images, fusing them in a higher order conditional random fields approach. Merging optical and Lidar data improved the semantic segmentation of 3D point clouds [49] using a set of convolutions to merge both feature sets. The fusion of Lidar and multispectral imagery was also addressed [48], in which the authors report the complementarity of such data in semantic segmentation. CNN and hand-crafted image features were concatenated to generate a stronger segmentation network in the case of aerial images [44]. In the damage mapping domain, Vetrivel et al. [17] merged both the CNN and 3D features (derived from a dense image-matching point cloud) in a multiple-kernel-learning approach for the image classification of building damages using airborne (manned and unmanned vehicles) images. The most relevant finding in this work was that the CNN features were so meaningful that, in some cases, the combined use of 3D information with CNN features only degraded the result, when compared to using only CNN features. The authors also found that CNNs still cannot optimally deal with the model geographical transferability in the specific case of the image classification of building damages because differences in urban morphology, architectural design, image capture settings, among others, may hinder this transferability.

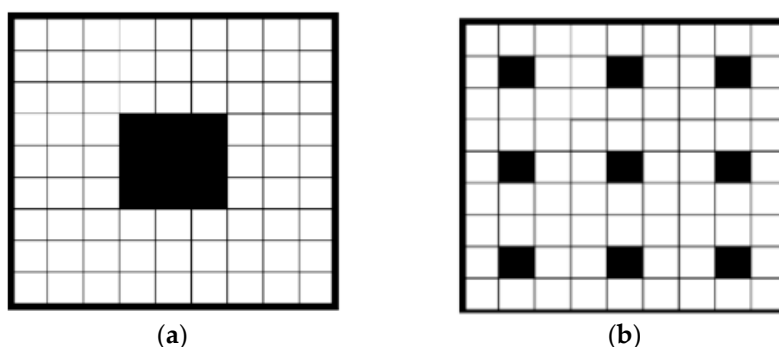
The fusion of multi-resolution imagery coming from different resolution levels, such as satellite and airborne (manned and unmanned) imagery, had already been tested only for the specific case of satellite image classification of building damages [30]. The authors reported that it is more meaningful to perform a fusion of the feature maps coming from each of the resolutions than to have all the multi-resolution imagery share features in a single CNN. Nonetheless, the multi-resolution feature fusion approach was (1) not tested for the airborne (manned and unmanned) resolution levels and (2) not tested for the model transferability when a new region was only considered in the validation step.

### 3. Methodology

Three different CNN feature fusion approaches were used to assess the multi-resolution capabilities of CNN in performing the image classification of building damages. These multi-resolution experiments were compared with two baseline approaches. These baselines followed the traditional image classification pipeline using CNN, where each imagery resolution level was fed to a single network.

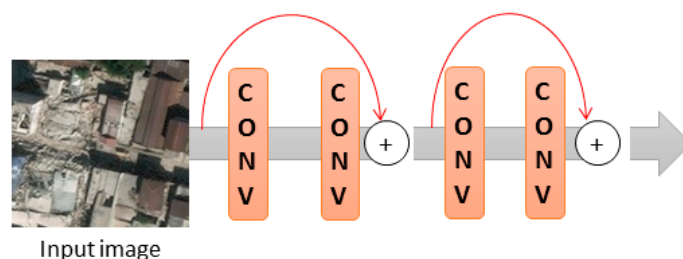
The network used in the experiments is presented in Section 3.1. This network exploited two main characteristics: residual connections and dilated convolutions (presented in the following paragraphs). The baseline experiments are presented in Section 3.2, while the feature fusion approaches are presented in Section 3.3.

A central aspect of a network capable of capturing multi-resolution information present in the images is its ability to capture spatial context. Yu and Koltun [51] introduced the concept of dilated convolutions in CNN with the aim of capturing the context in image recognition tasks. Dilated convolutions are applied to a given input image using a kernel with defined gaps (Figure 2). Due to the gaps, the receptive field of the network is bigger, capturing more contextual information [51]. Moreover, the receptive field size of the dilated convolutions also enables the capture of finer details since there is no need to perform an aggressive down-sampling of the feature maps throughout the network, better preserving the original spatial resolution [52]. Looking at the specific task of building damage detection, the visual depiction of a collapsed building in a nadir aerial image patch may not appear in the form of a single rubble pile. Often, only smaller damage cues such as blown out debris or smaller portions of rubble are found in the vicinity of such collapsed buildings. Hence, by using dilated convolutions in this study, we aim to learn the relationship between damaged areas and their context, relating these among all the levels of resolution.



**Figure 2.** The scheme of (a) a  $3 \times 3$  kernel with dilation 1, (b) a  $3 \times 3$  kernel with dilation 3 [30].

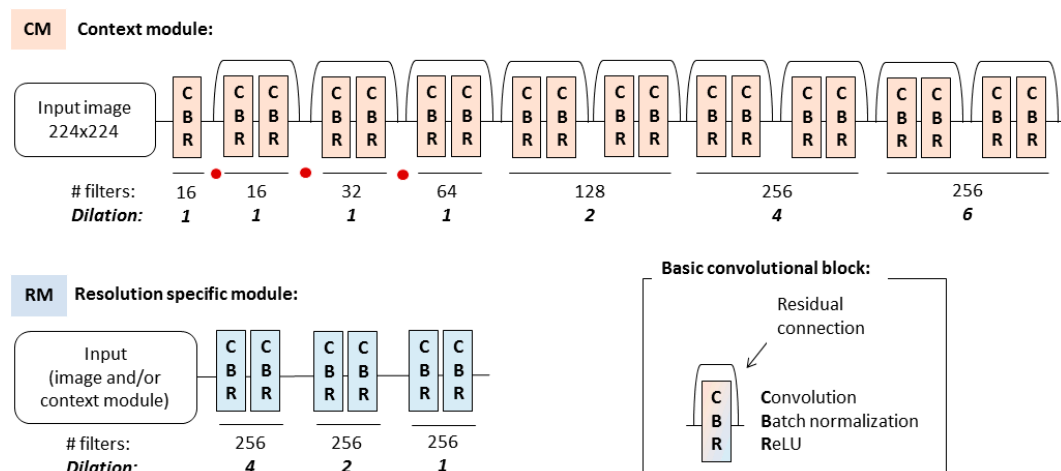
From the shallow *alexnet* [22], to the *VGG* [23], and the more recently proposed *resnet* [21], the depth of the proposed networks for image classification has increased. Unfortunately, the deeper the network, the harder it is to train [23]. CNNs are usually built by the stacking of convolution layers, which allows a given network to learn from lower level features to higher levels of abstraction in a hierarchical setting. Nonetheless, a given layer  $l$  is only connected with the layers adjacent to it (i.e., layers  $l-1$  and  $l+1$ ). This assumption has shown to be not optimal since the information from earlier layers may be lost during backpropagation [21]. Residual connections were then proposed [21], where the input of a given layer may be a summation of previous layers. These residual connections allow us to (1) have deeper networks while maintaining a low number of parameters and (2) to preserve the feature information across all layers (Figure 3) [21]. The latter aspect is particularly important for a multi-resolution approach since a given feature may have a different degree of relevance for each of the considered levels of resolution. The preservation of this feature information is therefore critical when aggregating the feature maps generated using different resolution data.



**Figure 3.** The scheme of a possible residual connection in a CNN. The grey arrows indicate a classical approach, while the red arrows on top show the new added residual connection [30].

### 3.1. Basic Convolutional Set and Modules Definition

The main network configuration was built by considering two main modules: (1) the context module and (2) the resolution-specific module (Figure 4). This structure was inspired by the works of References [21,52,53]. The general idea regarding the use of these two modules was that while the dilated convolutions capture the wider context (context module), more local features may be lost in the dilation process, hence the use of the resolution-specific module [51,53] with the decreasing dilation. In this way, the context is harnessed through the context module, while the resolution-specific module brings back the feature information related to a given resolution. The modules were built by stacking basic convolutional sets that were defined by convolution, batch normalization, and ReLU (rectified linear unit) (called CBR in Figure 4) [54]. As depicted in Figure 4, a pair of these basic convolutional sets bridged by a residual connection formed the simplest component of the network, which were then used to build the indicated modules.



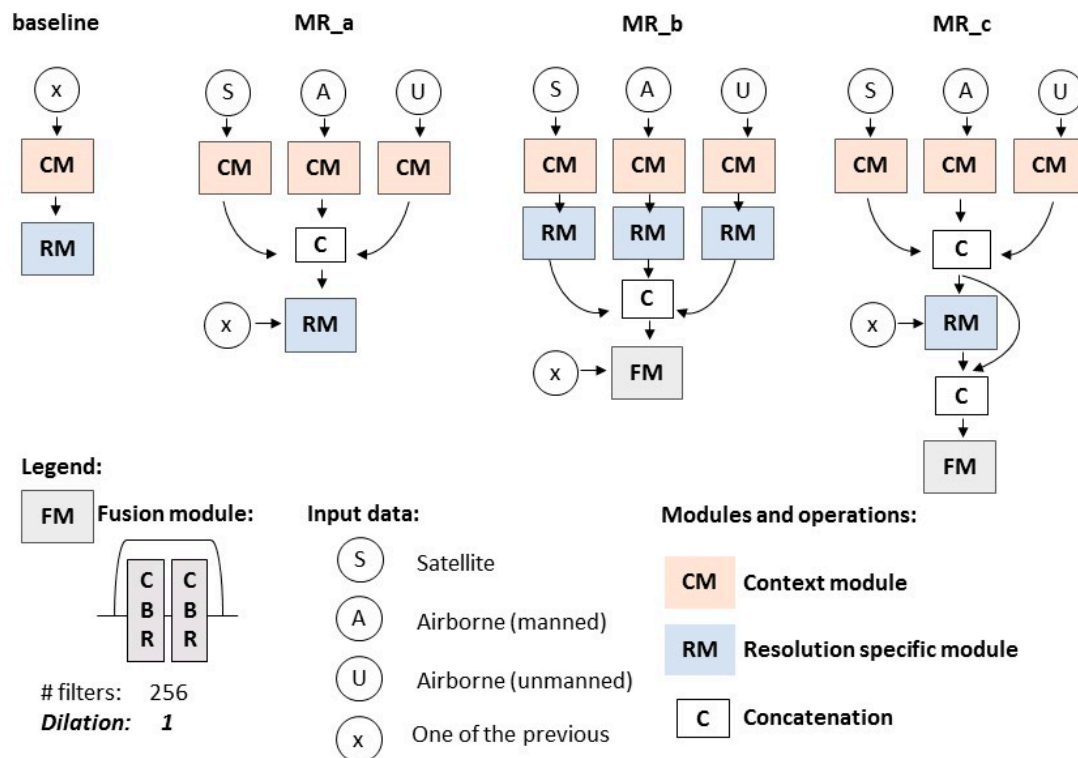
**Figure 4.** The basic convolution block is defined by convolution, batch-normalization, and ReLU (CBR). The CBR is used to define both the context and resolution-specific modules. It contains the number of filters used at each level of the modules and also the dilation factor. The red dot in the context module indicates when a striding of 2, instead of 1 was used.

The context module was built by stacking 19 CBRs with an increasing number of filters and a dilation factor. For our tests, a lower number of CBRs would make the network weaker while deeper networks would give no improvements and slow the network runtime (increasing the risk of overfitting). The growing number of filters is commonly used in CNN approaches, following the general assumption that more filters are needed to represent more complex features [21–23]. The increasing dilation factor in the context module is aimed at gradually capturing feature representations over a larger context area [51]. The red dots in Figure 4 indicate when a striding of 2, instead of 1, was applied. The striding reduced the size of the feature map (from the initial  $224 \times 224$  px to the final  $28 \times 28$  px) without performing max pooling. Larger striding has been shown to be beneficial when dilated convolutions are considered [52]. The kernel size was  $3 \times 3$  [55] and only the first CBR block of the context module had a kernel size of  $7 \times 7$  [52]. The increase in the dilation factor can generate artifacts (aliasing effect) on the resulting feature maps due to the gaps introduced by the dilated kernels [52,53]. To attenuate this drawback, the dilation increase in the context module was compensated in the resolution-specific module with a gradual reduction of the dilation value [53] and the removal of the residual connections from the basic CBR blocks [52]. This also allowed us to recapture the more local features [53], which might have been lost due to the increasing dilations in the context module. For the classification part of the network, global average pooling followed by a convolution which maps the feature map size to the number of classes was applied [52,56]. Since this was a binary classification problem, a sigmoid function was used as the activation.

### 3.2. Baseline Method

As already mentioned, the multi-resolution tests were compared against two baseline networks. These followed the traditional pipelines for the image classification of building damages [17,27]. In the first baseline network (Figure 5), the training samples of a single resolution (i.e., only airborne—manned or unmanned—or satellite) were fed into a network composed of the context and the resolution-specific module like in a single resolution approach. The second baseline (hereafter referred to as baseline\_ft) used the same architecture as defined for the baseline (Figure 5). It fed generic image samples of a given level of resolution (Tables 2 and 3) into the context module, while the resolution-specific one was only fed with the damage domain image samples of that same level of resolution. Fine-tuning a network that used a generic image dataset for training may improve the image classification process [25], especially in cases with a low number of image samples for the specific classification problem [57]. The generic resolution-specific image samples were used to train a network considering

two classes: built and non-built environments. Its weights were used as a starting point in the fine-tuning experiments for the specific case of the image classification of building damages. This led to two baseline tests for each resolution level (one trained from scratch and one fine-tuned on generic resolution-specific image samples).



**Figure 5.** The baseline and multi-resolution feature fusion approaches (MR\_a, MR\_b, and MR\_c). The fusion module is also defined.

### 3.3. Feature Fusion Methods

The multi-resolution feature fusion approaches used different combinations of the baseline modules and their computed features (Section 3.2). Three different approaches have been defined: MR\_a, MR\_b, and MR\_c, as shown in Figure 5. The three types of fusion were inspired by previous studies in computer vision [58] and remote sensing [30,43,48,49]. In the presented implementation, the baselines were independently computed for each level of resolution without sharing the weights among them [49]. The used image samples have different resolutions and they were acquired in different locations: the multi-modal approaches (e.g., [48]), dealing with heterogeneous data fusions (synchronized and in overlap), could not be directly adopted in this case as there was no correspondence between the areas captured by the different sensors. Moreover, in a disaster scenario, time is critical. Acquisitions with three different sensors (mounted on three different platforms) and resolutions would not be easily doable.

A fusion module (presented in Figure 5) was used in two of the fusion strategies, MR\_b and MR\_c, while MR\_a followed the fusion approach used in Reference [30]. This fusion module aimed to learn from all the different feature representations, blending their heterogeneity [48,58] through a set of convolutions. The objective behind the three different fusion approaches was to understand (i) which layers (and its features) were contributing more to the image classification of building damages in a certain resolution level and (ii) which was the best approach to fuse the different modules with multi-resolution information. The networks were then fine-tuned with the image data (X in Figure 5) of the resolution level of interest. For example, in MR\_a, the features from the context modules of the



three baseline networks were concatenated. Then, the resolution-specific module was fine-tuned with the image data  $X$  of a given resolution level (e.g., satellite imagery).

The concatenation indicated in Figure 5 had as input the feature maps which had the same width and height, merging them along the channel dimension. Other merging approaches were tested such as summation, addition, and the averaging of the convolutional modules, however, they underperformed when compared to concatenation. In the bullet points below, each of the fusion approaches is defined in detail. Three fusions (MR\_a, MR\_b, and MR\_c) were performed for each resolution level.

1. **MR\_a:** in this fusion approach, the features of the context modules of each of the baseline experiments were concatenated. The resolution-specific module was then fine-tuned using the image data of a given resolution level ( $X$ , in Figure 5). This approach followed a general fusion approach already used in computer vision to merge the artificial multi-scale branches of a network [28,59] or to fuse remote sensing image data [60]. Furthermore, this simple fusion approach has already been tested in another multi-resolution study [30].
2. **MR\_b:** in this fusion approach, the features of the context followed by the resolution-specific modules of the baseline experiments were concatenated. The fusion module considered as input the previous concatenation and it was fine-tuned using the image data of a given resolution level ( $X$ , in Figure 5). While only the context module of each resolution level was considered for the fusion in MR\_a, MR\_b considered the feature information of the resolution-specific module. In this case, the fusion model aimed at blending all these heterogeneous feature maps and building the final classifier for each of the resolution levels separately (Figure 5). This fusion approach allows the use of traditional (i.e., mono resolution) pre-trained networks as only the last set of convolutions need to be run (i.e., fusion module).
3. **MR\_c:** this approach builds on MR\_a. However, in this case, the feature information from the concatenation of several context modules is maintained in a later stage of the fusion approach. This was performed by further concatenating this feature information with the output of the resolution-specific module that was fine-tuned with a given resolution image data ( $X$  in Figure 5). Like MR\_b, the feature information coming from the context modules and resolution-specific module were blended using the fusion module.

## 4. Experiments and Results

The experiments, results, and used datasets are described in this section. The first set of experiments was performed to assess the classification results combining the multi-resolution data. In the second set of experiments, the model geographical transferability was assessed; i.e., when considering a new image dataset only for the validation (not used in training) of the networks.

### 4.1. Datasets and Training Samples

This subsection describes the datasets used in the experiments for each resolution level. It also describes the image sample generation from the raw images to image patches of a given resolution, which were then used in the experiments (Section 4.2). The data were divided into two main subsets: (a) a multi-resolution dataset formed by three sets of images corresponding to satellite and airborne (manned and unmanned) images containing damage image samples, and (b) three sets of generic resolution-specific image samples used in the fine-tuning baseline approach for the considered levels of resolution.

#### 4.1.1. Damage Domain Image Samples for the Three Resolution Levels Considered

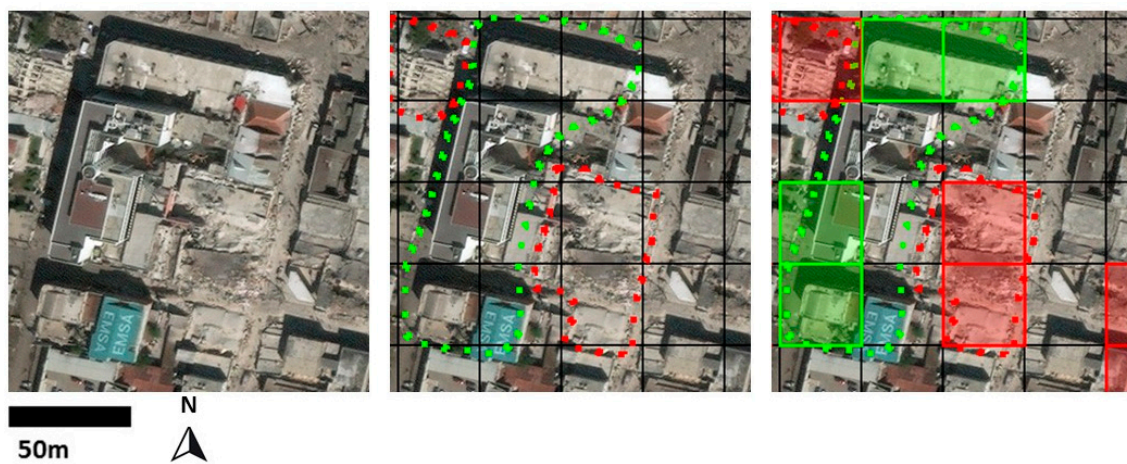
Most of the datasets depict real earthquake-induced building damages; however, there are also images of controlled demolitions (Table 1). The satellite images cover five different geographical locations in Italy, Ecuador, and Haiti. The satellite imagery was collected with WorldView-3 (Amatrice (Italy), Pescara del Tronto (Italy), and Portoviejo (Ecuador)) and GeoEye-1 (L'Aquila (Italy)),

Port-au-Prince (Haiti)). These data were pansharpened and have a variable resolution between 0.4 and 0.6 m. The airborne (manned platforms) images cover seven different geographic locations in Italy, Haiti, and New Zealand. These sets of airborne data consist of nadir and oblique views. These were captured with the PentaView capture (Pictometry) and UltraCam Osprey (Microsoft) oblique imaging systems. Due to the oblique views, the ground sampling distance varies between 8 and 18 cm. These are usually captured with similar image capture specifications (flying height, overlap, etc.). The airborne (unmanned platforms) images cover nine locations in France, Italy, Haiti, Ecuador, Nepal, Germany, and China. These are composed of both the nadir and oblique views that were captured using both fixed wing and rotary wing aircraft mounted with consumer grade cameras. The ground sampling distance ranges from <1 cm up to 12 cm, where the image capture specifications (flying height, overlap, etc.) are related to the specific objective of each of the surveys, which changes significantly between the different datasets

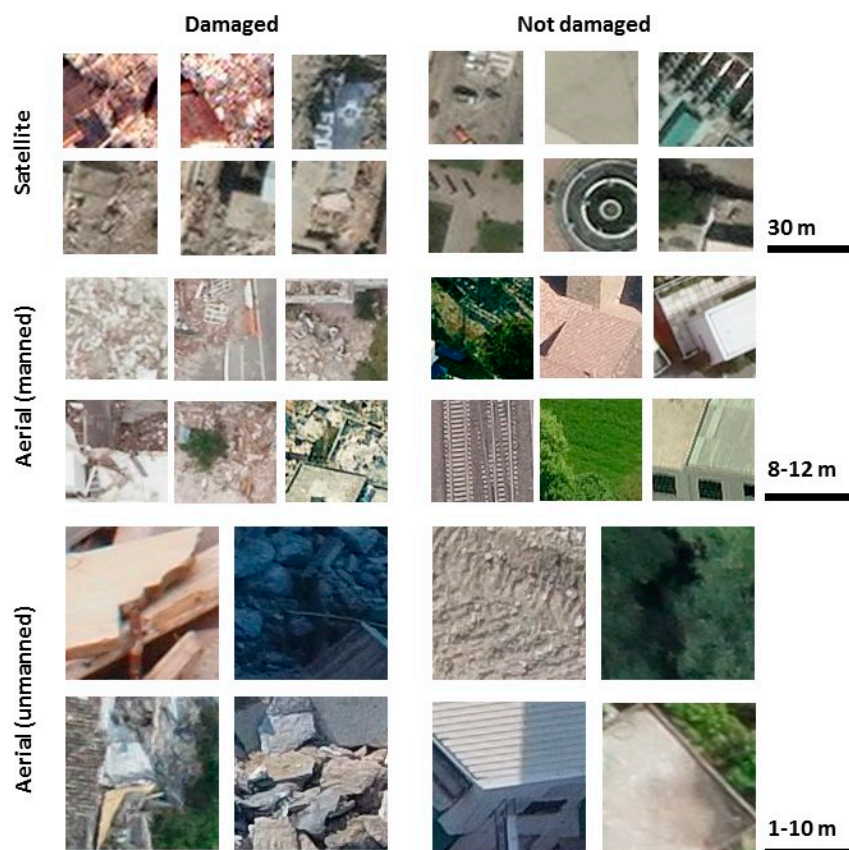
**Table 1.** An overview of the location and quantity of the satellite and airborne image samples. The ++ locations indicate the controlled demolitions of buildings.

Location	No. of Samples		Month/Year of Event	Sensor/System
	Damaged	Not Damaged		
Satellite				
L'Aquila (Italy)	115	108	April 2009	GeoEye-1
Port-au-Prince (Haiti)	701	681	January 2010	GeoEye-1
Portoviejo (Ecuador)	125	110	April 2016	WorldView-3
Amatrice (Italy)	135	159	August 2016	WorldView-3
Pesc. Tronto (Italy)	91	94	August 2016	WorldView-3
Total	1169	1152		
Airborne (manned)				
L'Aquila (Italy)	242	235	April 2009	PentaView
St Felice (Italy)	337	366	May 2012	PentaView
Amatrice (Italy)	387	262	September 2016	PentaView
Tempera (Italy)	151	260	April 2009	PentaView
Port-au-Prince (slums) (Haiti)	409	329	January 2010	PentaView
Port-au-Prince (Haiti)	302	335	January 2010	PentaView
Onna (Italy)	293	265	April 2009	PentaView
Christchurch (New Zealand)	603	649	February 2011	Vexcel UCXp
Total	2754	2701		
Airborne (unmanned)				
L'Aquila (Italy)	103	99	April 2009	Sony ILCE-6000
Wesel (Germany)	175	175	June 2016++	Canon EOS 600D
Portoviejo (Ecuador)	306	200	April 2016	DJI FC300S
Pesc. Tronto (Italy)	197	262	August 2016	Canon Powershot S110
Katmandu (Nepal)	388	288	April 2015	Canon IXUS 127 HS
Taiwan (China)	257	479	February 2016	DJI FC300S
Gronau (Germany)	437	501	October 2013++	Canon EOS 600D
Mirabello (Italy)	412	246	May 2012	Olympus E-P2
Lyon (France)	230	242	May 2017++	DJI FC330
Total	2505	2692		

The image samples were derived from the set of images indicated before. First, the damaged and undamaged image regions were manually delineated, see Figure 6. A regular grid was then applied to each of the images and every cell that contained more than 40% of its area masked by the damage class was cropped from the image and used as an image sample for the damage class. The low value of 40% to consider a patch as damaged, aimed at forcing the networks to detect damage on an image patch even if it did not occupy the majority of the area of the said patch. This is motivated by practical reasons as an image patch should be considered damaged even if just a small area contains evidence of damage. On the other hand, a patch is considered intact only if no damage can be detected (Figure 6). The grid size varied according to the resolution: satellite =  $80 \times 80$  px, airborne (manned vehicles) =  $100 \times 100$  px, and airborne (unmanned) =  $120 \times 120$  px (examples in Figure 7). The variable size of the image patches according to the resolution aimed to attenuate the captured extent by each of the resolution levels. The use of smaller patches also allowed us to increase the number of samples, compensating for the rare availability of these data.



**Figure 6.** An example of the extracted samples considering a satellite image (GeoEye-1, Port-au-Prince, Haiti, 2010) on the left. The center image contains the grid for the satellite resolution level ( $80 \times 80$  px) where the damaged (red) and non-damaged (green) areas were manually digitized. The right patch indicates which squares of the grid are considered damaged and non-damaged after the selection process.



**Figure 7.** Examples of image samples derived from the procedure illustrated in Figure 6. These were used as the input for both the baseline and multi-resolution feature fusion experiments. (**Left side**) damaged samples; (**Right side**) non-damaged samples. From top to bottom: 2 rows of satellite, aerial (manned), and aerial (unmanned) image samples. The approximate scale is indicated for each resolution level.

The number of image samples between the classes was approximately the same, while the number of image samples between the three different resolution levels was not balanced. The number of satellite image samples was two-fold lower when compared to the other two levels of resolution.

#### 4.1.2. Generic Image Samples for the Three Levels of Resolution

Generic image samples for each of the levels of resolution are presented in this sub-section. These were used in one of the baseline approaches (baseline\_ft).

The generic satellite image samples were taken from a freely available baseline dataset: NWPU-RESISC45 (Cheng et al., 2017). This baseline dataset contained 45 classes with 700 satellite image samples per class. From these, fourteen classes were selected and divided into two broader classes: built and non-built (Table 2).

**Table 2.** The 14 classes of the benchmark dataset (NWPU-RESISC45) divided into the built and non-built classes. Each class contains 700 samples, with a total of 9800 image samples.

Built	Non-Built
Airport	Beach
Commercial area	Circular farmland
Dense residential	Desert
Freeway	Forest
Industrial area	Mountain
Medium residential	Rectangular farm
Sparse residential	Terrace

To derive the generic image samples from the airborne images (manned and unmanned), the same sample extraction procedure used for the damage and non-damaged samples was adopted. In this case, the division was between the built and non-built environments, while the rest of the procedure was the same: a mask for the built and non-built environments was applied by considering a 60% threshold for each given class. This threshold was adopted to ensure that one of the two classes (the built and non-built environment classes) occupied the larger area of the image patch.

Table 3 shows the origin of the data for this generic image samples generation, the quantity of image samples and the considered camera for each location.

**Table 3.** The generic airborne image samples used in one of the baselines. The \* indicates that in the aerial (manned) case, three different locations from the Netherlands were considered.

Location	Generic Airborne (Unmanned) Image Samples			Generic Airborne (Manned) Image Samples		
	Built	Non-Built	Sensor/System	Built	Non-Built	Sensor/System
Netherlands *	971	581	Olympus E-P3	1758	878	PentaView and Vexcel Ultra-CamXP
France	697	690	Canon IXUS220 HS	1110	1953	Vexcel Ultra-Cam D
Germany	681	618	DJI FC330			
Italy	578	405	Pentax OPTIO A40			
Switzerland	107	688	Canon IXUS220 HS	2868	2831	
<b>Total</b>	3034	2982				

During the training of every network, data augmentation was performed (Table 4) since this was shown to decrease overfitting and improve the overall image classification [22,23]. The used data augmentation consisted of random translations and rotations, image normalization, and the up-/down-sampling of the images (examples in Figure 8). Since we were dealing with oblique imagery in the airborne data, the performed flips were only horizontal and both the rotation value and the scale factor were low. Furthermore, light data augmentation is usually considered when batch normalization is used in a CNN since the network should be trained by focusing on less distorted images [54].



**Table 4.** The data augmentation used: image normalization, the interval of the scale factor to be multiplied by the original size of the image sample, the rotation interval to be applied to the image samples, and the horizontal flip.

Data Augmentation	Value
Image normalization	1/255
Scale factor	[0.8,1.2]
Rotation	[−12,12] deg
Horizontal flip	true



**Figure 8.** Several random data augmentation examples from an original aerial (unmanned) image sample with the scale, left.

The image samples were zero padded to fit in the  $224 \times 224$  px input size, instead of being resized; this has been demonstrated to perform better [27] in the specific image classification of building damages using CNNs.

Two main sets of experiments were performed using the multi-resolution feature fusion approaches indicated in Figure 5: (1) general multi-resolution feature fusion experiments, where the training was performed using 70% of the image samples of each resolution and using the remaining 30% of the image samples for validation. This ratio was applied to each location separately. The training/validation data splits were performed randomly three times, enforcing the validation sets to contain different image samples on each data split; (2) model transferability, where the training of each of the multi-resolution feature fusion approaches was performed by considering all the locations except the one that was used for the validation. This experiment aimed at assessing the behavior of the approaches in a realistic scenario wherein the image data from a new event were classified without extracting any training samples from this location.

For both sets of experiments, the accuracy, recall, and precision were calculated for the validation image datasets described before and the following equations were considered:

$$accuracy = \frac{TP + FN}{\# \text{ validation samples}} \quad (1)$$

$$recall = \frac{TP}{TP + FN} \quad (2)$$

$$precision = \frac{TP}{TP + FP} \quad (3)$$

where, in Equations (1)–(3), TP are the true positives, FN are the false negatives, and FP are the false positives.

#### 4.2. Results

In this sub-section, the results of the multi-resolution fusion approaches are shown. The results are divided into two sub-sections for each of the resolution levels: the general multi-resolution fusion experiment and the model transferability experiment (using a dataset from a location not used in the training). To understand the behavior of the networks better, the activations from the last set of filters of the networks are visualized when classifying a new and unused image patch depicting a



damaged scene. These activations show the per pixel probability of a pixel being damaged (white) or not damaged (black). Furthermore, in the model transferability sub-section, larger image patches were considered and classified with the best baseline and multi-resolution feature fusion approach.

#### 4.2.1. Multi-Resolution Fusion Approaches

The achieved accuracies, recalls, and precisions for the baselines and for the different multi-resolution feature fusion approaches are presented in Table 5.

**Table 5.** The accuracy, recall, and precision results when considering the multi-resolution image data in the image classification of building damage of the given resolutions. Overall, the multi-resolution feature fusion approaches present the best results.

Network	Satellite			
	Accuracy	Recall	Precision	Training Samples
baseline	87.7 ± 0.7	88.4 ± 0.9	87.4 ± 1.0	1602
baseline_ft	84.3 ± 0.8	84.1 ± 1.2	87.5 ± 1.8	11,402
MR_a	89.2 ± 1.0	87.0 ± 1.2	<b>91.0 ± 1.3</b>	8968
MR_b	89.3 ± 0.9	91.0 ± 0.9	86.5 ± 0.6	8968
MR_c	<b>89.7 ± 0.9</b>	<b>93.1 ± 1.1</b>	82.3 ± 1.6	8968
Network	Airborne (Manned)			
	Accuracy	Recall	Precision	Training Samples
baseline	91.1 ± 0.1	92.4 ± 1.5	91.1 ± 0.4	3736
baseline_ft	90.0 ± 0.4	89.8 ± 2.4	<b>90.5 ± 0.3</b>	9752
MR_a	<b>91.4 ± 0.2</b>	<b>94.0 ± 0.6</b>	88.0 ± 0.7	8968
MR_b	90.7 ± 0.4	91.9 ± 2.2	90.0 ± 1.2	8968
MR_c	<b>91.4 ± 0.2</b>	92.4 ± 0.7	89.4 ± 1.3	8968
Network	Airborne (Unmanned)			
	Accuracy	Recall	Precision	Training Samples
baseline	94.2 ± 1.0	93.1 ± 2.6	95.0 ± 0.7	3630
baseline_ft	91.3 ± 1.0	91.8 ± 2.0	89.9 ± 2.0	9329
MR_a	94.3 ± 0.7	94.1 ± 1.9	<b>95.7 ± 1.9</b>	8968
MR_b	95.3 ± 1.2	95.2 ± 0.7	95.3 ± 1.5	8968
MR_c	<b>95.4 ± 0.6</b>	<b>95.5 ± 1.7</b>	95.1 ± 1.2	8968

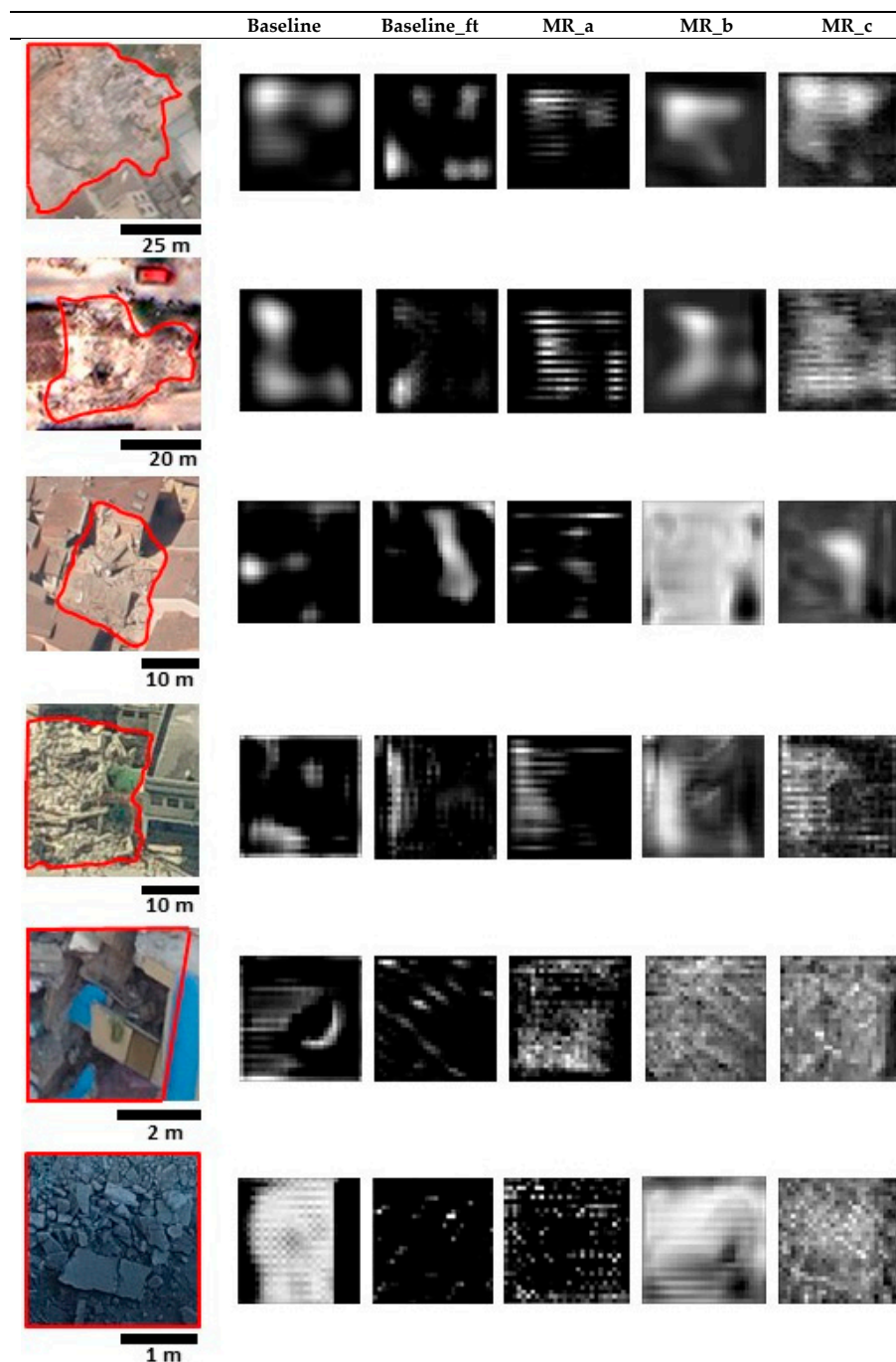
Considering the satellite resolution, the multi-resolution approaches improved the overall image classification of building damages when compared with the baselines by 2%. However, these also presented a slightly higher standard deviation between different runs. The MR\_c presents the best results even though the improvement was marginal when compared with the other multi-resolution approaches. In comparison to the baseline experiment, the recall was higher in 2 of the 3 fusion approaches, while the precision was only higher in MR\_a.

In the aerial (manned) case, the accuracy improvement was only marginal compared with the best performing baseline experiment. One of the multi-resolution approaches (MR\_b) presented the worst results compared to the baseline network. Baseline\_ft was the experiment with the weakest performance as happened in the satellite case. MR\_a had the highest recall and it also had the lower precision compared with the baseline experiment. MR\_c increased the precision of the baseline test.

The airborne (unmanned) case also presented a marginal improvement using the proposed fusion approaches (MR\_c and MR\_b). Furthermore, in MR\_c, the standard deviations of the experiments were lower. The baseline\_ft was the experiment with the weakest performance. Overall, the best performing network regarding the classification accuracy was MR\_c. This was further confirmed by the recall and precision values where all the fusion approaches had higher values for both the recall and precision than the baseline experiment.

The activations are shown in Figure 9. On the left, the input image patches are shown; on the right, the activations with the higher average activation value for each of the baseline and feature fusion approaches are shown. Overall, the multi-resolution fusion approaches presented better localization

capabilities. These usually detected larger damaged areas than the baseline experiments. Namely, MR\_c was the fusion approach with the better overall localization, even if it was noisier. The Figure 9 activations also present several striped patterns and gridding artifacts, where MR\_b seems to be the network which better attenuates this issue.



**Figure 9.** The image samples (left) and activations from the last set of feature maps (right) for each of the networks in the general multi-resolution feature fusion experiments. From top to bottom: 2 image samples of the satellite and aerial (manned and unmanned) resolutions. Overall, the multi-resolution feature fusion approaches have better localization capabilities than the baseline experiments.

#### 4.2.2. Multi-Resolution Fusion Approaches' Impact on the Model Transferability

Table 6 shows the accuracies, recalls, and precisions of the multi-resolution and baseline approaches when using a single location in the validation which was not used in the training. In the satellite case, only the image data from Portoviejo were used as its validation data. In the airborne (manned) case, the Port-au-Prince image data were used as validation while in the airborne (unmanned) case, the Lyon image data were used for validation.

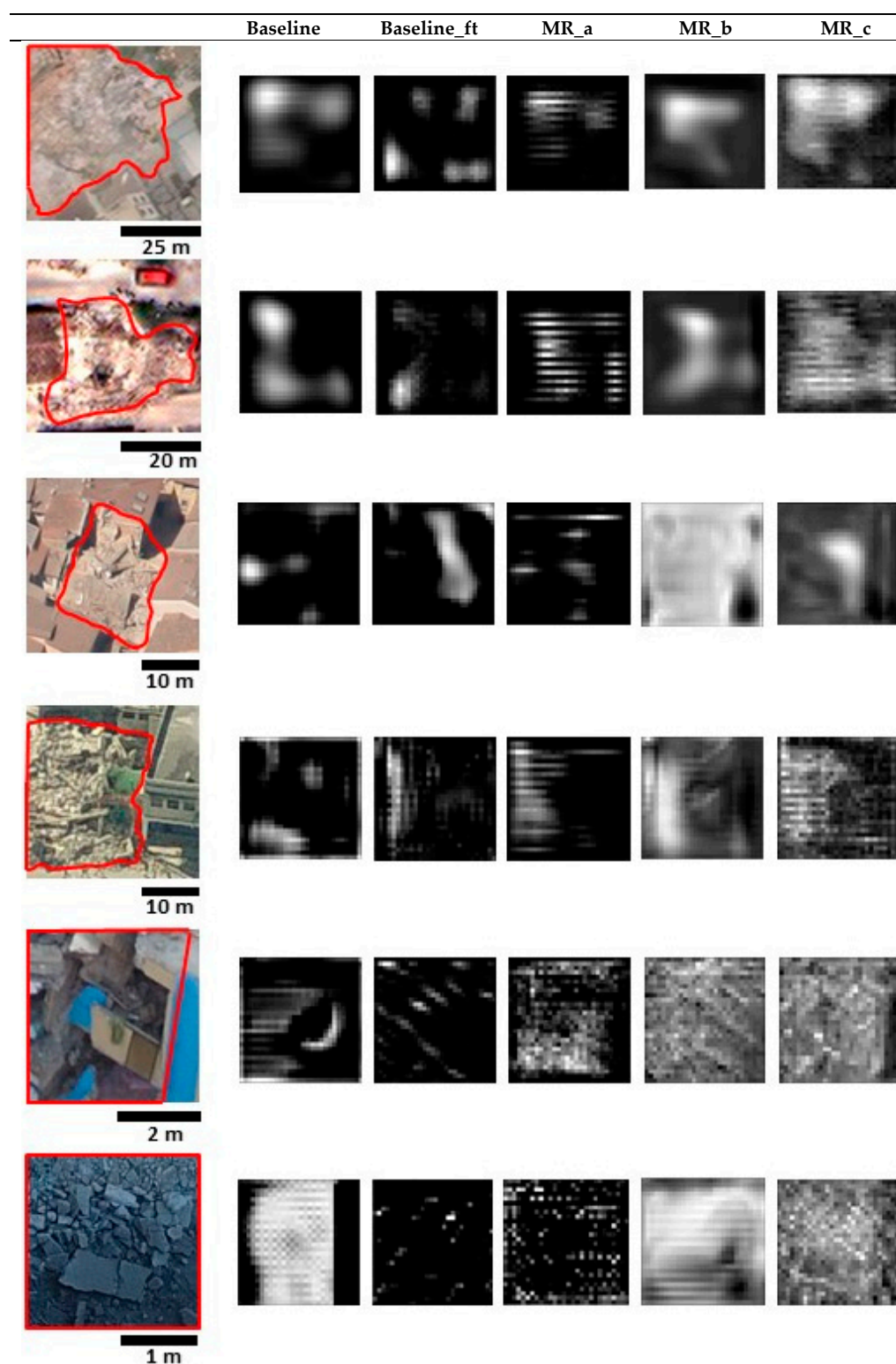
**Table 6.** The accuracy, recall and precision results when considering the multi-resolution feature fusion approaches for the model transferability. One of the locations for each of the resolutions is only used in the validation of the network: satellite = Portoviejo; aerial (manned) = Haiti; aerial (unmanned) = Lyon. Overall, the multi-resolution feature fusion approaches outperform the baseline experiments, where the baseline\_ft present better results only in the aerial (manned) case.

Network	Satellite (Portoviejo)			
	Accuracy (%)	Recall (%)	Precision (%)	Training Samples
baseline	81.5	84	78	2160
baseline_ft	79.4	76	85	11,960
MR_a	81.5 $\pm$ 0.9	83.5 $\pm$ 0.1	83.5 $\pm$ 1.7	9526
MR_b	82.1 $\pm$ 0.6	77.7 $\pm$ 0.8	<b>90.5 <math>\pm</math> 1.5</b>	9526
MR_c	<b>83.4 <math>\pm</math> 0.4</b>	<b>86.5 <math>\pm</math> 0.9</b>	82.9 $\pm$ 0.6	9526
Network	Aerial (Manned, Port-au-Prince)			
	Accuracy (%)	Recall (%)	Precision (%)	Training Samples
baseline	84.3	80.2	83.4	4406
baseline_ft	<b>84.7</b>	83.2	85.1	10,442
MR_a	81.9 $\pm$ 0.4	85.0 $\pm$ 0.3	78.6 $\pm$ 2.0	9638
MR_b	83.9 $\pm$ 0.4	80.3 $\pm$ 0.9	<b>84.1 <math>\pm</math> 2.1</b>	9638
MR_c	84.2 $\pm$ 0.2	<b>85.0 <math>\pm</math> 0.5</b>	80.0 $\pm$ 1.4	9638
Network	Aerial (Unmanned, Lyon)			
	Accuracy (%)	Recall (%)	Precision (%)	Training Samples
baseline	87.2	79.5	<b>95.1</b>	4711
baseline_ft	83.0	70.0	94.6	10,442
MR_a	85.7 $\pm$ 3.2	85.2 $\pm$ 3.6	90.0 $\pm$ 3.4	9943
MR_b	83.6 $\pm$ 2.1	86.2 $\pm$ 1.4	83.2 $\pm$ 3.3	9943
MR_c	<b>88.7 <math>\pm</math> 1.7</b>	<b>89.6 <math>\pm</math> 2.0</b>	82.4 $\pm$ 3.3	9943

Overall, the results followed the tendency of the previous experiments. The multi-resolution fusion approaches were the networks that performed better. Only in the aerial (manned) case was the baseline\_ft accuracy superior to that of the multi-resolution experiments. In the rest of the experiments, the baseline networks performed the worst.

In the airborne (unmanned) experiments, while the accuracy also increased with the MR\_c feature fusion approach, the standard deviation was also considerably higher when compared with the rest of the experiments. Overall, the recall was higher in the fusion approaches, while the precision was lower when compared to the baseline experiments.

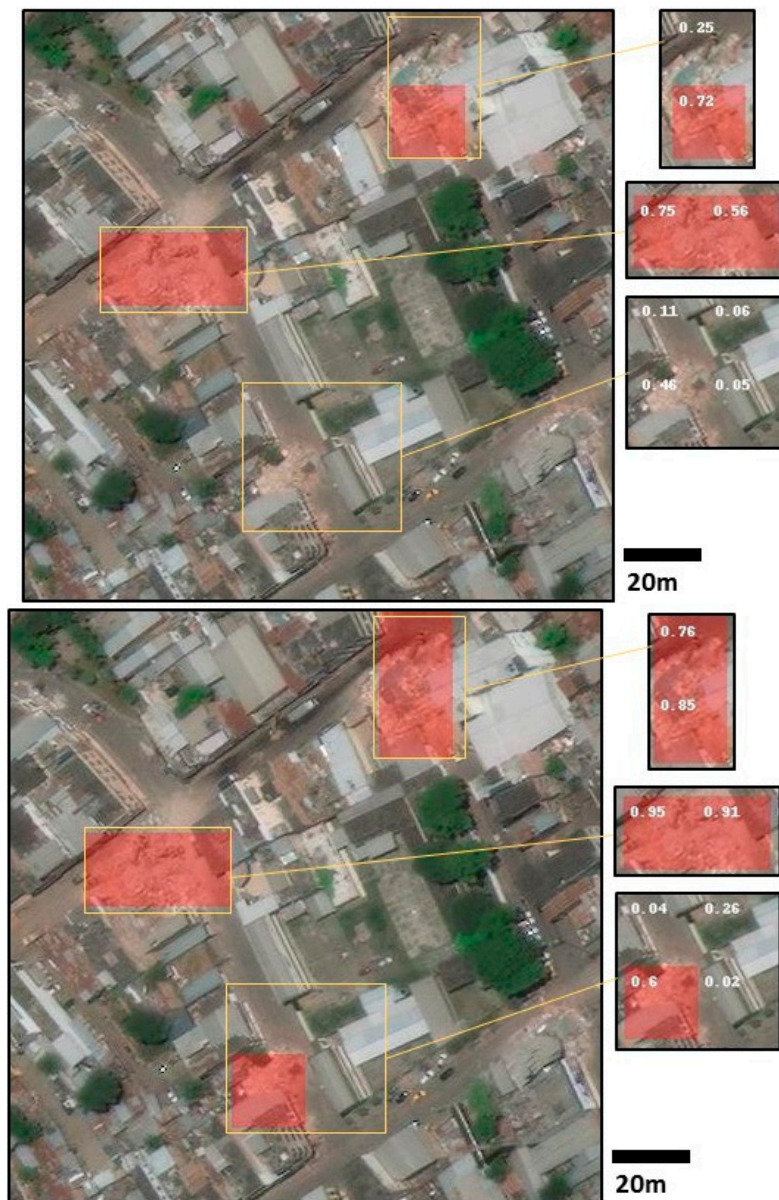
The activations are shown in Figure 10. On the left, the input image patches are shown; on the right, the activations with the highest average activation value per network are shown. Overall, the activations of the model transferability test presented the worst results when compared to the previous set of experiments. Striped patterns and gridding artifacts can also be noticed in this case. MR\_b was the network which presented a lower amount of artifacts compared to the rest of the experiments. In the aerial (unmanned) case, the localization capability decreased drastically. Nonetheless, the multi-resolution experiments, in general, could better localize the damaged area.



**Figure 10.** The image samples (*left*) and activations from the last set of the feature maps (*right*) for each of the networks in the model transferability experiments. From top to bottom: the 2 image samples of the satellite and aerial (manned and unmanned) resolutions. Overall, the multi-resolution feature fusion approaches have better localization capabilities than the baseline experiments.



In Figures 11–13, larger image patches are shown for each of the locations considered for model transferability. These image patches were divided into smaller regions ( $80 \times 80$  px for the satellite,  $100 \times 100$  px for the aerial manned, and  $120 \times 120$  px for the aerial unmanned) and classified using the best performing baseline and multi-resolution feature fusion approaches (Table 6). The red overlay in these larger image patches indicates when a patch was classified as damaged (with a  $>0.5$  probability of being damaged). The details (on the right) of these figures indicate the areas where differences between the baseline and the multi-resolution feature fusion methods were more significant. In these details, the probability of each of the smaller image patches being damaged is indicated.



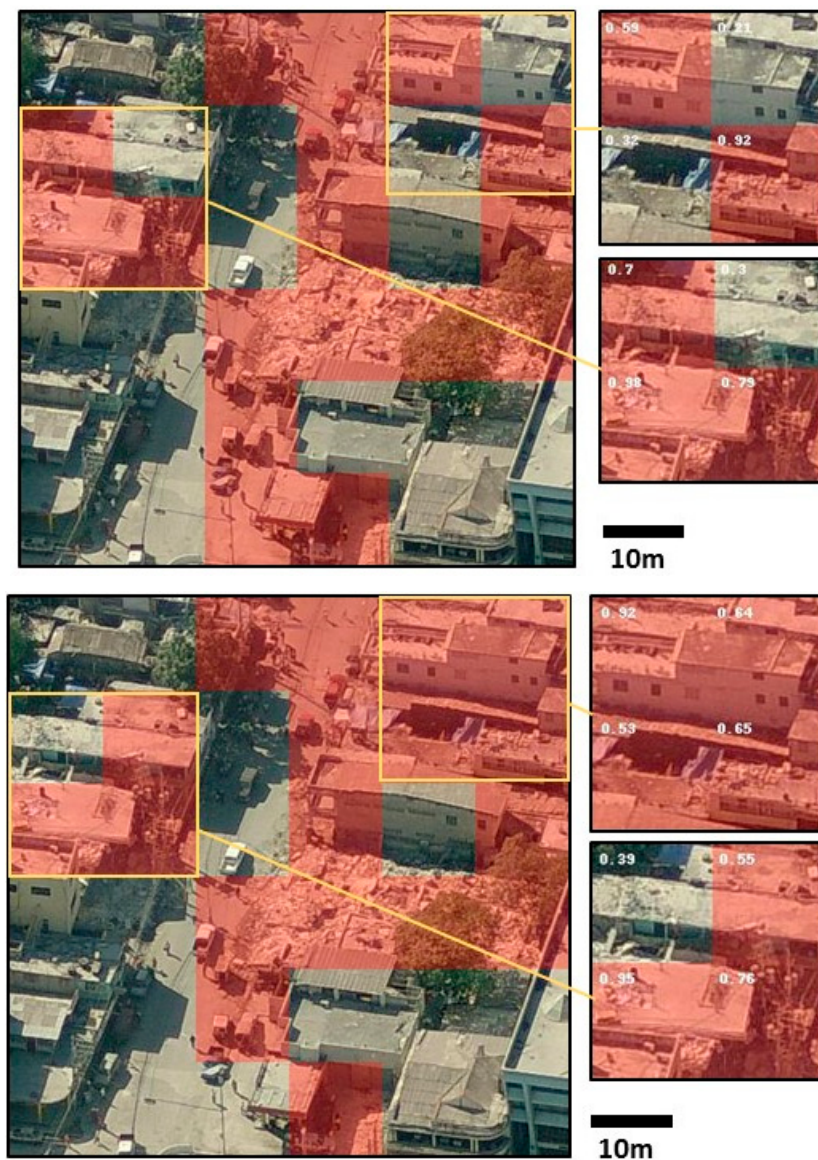
**Figure 11.** The large satellite image patch classified for damage using (top) the baseline and (bottom) the MR\_c models on the Portoviejo dataset. The red overlay shows the image patches ( $80 \times 80$  px) considered as damaged (the probability of being damaged =  $>0.5$ ). The right part with the details contains the probability of a given patch being damaged. The scale is relative to the large image patch on the left.

Figure 11 contains the image patch considered for the satellite level of resolution (Portoviejo). Besides correctly classifying 2 more patches as damaged, MR\_c also increased the certainty of the



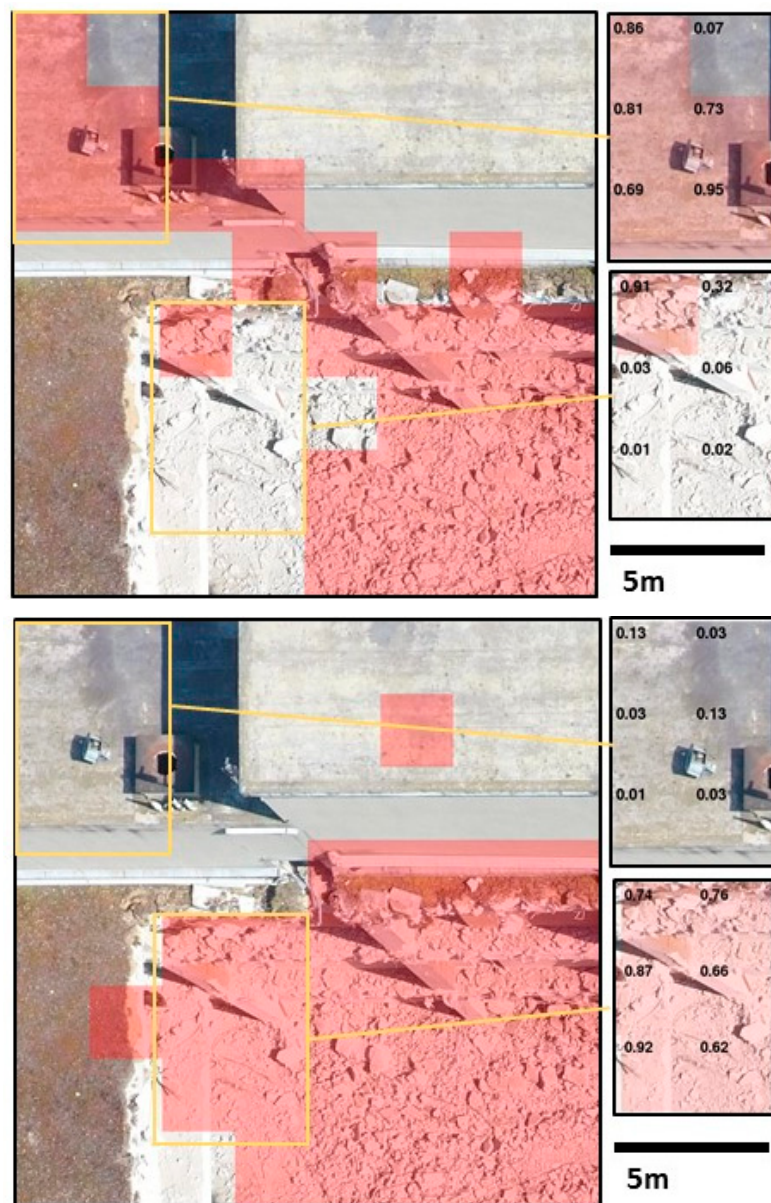
already correctly classified patches in the baseline experiments. Nonetheless, none of the approaches was able to correctly classify the patch on the lower right corner of the larger image patch as damage.

Figure 12 shows a larger image patch for the aerial (manned) case (Port-au-Prince). The best performing networks were the baseline\_ft and MR\_c networks and the classification results are shown in the figure. In general, the results followed the accuracy assessment presented in Table 6. In this case, MR\_c introduced more false positives (the details are on the right and on the bottom of the patch), even if it correctly classified more damaged patches.



**Figure 12.** The large aerial (manned) image patch classified for damage using the (top) baseline\_ft and (bottom) the MR\_c models on the Port-au-Prince dataset. The red overlay shows the image patches ( $100 \times 100$  px) considered as damaged (the probability of being damaged =  $>0.5$ ). The right part with the details contains the probability of a given patch being damaged. The legend is relative to the large image patch on the left.

Figure 13 shows a larger patch of the Lyon dataset classified with the benchmark and MR\_c networks. In this case, the MR\_c is clearly more generalizable. It reduced the false positives of the baseline approach and correctly classified the patches that were not considered damaged by the baseline.



**Figure 13.** The large aerial (unmanned) image patch classified using (top) the baseline and (bottom) the MR\_c models on the Lyon dataset. The red overlay shows the image patches ( $120 \times 120$  px) considered as damaged (the probability of being damaged =  $>0.5$ ). The right part of the figure shows the probability of each patch being damaged. The scale is relative to the large image patch on the left.

## 5. Discussion

The results show an improvement in the classification accuracy and the localization capabilities of a CNN for the image classification of building damages using the multi-resolution feature maps. However, each of the different feature fusion approaches behaved differently. The overall best multi-resolution feature fusion approach (MR\_c) concatenates the feature maps from intermediate layers, confirming the need for preserving feature information from the intermediate layers at a later stage of the network [25,28]. This feature fusion approach also considers a fusion module (Figure 5) that is able to merge and blend the multi-resolution feature maps. Other feature fusion studies using small convolutional sets to merge audio and video features [58] or remote sensing multi-modal feature maps [44,48,50] have underlined the same aspect. In general, the satellite and aerial (unmanned) resolutions were the ones which presented the most improvements when using multi-resolution

feature fusion approaches. The aerial (unmanned) resolution also improved their image classification accuracy and localization capabilities (although marginally). In the aerial (manned) case, the resolution level had the least improvement with the multi-resolution feature fusion approach. This will be discussed in detail below.

The model transferability experiments generally had a lower accuracy, indicating the need for in situ image acquisitions to get optimal classifiers, as shown in [17]. In the satellite case, both the precision and recall were higher in the multi-resolution feature fusion approaches, and the models captured fewer false positives and fewer false negatives. In the aerial (manned and unmanned) cases, the recall was higher and the precision was lower, reflecting that a higher number of image patches were correctly classified as damaged but more false positives were also present. In the aerial (manned) resolution tests, the multi-resolution feature fusion approaches had worse accuracies than the baselines. In this case, the best approach was to fine-tune a network which used generic aerial (manned) image samples during the training. In the aerial (manned) case, the image quality was better (high-end calibrated cameras), with more homogenous captures throughout different geographical regions. The aerial (unmanned) platform image captures were usually performed with a wide variety of compact grade cameras which presented a higher variability both in the sensor characteristics and in their image capture specifications. Consequently, there was a variable image quality compared to the aerial (manned) platforms.

The transferability tests of aerial (unmanned) imagery, contemporarily deal with geographical transferability aspects and also with very different image quality and image capture specifications. In such cases, the presented results indicate that the multi-resolution feature fusion approaches helped the model to be more generalizable than when using traditional mono-resolution methods.

The activations shown in the results are in agreement with the accuracy results. The multi-resolution feature fusion approaches presented better localization capabilities compared with the baseline experiments. Strike patterns and gridding artifacts can be seen in the activations. This could be due to the use of a dilated kernel in the presented convolutional modules, as indicated in [52,53].

The large image patches shown in Figures 11–13 show that both the satellite and aerial (unmanned) resolution levels can benefit more from the multi-resolution feature fusion approach in comparison to the baseline experiments. Furthermore, the aerial (unmanned) multi-resolution feature fusion identifies only one of the patches as a false positive, while correctly classifying more damaged image patches.

The previous study on multi-resolution feature fusion [30], using both a baseline and a feature fusion approach similar to the MR\_a, had better accuracies than the ones presented in this paper, although both contributions reflect a general improvement. The differences in the two works is in the training data that were extracted from the same dataset but considering different images and different damage thresholds for the image patches labelling (40% in this paper, 60% in Reference [30]). The different results confirm the difficulties and subjectivity inherent in the manual identification of building damages from any type of remote sensing imagery [12,58]. Moreover, it also indicates the sensibility of the damage detection with CNN according to the input used for training.

## 6. Conclusions and Future Work

This paper assessed the combined use of multi-resolution remote sensing imagery coming from sensors mounted on different platforms within a CNN feature fusion approach to perform the image classification of building damages (rubble piles and debris). Both a context and a resolution-specific network module were defined by using dilated convolutions and residual connections. Subsequently, the feature information of these modules was fused using three different approaches. These were further compared against two baseline experiments.

Overall, the multi-resolution feature fusion approaches outperformed the traditional image classification of building damages, especially in the satellite and aerial (unmanned) cases. Two relevant aspects have been highlighted by the performed experiments on the multi-resolution feature fusion



approaches: (1) the importance of the fusion module, as it allowed both MR\_b and MR\_c to outperform MR\_a (2) the beneficial effect of considering the feature information from the intermediate layers of each of the resolution levels in the later stages of the network, as in MR\_c.

These results were also confirmed in the classification of larger image patches in the satellite and aerial (unmanned) cases. Gridding artifacts and stripe patterns could be seen in the activations of the several fusion and baseline experiments due to the use of dilated kernels, however, in the multi-resolution feature fusion experiments, the activations were often more detailed than in the traditional approaches.

The model transferability experiments in the multi-resolution feature fusion approaches also improved the accuracy of the satellite and aerial (unmanned) imagery. On the contrary, fine-tuning a network by training it with generic aerial (manned) images was preferable in the aerial (manned) case. The different behavior in the aerial (manned) case could be explained by the use of images captured with high-end calibrated cameras and with more homogenous data capture settings. The characteristics of the aerial (manned) resolution level contrasted with the aerial (unmanned) case, where the acquisition settings were more heterogeneous and a number of different sensors with a generally lower quality were used. In the aerial (manned) case, the model transferability to a new geographical region was, therefore, more related with the scene characteristics of that same region (e.g., urban morphology) and less related with the sensor or capture settings. In the aerial (unmanned) case, the higher variability of the image datasets allowed to better generalize the model.

The transferability test also indicated that the highest improvements of the multi-resolution approach were visible in the satellite resolution, with a substantial reduction of both false positives and false negatives. This was not the case in the aerial (unmanned) resolution level, where a higher number of false positives balanced the decrease in the number of false negatives. In a disaster scenario, the objective is to identify which buildings are damaged (hence, having potential victims). Therefore, it is preferable to lower the number of false negatives, maybe at the cost of a slight increase in false positives.

Despite the successful multi-resolution feature fusion approach for the image classification of building damages, there is no information regarding the individual contribution of each of the levels of resolution in the image classification task. Moreover, the presented results are mainly related to the overall accuracy and behavior of the multi-resolution feature fusion and baseline experiments. More research is needed to assess which signs of damage are better captured with this multi-resolution feature fusion approach, for each of the resolution levels. The focus of this work was on the fusion of the several multi-resolution feature maps. However, other networks can be assessed to perform the same task. In this regard, MR\_b, for example, can be directly applied to pre-trained modules, where the last set of activations can be concatenated and posteriorly fed to the fusion module. In this case, there is no need to re-train a new network for a specific multi-resolution feature fusion approach. There is an ongoing increase in the amount of collected image data, where a multi-resolution approach could harness this vast amount of information and help build stronger classifiers for the image classification of building damages. Moreover, given the recent contributions focusing on online learning [27], the initial satellite images from a given disastrous event could be continuously refined with location-specific image samples that come from other resolutions. In such conditions, the use of a multi-resolution feature fusion approach would be optimal. This is especially relevant in an early post-disaster setting, where all these multi-resolution data would be captured independently with different sensors and at different stages of the disaster management cycle.

This multi-resolution feature fusion approach can also be assessed when considering other image classification problems with more classes. There is an ever-growing amount of collected remote sensing imagery and taking advantage of this large quantity of data would be optimal.

**Author Contributions:** D.D., F.N. conceived, designed the experiments and drafted the first release of the paper. D.D. implemented the experiments. N.K. and G.V. helped improving the method and the experimental setup, and contributed to the improvement of the paper writing and structure.

**Funding:** This work was funded by INACHUS (Technological and Methodological Solutions for Integrated Wide Area Situation Awareness and Survivor Localisation to Support Search and Rescue Teams), an EU-FP7 project with grant number 607522.

**Acknowledgments:** The authors would like to thank the DigitalGlobe Foundation ([www.digitalglobefoundation.com](http://www.digitalglobefoundation.com)) for providing satellite images on Italy and Ecuador.

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

## References

1. Dell’Acqua, F.; Gamba, P. Remote sensing and earthquake damage assessment: Experiences, limits, and perspectives. *Proc. IEEE* **2012**, *100*, 2876–2890. [[CrossRef](#)]
2. Eguchi, R.T.; Huyck, C.K.; Ghosh, S.; Adams, B.J.; McMillan, A. Utilizing new technologies in managing hazards and disasters. In *Geospatial Techniques in Urban Hazard and Disaster Analysis*; Showalter, P.S., Lu, Y., Eds.; Springer: Dordrecht, The Netherlands, 2009; pp. 295–323. ISBN 978-90-481-2237-0.
3. United Nations. INSARAG Guidelines, Volume II: Preparedness and Response, Manual B: Operations. 2015. Available online: <https://www.insarag.org/methodology/guidelines> (accessed on 6 June 2018).
4. Curtis, A.; Fagan, W.F. Capturing damage assessment with a spatial video: An example of a building and street-scale analysis of tornado-related mortality in Joplin, Missouri, 2011. *Ann. Assoc. Am. Geogr.* **2013**, *103*, 1522–1538. [[CrossRef](#)]
5. Ishii, M.; Goto, T.; Sugiyama, T.; Saji, H.; Abe, K. Detection of earthquake damaged areas from aerial photographs by using color and edge information. In Proceedings of the ACCV2002: The 5th Asian Conference on Computer Vision, Melbourne, Australia, 23–25 January 2002.
6. Vu, T.T.; Matsuoka, M.; Yamazaki, F. Detection and Animation of Damage Using Very High-Resolution Satellite Data Following the 2003 Bam, Iran, Earthquake. *Earthq. Spectra* **2005**, *21*, 319–327. [[CrossRef](#)]
7. Balz, T.; Liao, M. Building-damage detection using post-seismic high-resolution SAR satellite data. *Int. J. Remote Sens.* **2010**, *31*, 3369–3391. [[CrossRef](#)]
8. Brunner, D.; Schulz, K.; Brehm, T. Building damage assessment in decimeter resolution SAR imagery: A future perspective. In Proceedings of the 2011 Joint Urban Remote Sensing Event, Munich, Germany, 11–13 April 2011; pp. 217–220.
9. Armesto-González, J.; Riveiro-Rodríguez, B.; González-Aguilera, D.; Rivas-Brea, M.T. Terrestrial laser scanning intensity data applied to damage detection for historical buildings. *J. Archaeol. Sci.* **2010**, *37*, 3037–3047. [[CrossRef](#)]
10. Khoshelham, K.; Oude Elberink, S.; Xu, S. Segment-based classification of damaged building roofs in aerial laser scanning data. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 1258–1262. [[CrossRef](#)]
11. Kerle, N.; Hoffman, R.R. Collaborative damage mapping for emergency response: The role of Cognitive Systems Engineering. *Nat. Hazards Earth Syst. Sci.* **2013**, *13*, 97–113. [[CrossRef](#)]
12. Saito, K.; Spence, R.; Booth, E.; Madabhushi, G.; Eguchi, R.; Gill, S. Damage assessment of Port-au-Prince using Pictometry. In Proceedings of the 8th International Conference on Remote Sensing for Disaster Response, Tokyo, Japan, 30 September–1 October 2010.
13. Kerle, N. Satellite-based damage mapping following the 2006 Indonesia earthquake—How accurate was it? *Int. J. Appl. Earth Obs. Geoinf.* **2010**, *12*, 466–476. [[CrossRef](#)]
14. Gerke, M.; Kerle, N. Automatic structural seismic damage assessment with airborne oblique Pictometry® imagery. *Photogramm. Eng. Remote Sens.* **2011**, *77*, 885–898. [[CrossRef](#)]
15. Murtiyoso, A.; Remondino, F.; Rupnik, E.; Nex, F.; Grussenmeyer, P. Oblique aerial photography tool for building inspection and damage assessment. In Proceedings of the ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Denver, CO, USA, 17–20 November 2014; Volume XL–1, pp. 309–313.



16. Nex, F.; Rupnik, E.; Toschi, I.; Remondino, F. Automated processing of high resolution airborne images for earthquake damage assessment. In Proceedings of the ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Denver, CO, USA, 17–20 November 2014; Volume XL–1, pp. 315–321.
17. Vetrivel, A.; Gerke, M.; Kerle, N.; Nex, F.; Vosselman, G. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 45–59. [[CrossRef](#)]
18. CGR Supplies Aerial Survey to JRC for Emergency. Available online: <http://www.cgrspa.com/news/cgr-fornira-il-jrc-con-immagini-aeree-per-le-emergenze/> (accessed on 9 November 2015).
19. Fernandez Galarreta, J.; Kerle, N.; Gerke, M. UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. *Nat. Hazards Earth Syst. Sci.* **2015**, *15*, 1087–1101. [[CrossRef](#)]
20. Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Towards a more efficient detection of earthquake induced facade damages using oblique UAV imagery. In Proceedings of the ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Bonn, Germany, 4–7 September 2017; Volume XLII-2/W6, pp. 93–100.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the ICLR 2015, San Diego, CA, USA, 7–9 May 2015; pp. 1–13.
24. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for high resolution remote sensing imagery using a fully convolutional network. *Remote Sens.* **2017**, *9*, 498. [[CrossRef](#)]
25. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 645–657. [[CrossRef](#)]
26. Wei, Y.; Wang, Z.; Xu, M. Road structure refined CNN for road extraction in aerial images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 709–713. [[CrossRef](#)]
27. Vetrivel, A.; Kerle, N.; Gerke, M.; Nex, F.; Vosselman, G. Towards automated satellite image segmentation and classification for assessing disaster damage using data-specific features with incremental learning. In Proceedings of the GEOBIA 2016, Enschede, The Netherlands, 14–16 September 2016.
28. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Washington, DC, USA, 7–13 December 2015; pp. 2650–2658.
29. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [[CrossRef](#)]
30. Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach. In Proceedings of the ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, Riva del Garda, Italy, 4–7 June 2018; Volume IV–2, pp. 89–96.
31. Vu, T.T.; Ban, Y. Context-based mapping of damaged buildings from high-resolution optical satellite images. *Int. J. Remote Sens.* **2010**, *31*, 3411–3425. [[CrossRef](#)]
32. Yamazaki, F.; Vu, T.T.; Matsuoka, M. Context-based detection of post-disaster damaged buildings in urban areas from satellite images. In Proceedings of the 2007 Urban Remote Sensing Joint Event, Paris, France, 11–13 April 2007; pp. 1–5.
33. Miura, H.; Yamazaki, F.; Matsuoka, M. Identification of damaged areas due to the 2006 Central Java, Indonesia earthquake using satellite optical images. In Proceedings of the 2007 Urban Remote Sensing Joint Event, Paris, France, 11–13 April 2007; pp. 1–5.
34. Dong, L.; Shan, J. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS J. Photogramm. Remote Sens.* **2013**, *84*, 85–99. [[CrossRef](#)]
35. Li, X.; Yang, W.; Ao, T.; Li, H.; Chen, W. An improved approach of information extraction for earthquake-damaged buildings using high-resolution imagery. *J. Earthq. Tsunami* **2011**, *5*, 389–399. [[CrossRef](#)]

36. Ma, J.; Qin, S. Automatic depicting algorithm of earthquake collapsed buildings with airborne high resolution image. In Proceedings of the International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012; pp. 939–942.
37. Vetrivel, A.; Gerke, M.; Kerle, N.; Vosselman, G. Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images. *ISPRS J. Photogramm. Remote Sens.* **2015**, *105*, 61–78. [[CrossRef](#)]
38. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [[CrossRef](#)]
39. Cusicanqui, J.; Kerle, N.; Nex, F. Usability of aerial video footage for 3D-scene reconstruction and structural damage assessment. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 1583–1598. [[CrossRef](#)]
40. Hasegawa, H.; Aoki, H.; Yamazaki, F.; Matsuoka, M.; Sekimoto, I. Automated detection of damaged buildings using aerial HDTV images. In Proceedings of the IGARSS 2000, Honolulu, HI, USA, 24–28 July 2000; Volume 1, pp. 310–312.
41. Mitomi, H.; Matsuoka, M.; Yamazaki, F. Application of automated damage detection of buildings due to earthquakes by panchromatic television images. In Proceedings of the 7th US National Conference on Earthquake Engineering, Boston, MA, USA, 21–25 July 2002.
42. Vetrivel, A.; Gerke, M.; Kerle, N.; Vosselman, G. Identification of structurally damaged areas in airborne oblique images using a Visual-Bag-of-Words approach. *Remote Sens.* **2016**, *8*, 231. [[CrossRef](#)]
43. Gomez-Chova, L.; Tuia, D.; Moser, G.; Camps-Valls, G. Multimodal classification of remote sensing images: A review and future directions. *Proc. IEEE* **2015**, *103*, 1560–1584. [[CrossRef](#)]
44. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Van-Den Hengel, A. Effective semantic pixel labelling with convolutional networks and Conditional Random Fields. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 36–43.
45. Hermosilla, T.; Ruiz, L.A.; Recio, J.A.; Estornell, J. Evaluation of Automatic Building Detection Approaches Combining High Resolution Images and LiDAR Data. *Remote Sens.* **2011**, *3*, 1188–1210. [[CrossRef](#)]
46. Sohn, G.; Dowman, I. Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction. *ISPRS J. Photogramm. Remote Sens.* **2007**, *62*, 43–63. [[CrossRef](#)]
47. Prince, D.; Sidike, P.; Essa, A.; Asari, V. Multifeature fusion for automatic building change detection in wide-area imagery. *J. Appl. Remote Sens.* **2017**, *11*, 026040. [[CrossRef](#)]
48. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [[CrossRef](#)]
49. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *Computer Vision—ACCV 2016*; Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2017; Volume 10111, pp. 180–196. ISBN 978-3-319-54180-8.
50. Liu, Y.; Piramanayagam, S.; Monteiro, S.T.; Saber, E. Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fully-convolutional neural networks and higher-order CRFs. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1561–1570.
51. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. In Proceedings of the ICLR 2016, Caribe Hilton, San Juan, Puerto Rico, 2–4 May 2016.
52. Yu, F.; Koltun, V.; Funkhouser, T. Dilated residual networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
53. Hamaguchi, R.; Fujita, A.; Nemoto, K.; Imaizumi, T.; Hikosaka, S. Effective use of dilated convolutions for segmenting small object instances in remote sensing images. *arXiv*, 2017; arXiv:1709.00179.
54. 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, 7–9 July 2015.
55. Springenberg, J.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. In Proceedings of the ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
56. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

57. Tajbakhsh, N.; Shin, J.Y.; Gurudu, S.R.; Hurst, R.T.; Kendall, C.B.; Gotway, M.B.; Liang, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans. Med. Imaging* **2016**, *35*, 1299–1312. [[CrossRef](#)] [[PubMed](#)]
58. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011.
59. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A convolutional neural network cascade for face detection. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5325–5334.
60. Boulch, A.; Saux, B.L.; Audebert, N. Unstructured point cloud semantic labeling using deep segmentation networks. In Proceedings of the Eurographics Workshop on 3D Object Retrieval, Lyon, France, 23–24 April 2017; The Eurographics Association: Lyon, France, 2017.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).