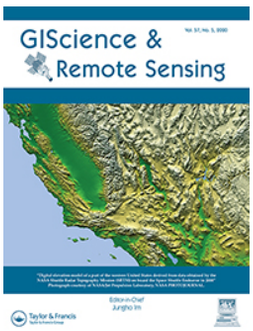

GIScience & Remote Sensing

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgrs20

Detection of seismic façade damages with multi-temporal oblique aerial imagery

Diogo Duarte , Francesco Nex , Norman Kerle & George Vosselman

To cite this article: Diogo Duarte , Francesco Nex , Norman Kerle & George Vosselman (2020) Detection of seismic façade damages with multi-temporal oblique aerial imagery, GIScience & Remote Sensing, 57:5, 670-686, DOI: 10.1080/15481603.2020.1768768

To link to this article: https://doi.org/10.1080/15481603.2020.1768768

Published online: 25 May 2020.

GISCIENCE & REMOTE SENSING
2020, VOL. 57, NO. 5, 670–686
https://doi.org/10.1080/15481603.2020.1768768

Taylor & Francis
Taylor & Francis Group

Detection of seismic façade damages with multi-temporal oblique aerial imagery

Diogo Duarte [a,b,c], Francesco Nex [a], Norman Kerle [a] and George Vosselman [a]

[a]Faculty of Geo-information Science and Earth Observation, University of Twente, Enschede, Netherlands; [b]Department of Mathematics, University of Coimbra, Coimbra, Portugal; [c]Institute for Systems Engineering and Computers of Coimbra, University of Coimbra, Coimbra, Portugal

ABSTRACT

Remote sensing images have long been recognized as useful for the detection of building damages, mainly due to their wide coverage, revisit capabilities and high spatial resolution. The majority of contributions aimed at identifying debris and rubble piles, as the main focus is to assess collapsed and partially collapsed structures. However, these approaches might not be optimal for the image classification of façade damages, where damages might appear in the form of spalling, cracks and collapse of small segments of the façade. A few studies focused their damage detection on the façades using only post-event images. Nonetheless, several studies achieved better performances in damage detection approaches when considering multi-temporal image data. Hence, in this work a multi-temporal façade damage detection is tested. The first objective is to optimally merge pre- and post-event aerial oblique imagery within a supervised classification approach using convolutional neural networks to detect façade damages. The second objective is related to the fact that façades are normally depicted in several views in aerial manned photogrammetric surveys; hence, different procedures combining these multi-view image data are also proposed and embedded in the image classification approach. Six multi-temporal approaches are compared against 3 mono-temporal ones. The results indicate the superiority of multi-temporal approaches (up to ~25% in f1-score) when compared to the mono-temporal ones. The best performing multi-temporal approach takes as input sextuples (3 views per epoch, per façade) within a late fusion approach to perform the image classification of façade damages. However, the detection of small damages, such as smaller cracks or smaller areas of spalling, remains challenging in this approach, mainly due to the low resolution (~0.14 m ground sampling distance) of the dataset used.

ARTICLE HISTORY
Received 14 October 2019
Accepted 8 May 2020

KEYWORDS
Deep learning; change detection; remote sensing; convolutional neural networks; Pictometry; CNN

1. Introduction

Earthquakes are the deadliest natural hazard, and are responsible for almost a quarter of the recorded economic losses by disasters in the last 20 years (Wallemacq and House 2018). The built-up environment plays a major role in both of the latter issues, where a growing migration to megacities is further increasing the risk associated with earthquakes (Dong and Shan 2013). A synoptic assessment of the damaged buildings over an affected region is therefore useful in the several steps of the disaster management cycle. The localization of collapsed and partially collapsed buildings is mandatory for an efficient deployment of first responders immediately after an event occurs (United Nations 2015). On the other hand, the thorough damage assessment of a building can be also valuable for recovery and insurance purposes (United Nations 2009) performed at a later stage of the disaster management cycle.

The manual inspection of damaged buildings is a time and resource consuming procedure. Many approaches using remote sensing have been proposed for building damage assessment at several scales and with different platforms and sensors. Satellite, airborne, and terrestrial platforms coupled with optical (Curtis and Fagan 2013; Cusicanqui, Kerle, and Nex 2018; Dubois and Lepage 2014; Sui et al. 2014), radar (Brunner, Schulz, and Brehm 2011; Jung et al. 2018; Li et al. 2012) or laser instruments (Armesto-González et al. 2010; Khoshelham, Oude Elberink, and Sudan 2013) have already been proposed as remotely sensed data for building damage assessment. However, the largest effort has been focused on the methods using optical images as input (Cusicanqui, Kerle, and Nex 2018; Dell'Acqua and Gamba 2012; Duarte et al. 2018a; Dubois and Lepage 2014; Vetrivel et al. 2017). This is due to several factors, among them the availability of images being collected from satellite and aerial platforms when compared with laser scanners for example, and their frequent use in photogrammetric processes to generate 3D models (Gerke and Kerle 2011; Vetrivel et al. 2017).

Many approaches have been proposed to detect damaged regions in remote sensing imagery (Duarte et al. 2018a; Fernandez Galarreta, Kerle, and Gerke

CONTACT Diogo Duarte d.duarte@utwente.nl

2015; Gerke and Kerle 2011; Sui et al. 2014; Vetrivel et al. 2017). Often these approaches rely on features extracted from images which are later used as input for a given classifier. Convolutional neural networks (CNN) have been shown to outperform the image classification with traditional handcrafted features in many applications (Krizhevsky, Sutskever, and Hinton 2012; Long, Shelhamer, and Darrell 2015), and this has been confirmed in the detection of building damages in remote sensing images (Duarte et al. 2018a; Vetrivel et al. 2016), too.

Most of the recent image-based damage detection frameworks rely on CNN to determine if a given image patch contains a damage region in a binary classification approach (Duarte et al. 2018a; Nex et al. 2019; Vetrivel et al. 2017). Such frameworks were designed to detect rubble piles and/or debris from satellite (Duarte et al. 2018b) and aerial images (Vetrivel et al. 2017). The details visible in satellite images and the (near) nadir view only allow a rough analysis and identification of collapsed buildings (Kerle and Hoffman 2013).

In the literature most of the contributions consider the detection of partially or completely collapsed buildings. Given that these are trained with image samples containing rubble piles and debris, these are not optimal for the façade case (Duarte et al. 2017), see Figure 1. The specific case of façade damage detection is only discussed in a few contributions. Fernandez Galarreta, Kerle, and Gerke (2015) extracted cracks and spalling from façades from unmanned aerial vehicle (UAV) imagery, relying both on the image and 3D features. Gerke and Kerle (2011) used multi-view aerial imagery and derived a 3D point cloud to extract features and identify damaged buildings, and at the same time classified the damage of a given building into three classes, based on the European Macroseismic Scale (EMS-98). More recently, Tu et al. (2017) identified damaged façades using local symmetry features and the Gini Index extracted from aerial oblique images. The authors assumed symmetric façades and considered the deviations from that symmetry to be façade damage proxies. These studies only used post-event image data, while the potential of using multi-temporal image information has already been demonstrated in works using both nadir (Dong and Shan 2013; Murtiyoso et al. 2014) and oblique imagery (Duarte et al. 2019; Vetrivel et al. 2016). Vetrivel et al. (2016) tested the potential of multi-temporal aerial imagery by using a correlation coefficient to determine the similarity between two rectified façade image patches. Duarte et al. (2019) reported preliminary results regarding the use of a supervised classifier to detect damaged façades using multi-temporal oblique imagery. The authors used two different approaches to merge the multi-temporal oblique image data, which clearly outperformed mono-temporal approaches. Nonetheless, the results only achieved ~66% accuracy in the best performing multi-temporal approach.

The use of airborne oblique imagery has substantially increased in the last decade, allowing the efficient collection of detailed high-resolution information over urban areas. Aerial surveys are regularly performed in

Figure 1. Examples of nadir images depicting rubble piles and debris, left. Damaged façades shown in oblique imagery, right.

many countries and enable their use to detect changes over time and after sudden events.

This paper is derived from a dissertation thesis chapter (Duarte 2020) and extends on the previously reported work by Duarte et al. (2019). Namely, it proposes different methods for the use of multi-temporal aerial oblique image data to detect façade damages caused by a catastrophic event (i.e. an earthquake). Exploiting the availability of multi-temporal datasets, six different approaches to detect façade damages from pre- and post-event are discussed. Three mono-temporal approaches (using only post-event data) are used as reference.

The focus on the multi-temporal experiments is twofold:

(1) To determine the optimal approach to merge the multi-temporal information within deep learning framework for the image classification of façade damages;
(2) To leverage the redundancy present in aerial (manned) surveys to extract several façade image patches per façade in each epoch, and to combine these within the frameworks presented in (1).

An additional effort is made to conceive methods exploiting only image information and pre-event 3D models to be (potentially) used in near-real time conditions (assuming the availability of multi-temporal data), when fast and automated methods are needed.

The following section presents a short background. The datasets used in the experiments are presented in section 3. This section also addresses the façade extraction from the aerial oblique imagery. Section 4 presents the methodology for the multi-temporal image classification of façade damages. Section 5 presents the experiments and results, which are followed by the discussion and conclusions.

2. Background

This sub-section focuses on CNN and its role in multi-temporal studies using remote sensing imagery. It starts with a brief description of recent developments in CNN that were adapted to this work. An overview of multi-temporal approaches using remote sensing imagery is also given.

Supervised deep learning methods have become an established machine learning technique for image-based tasks, where CNN play a central role. CNN usually achieve high discriminative capacity by stacking convolutions in a hierarchical manner, learning from lower level features to higher levels of abstraction (Krizhevsky, Sutskever, and Hinton 2012). However, in this way each layer is only connected with the previous and posterior layer. Hence, there is feature information that may be lost during back-propagation, especially from earlier layers (Yu, Koltun, and Funkhouser 2017). To tackle this, short connections between non-adjacent layers started to be used (He et al. 2016; Huang et al. 2017). (He et al. 2016), introduced the concept of residual connection, in which the authors used short connections through element-wise addition of nonconsecutive layers. This allowed for the use of deeper networks while maintaining their efficiency, which is often translated into more accurate predictions.

More recently it was found to be preferable to concatenate the feature information instead of performing element-wise addition. (Huang et al. 2017) proposed the densely connected convolutional network, introducing short connections in the form of the concatenation of feature maps. This difference allows the model to be more compact, given that every layer receives feature information from the layers preceding it. Thus, features of a given layer may be re-used in later stages of the network, which offers them more representability.

Another aspect of CNN that is closely related with remote sensing is the use of dilated convolutions. These were proposed by Yu and Koltun (Yu, Koltun, and Funkhouser 2017) and are convolutions with a kernel with pre-defined gaps. This is translated into a wider receptive field, capturing more contextual information. Given that the receptive field of the dilated convolutions is larger, it can capture features over a larger image region, while maintaining a low number of parameters due to the gapped kernel (Yu, Koltun, and Funkhouser 2017). This has been extensively used by researchers in remote sensing image recognition tasks (Hamaguchi et al. 2017; Jiang and Lu 2018; Persello and Stein 2017; Zhang et al. 2019).

The remote sensing community has been adapting and proposing CNN approaches for earth observation tasks and data. For example, such CNN have been directly used in image classification (Hu et al. 2015a, 2015b; Maggiori et al. 2017; Nogueira, Penatti, and Dos Santos 2017) and image segmentation (Kampffmeyer, Salberg, and Jenssen 2016; Längkvist et al. 2016; Volpi and Tuia 2017) approaches. However, CNN have for example also been used to merge different modalities of remote sensing data (e.g., 3D and images) (Audebert, Le Saux, and Lefèvre 2018, 2017; Duarte et al. 2018a), annotate aerial images (Xia et al. 2015; Zhuo et al. 2019) and perform multi-temporal studies (Daudt et al. 2018; Jung et al. 2018; Wang et al. 2018; Zhang et al. 2019).

One of the aspects of remote sensing imagery is the fact that often there are several views of a given scene (e.g. aerial surveys with high forward and side overlap). Hence, considering such multi-view image information has been shown useful in several applications (Koukal, Atzberger,

and Schneider 2014; Liu et al. 2018; Zhao et al. 2017). Land cover classification has been improved by considering multi-view images instead of an orthophoto (Liu et al. 2018), or instead a single image chosen following some predefined criteria (Koukal, Atzberger, and Schneider 2014). While these multi-view approaches are more computationally demanding given that more data is considered for training (Liu and Abd-Elrahman 2018a), they may allow for classification models to harness more information of a given scene (e.g. variations in perspective and illumination) (Liu and Abd-Elrahman 2018b) and consequently improving its recognition capabilities.

Multi-temporal studies using CNN often focus their attention on the optimal merging of the different epochs of imagery. Several approaches have been proposed, mostly using satellite imagery and nadir constrained images. Wang et al. (Wang et al. 2018) reported that for the task of change detection in satellite imagery it would be preferable to consider the subtraction of pre- and post-event imagery, with the new image being then fed to the CNN. Daudt et al. (Daudt et al. 2018) tested two approaches to detect changes in multi-temporal satellite imagery. One of the approaches considered two branches of convolutional layers with shared weights (also known as Siamese network), one for each epoch, while the other considered a single set of convolutions performed on the concatenation of the pre- and post- event data as the first stage of the network. The authors reported that early fusion of the inputs was preferable for the detection of changes from satellite imagery. In a different study with the objective of the detection of landslides, Chen et al. (Chen et al. 2018) used a two branch network (one for each epoch of image data), where the feature maps of these streams were then merged by computing a Manhattan distance between them.

3. Datasets and CNN input generation

This section presents the image datasets used in the experiments and the process from the original aerial oblique images to the input given to the approaches indicated in section 4.

The datasets used in this paper comprise two airborne oblique image captures of the city of L'Aquila and a smaller neighboring village, Tempera. These two locations were surveyed within an approximately 9-month interval, in August 2008 and in May 2009, the latter depicting the situation after the April 2009 earthquake that occurred in central Italy.

The image acquisition was performed using the Pictometry system that contains small format DSLR cameras, four obliques (one for each cardinal direction) and one nadir. The flying height was approximately 1000 m, which translated to an average sampling distance of 0.14 m on the oblique views. The flight was performed considering a forward overlap between 60 – 70% and a side overlap between 35–45%. The pre-event image data was collected on the 8th (between 11:28 h and 11:55 h) and 9th (between 10:03 h and 12:17 h) of August 2008. The post event image data was captured on the 31st (between 11: 11 h and 13:20 h) April 2009.

Figure 2 depicts the process between the original images and the final input to the experiments. Two types of input were generated, façade image patches extracted from the original images (Figure 3, top), and these same image patches rectified using the corresponding façade 3D information (Figure 3, bottom). These were the two types of input that were used, and compared, in the experiments.

The first step was to generate the 3D point cloud, which was used to define the façade planes and subsequently extract the façades from the images. To this end the first step was to perform the image orientation of both pre- and post-event images. These shared the tie point computation with the objective of aligning the datasets. However, only the pre-event images were used for dense matching.

Figure 2 presents the overview of the main steps to extract the façade image patches from the oblique views using the 3D point cloud generated from the pre-event images. The first step was to differentiate between *on* and *off* ground points, using *lasground* from the package *lastools* (Axelsson 2000). The point cloud, with the added attribute of the normalized height surface, was the input for a plane-based segmentation, which was followed by a connected component analysis, generating the final roof segments (Vosselman 2012).

The roof segments were then projected into the *xy* plane (see Figure 2 – Façade definition). The approach then assumed that each building segment contains 4 façades and that they are mutually perpendicular. With this assumption the roof points were fitted with a minimum-area bounding rectangle (Freeman and Shapira 1975) (red rectangle in Figure 2- Façade definition), defining the 4 main façade directions of a given building. The façade planes were then defined by the *xy* coordinates of the edges of the rectangle, where the *z* is obtained from the normalized height and from the difference between the mean *z* coordinate of the roof and the mean normalized height value. At this stage every façade was finally defined by 4 facade corners.

The projection matrices, coming from the orientation step, were then used to project the facade pixels into these 3D planes. This process was repeated for all images containing a given façade, in both epochs. The façade images were downscaled/upscaled at the same spatial resolution: gaps due to different viewpoints and scales

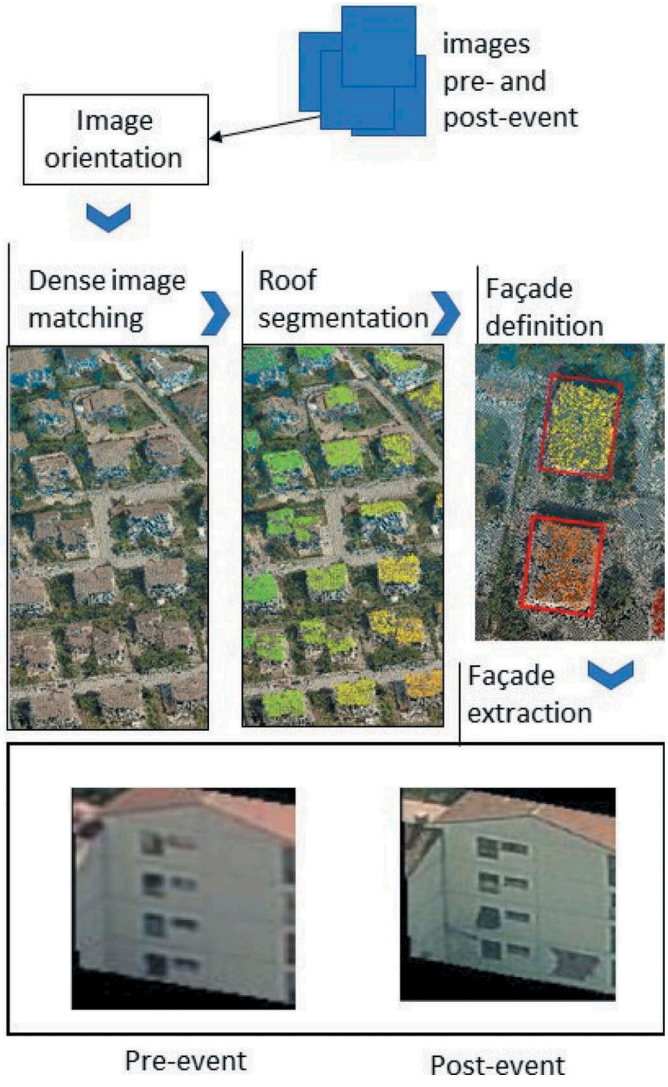

Figure 2. Overview of the main steps of the façade extraction from the aerial images. The segments in the Roof segmentation thumbnail are color coded. The red rectangle in the Facade definition thumbnail indicates the main 4 façades extracted from the roof points. Below, example of a façade, showing both pre- and post-event. These façade image patches (image pair) are one of the inputs to the experiments (see Figure 3).

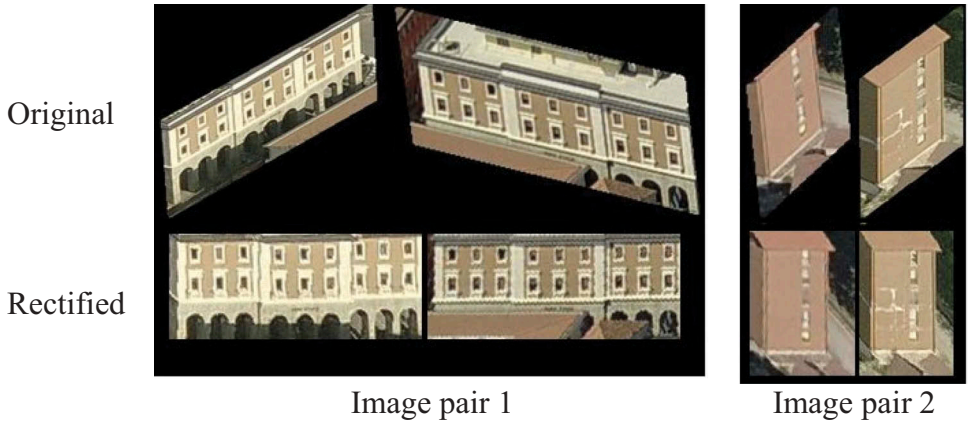

Figure 3. The two types of input used in the experiments, considering two views of two façades. Each of these pairs is an example of the input used in one set of experiments (see Figure 6). Top, original facade image patches. Bottom, corresponding rectified façade image patches.

of the oblique views were interpolated using a nearest neighbors' algorithm. This ensured the registration of the different views of the same façade. The visibility of the four façade corners in the images was used to detect occluded parts. The pre-event point cloud was used to perform the visibility analysis as it was assumed that all the buildings are still standing, and the number of occlusions is higher. A façade was considered occluded if at least two of the four corners were not visible in the image.

Examples of the two types of extracted (original and rectified façade image patches) data can be seen in Figure 3. All the experiments, both mono- and multi-temporal, were tested considering separately the original and the rectified façade image patches. The aim was to test if the approaches could leverage the rectification and registration of the façade image patches to perform a better image classification of façade damages, while at the same time assuming the interpolated areas which might modify the already small damage evidences present in the façades.

To take advantage of having several façade image patches per façade per epoch, these pre- and post-event image data (original and rectified) were combined in two distinct ways:

(1) Image pairs – these were created to associate each pre-event façade image patch to all the post-event façade image patches of a given façade (Figure 2). This was performed for all façades. This input is related with the experiments MTa (see section 4.3). Performing this combination between different views of the same façade allowed to generate more input data, instead of considering an image patch per façade, which would make the training dataset very small (only 88 damaged and 90 undamaged façades were possible to extract from the images).
(2) Image sextuples – these were created to combine three pre-event image patches, with three post-event image patches of a given façade. In this case the maximum amount of combinations allowed per façade was 50, given an unbalanced number of views per facade. This input is related with the experiments MTb (see section 4.3). Considering several views per façade per epoch enables the network to learn the similarities between different views of the same façade. This is due to the fact that these networks compute features that are shared across all the different views, instead of focusing on single image pairs like in (1).

This allowed the extraction of 4,546 image pairs and 5,179 image sextuples from a total of 178 façades (see Table 1).

Table 1. Number of image pairs and image sextuples extracted considering the 178 façades.

	Image pairs	Image sextuples	Façades
Damaged	2,274	2,559	88
Not damaged	2,272	2,610	90
Total	4,546	5179	178

4. Methodology

Six multi-temporal approaches were designed, tested, and compared with three mono-temporal approaches. The input into the multi-temporal approaches was pre- and post-event façade image patches captured from different oblique views (original and rectified), as described in the data section. The focus of the experiments was on the optimal merging of pre- and post-event image information within a supervised deep learning framework for the image classification of façade damages using CNN. Table 1 illustrates the small amount of data to perform this multi-temporal façade analysis using aerial manned imagery. This issue was central to the current work and is one of the main limitations of the experiments. Several measures were taken to attenuate the lack of data, and these are further detailed in this section and in the experiments.

This section starts by laying out the main characteristics of the used CNN, in the following paragraphs. The following sub-section formalizes the used network, while the final sub-sections explain the rationale behind each performed test.

4.1. Network definition

The basic network used in the experiments is presented in this sub-section, and it was a central component of the mono- and multi-temporal approaches (see *stream* in Figure 5, Figures 6 and 7). This network was composed by consecutively stacking of 2 modules, dense blocks and transitional layers. This composition was proposed in (Huang et al. 2017), where the authors derived several networks from different combinations of these modules. In the current work, the network used, was composed of 4 dense blocks, with transitional layers between these blocks. While maintaining the number of dense blocks presented in Huang et al. (2017), a lower number of layers per dense block was considered in this work. Given the small amount of data, decreasing the model complexity did not impact its representability and contributed to reduce overfitting.

Each dense block contained two sets of two convolutions, as indicated in Figure 4. In Figure 4, the conv field indicates the group: batch normalization, relu and convolution. A *dropout* layer (0.2) was also added after the

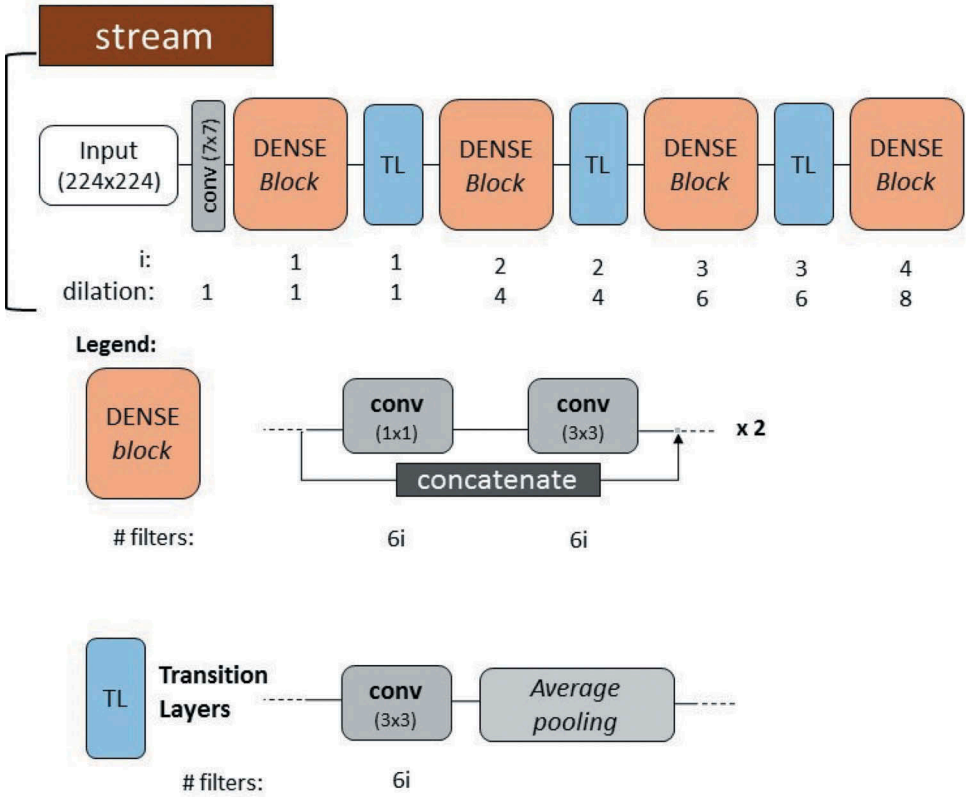

Figure 4. Network used in the experiments (stream), composed of dense blocks and transition layers. conv depicts the group batch normalization, relu and convolution. The number of filters and dilation value is affected by the number of dense block, transitional layer group, as indicated by *i*.

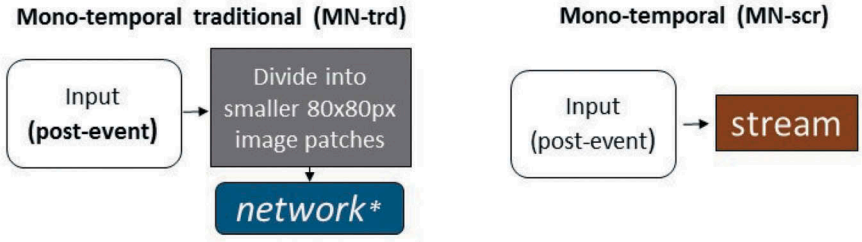

Figure 5. Mono-temporal approaches, MN-trd and MN-scr. * The network in italic refers to the aerial (manned) network presented in (Duarte et al. 2018a). The stream refers to the network presented in Figure 4. Input refers to façade image pairs.

first convolution, to further prevent overfitting (Clevert, Unterthiner, and Hochreiter 2016). Each transitional layer contained a convolution and it was followed by average pooling with stride 2 in order decrease the feature map size from the initial 224x224px to the final 28x28px. The façade image patch given as input (both original and rectified) was zero padded to fit the 224x224px size. In rare cases where the façade image patch was larger than the 224x224px, it was resized to fit the input size while keeping the aspect ratio. This input size was mainly chosen to fit the fine-tuning experiment.

The number of filters per convolution was tied to the growth rate (Huang et al. 2017), which was defined as 6 (see Figure 4). This growth was set in order not to overfit the small set of data for the current study, while following the general assumption that more filters are needed to represent more complex features later in the network (Szegedy et al. 2015).

Given the small damage evidences often present in façades that did not collapse (see introduction figure) it was mandatory that a given network would be able to detect such small details. In this way it was important to

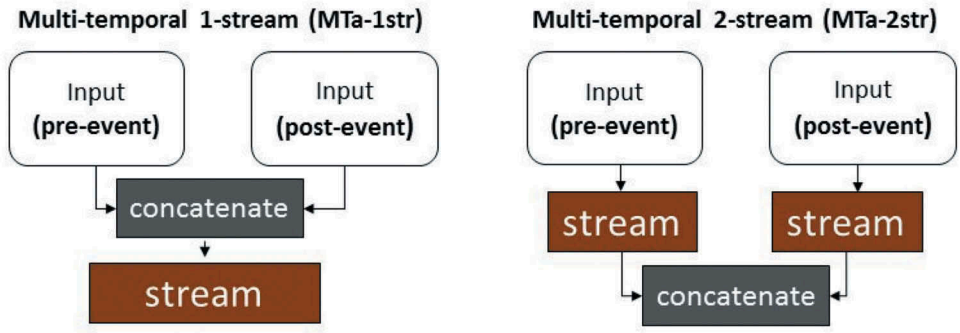

Figure 6. MTa group of experiments. Façade image pairs are fed to the experiments present in this figure.

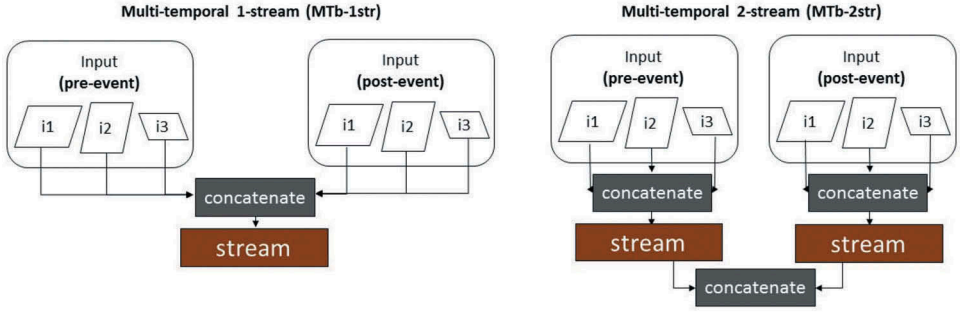

Figure 7. MTb group of experiments. Façade image sextuples are considered as input and indicated by i1-3 for each epoch.

retain contextual information, i.e. in the vicinity of the damaged area. Only then a network would be able to differentiate these small damage evidences from other areas with similar texture but in a different context. Aiming at capturing this context, dilated convolutions were used in the network. Such dilated convolutions use a kernel with defined gaps, aiming at the capture of more contextual information(Yu and Koltun 2016).As can be seen in Figure 4, the dilation factor is also growing with the number of filters even if at a smaller rate. The last set of dilated convolutions had a receptive field of 19×19.

The classification part of the network was performed by coupling batch normalization, relu, global average pooling and a dense layer of size 1 at the end of all networks.

4.2. Mono-temporal approaches

Three mono-temporal approaches were tested. These served as baseline for the multi-temporal methods (see Figure 5).

The mono-temporal traditional (MN-trd) directly used a network trained on aerial image patches containing debris and rubble piles, as in (Duarte et al. 2018a). This network was trained on aerial (manned) image samples of 7 different geographical locations and using approximately 5,400 image samples in total. These locations include cities with similar urban design as to L'Aquila and Tempera (e.g., Amatrice, Italian city). For this approach each post-event façade image patch was divided into smaller 50px squared patches. The latter were then classified for damage. In cases where a façade image patch contained at least one of these squared patches classified as damage, the whole façade image patch was considered damaged. This was performed for every façade image patch of a given façade. This experiment aimed at understanding how a network trained solely on debris and rubble piles and mostly using nadir imagery could be used for the specific case of the detection of façade damages.

The other two mono-temporal approaches also only used post-event façade image patches. The mono-temporal, MN-scr, was trained from scratch, while the MN-ft was fine-tuned on *densenet* (DenseNet121 as in (Huang et al. 2017)), where only the last dense block layers were re-trained with the façade image patches coming from the different oblique views. This experiment was deemed necessary given the low amount of

data and where the model could leverage the knowledge of the feature information learned on the ImageNET dataset.

4.3. Multi-temporal approaches

In this subsection, six multi-temporal approaches are presented. Overall, these experiments, aimed at: 1) better understanding how to merge the multi-temporal façade image patches within a CNN for the image classification of façade damages, and 2) embedding the façade image pairs and façade image sextuples defined in section 2.1 in the experiments. Figures 6 and 7, show the six different approaches. Two different ways to integrate the data from multiple perspectives and epochs were adopted and tested in the approaches, respectively considering the image pairs (1) and image sextuples (2):

(3) The group MTa (see Figure 6) considered as input only image pairs, as described in section 3.1. In MTa three different strategies were adopted. The MTa-1str concatenated the images in the channels dimension and subsequently fed this to the network defined previously. On the other hand, MT-2str, assumed one convolutional block for each epoch which were later concatenated. The MT-2str-ws (or Siamese) was similar to MT-2str, but in this case the convolution weights were shared between the two streams.

(4) The group MTb (Figure 7) considered as input the image sextuples defined in section 3.1. The rationale of these approaches followed the same concept of the group MTa. The MTb-1str concatenated the six images (three per epoch), where this 18-channel image was then fed to the network, while MTb-2str considered a convolutional set per epoch. In this case a concatenation of the three images per epoch was performed, where this 9-channel image was fed to an independent convolutional block. MT-2str-sw (or Siamese) was only different from MTb-2str, given that the convolutions were shared across streams. In this case features were computed not only across epochs but also across different views, given that for each epoch several image façade patches per façade were simultaneously considered.

The 1str set of experiments, both in MTa and MTb, forced the input image patches (both pre- and post-event) to go through a single convolutional set. On the other hand, 2str experiments had a set of epoch specific convolutions, where this information was later merged through concatenation. The 2str-sw (or Siamese) had the convolution weights shared across the epochs. These 3 different ways of considering the input data aimed at understanding which set of features were relevant for the image classification of façade damages. While the 1str approaches made use of inter-epoch features given that the inputs were concatenated at an early stage of the network, the 2str approaches gave more relevance to intra-epoch features which were merged at a later stage of the network. The 2str and 2str-sw only differed in the fact that the convolutions were shared across the epochs: in spite of having a set of convolutions for each epoch, these had the filters shared among them. Moreover, given the sharing of filters between epochs, this drastically decreased the number of parameters when compared with the 2str which did not share the convolutions.

While concatenation was used to merge the feature maps, other approaches were tested (e.g., element-wise multiplication or addition/subtraction of the feature maps). However, these did not perform as good as the simple concatenation.

All these approaches were tested using both the original and rectified façade images patches as described in 2.1.

5. Experiments and results

All the networks were trained with learning rate of 0.1 and weight decay of 10–4 (except for the fine tune experiment, where the learning rate was of 0.01) using stochastic gradient descent as optimizer (He et al. 2016; Huang et al. 2017). For each experiment one loss function, binary cross entropy, was used, given the binary classification problem being considered. The experiments were performed with early stopping, i.e. when the validation data loss stopped improving. This was performed to avoid overfitting given the small data set.

Data augmentation was performed to decrease overfitting and to give more generalization capabilities to the network (Krizhevsky, Sutskever, and Hinton 2012). This was applied only to the training data; the validation and testing sets did not consider any data augmentation other than image normalization. This augmentation consisted only of horizontal shifts and image normalization. This was due to two reasons: 1) shifts in the images could mask out the damaged area when it is close to the edge of the image; 2) rotations on the image patches could be cue for damage (e.g., slanted buildings which did not collapse). Given the small amount of data, this solution aimed at attenuating overfitting and helping generalization (Krizhevsky, Sutskever, and Hinton 2012).

The results were evaluated in terms of accuracy, recall, precision and f1 score. Recall aimed at capturing the proportion of actual damaged façades which were

Table 3. Precision, recall, accuracy and f1 score (mean) of the testing datasets for the mono- and multi-temporal approaches using the rectified (-r) façade image patches (range between brackets). These are presented at both an image pair/sextuple and at a façade level.

	Prec.	Rec.	Acc.	F1
Image/image-pair/image-sextuple level				
MN-trd-r	0.39 (0.34–0.80)	0.37 (0.26–0.38)	0.50 (0.47–0.64)	0.38 (0.30–0.52)
MN-scr.-r	0.65 (0.48–0.72)	0.83 (0.71–0.88)	0.70 (0.65–0.72)	0.71 (0.64–0.77)
MN-ft-r	0.69 (0.68–0.94)	0.68 (0.22–0.68)	0.69 (0.59–0.70)	0.68 (0.36–0.69)
MTa-1str-r	0.76 (0.70–0.80)	0.67 (0.52–0.77)	0.73 (0.67–0.76)	0.73 (0.62–0.73)
MTa-2str-r	0.70 (0.66–0.88)	0.72 (0.65–0.94)	0.73 (0.70–0.79)	0.78 (0.68–0.79)
MTa-2str-ws-r	0.86 (0.63–0.87)	0.64 (0.53–0.81)	0.73 (0.68–0.77)	0.71 (0.66–0.73)
MTb-1str-r	0.69 (0.67–0.7)	0.86 (0.72–0.96)	0.74 (0.71–0.76)	0.77 (0.71–0.79)
MTb-2str-r	0.71 (0.64–0.71)	0.64 (0.61–0.64)	0.69 (0.65–0.69)	0.66 (0.64–0.66)
MTb-2str-sw-r	0.76 (0.72–0.89)	0.75 (0.67–0.83)	0.76 (0.71–0.79)	0.75 (0.62–0.81)
Façade level				
MN-trd-r	0.47 (0.46–0.56)	0.47 (0.32–0.58)	0.58 (0.52–0.66)	0.51 (0.38–0.52)
MN-scr.-r	0.65 (0.53–0.66)	0.92 (0.62–1.00)	0.69 (0.69–0.70)	0.69 (0.64–0.76)
MN-ft-r	0.66 (0.66–1.00)	0.62 (0.22–0.62)	0.69 (0.56–0.70)	0.64 (0.30–0.64)
MTa-1str-r	0.80 (0.80–0.89)	0.66 (0.57–0.73)	0.74 (0.72–0.84)	0.72 (0.67–0.80)
MTa-2str-r	0.67 (0.67–1.00)	0.73 (0.67–1.00)	0.76 (0.72–0.83)	0.80 (0.70–0.80)
MTa-2str-ws-r	0.78 (0.60–1.00)	0.58 (0.67–1.00)	0.68 (0.66–0.70)	0.67 (0.55–0.70)
MTb-1str-r	0.69 (0.68–0.71)	0.84 (0.83–0.93)	0.74 (0.71–0.76)	0.77 (0.76–0.79)
MTb-2str-r	0.71 (0.67–0.83)	0.71 (0.55–0.77)	0.71 (0.65–0.79)	0.74 (0.60–0.77)
MTb-2str-sw-r	0.87 (0.76–0.94)	0.80 (0.45–0.95)	0.84 (0.76–0.85)	0.82 (0.62–0.87)

classified as such; precision to show the proportion of façades classified as damaged that were actually damaged. These metrics were computed three times for each experiment. In each run of the experiment the data were randomly divided in sets of training and testing (70% and 30%, respectively), where the validation data was a subset of the training set. This division was performed at a façade level, where both the image pairs and image sextuples datasets are relative to the same façades and hence comparable. In this way every façade was present in both training and testing when considering the three different splits. The mean and the range (min. and max.) of the different runs per experiment on the testing sets are shown in the results, too.

Tables 3 and 2 present the results on the testing sets for the approach using the original façade image patches and the approach using the rectified façade image patches, respectively (example of input to the approach considering the image pairs in Figure 3). A single façade is depicted in several façade image patches coming from different views (Table 1), hence several image sextuples and/or pairs per façade were considered as input for the multi-temporal approaches. Given that in both the mono- and multi-temporal approaches several façade image patches were considered for each façade, the results are presented at an image pair/sextuple (and images in mono-temporal approaches) and at a façade level. Regarding the façade level results, a façade was

Table 2. Precision, recall, accuracy and f1 score (mean) of the testing datasets for the mono- and multi-temporal approaches using the original façade image patches (range between brackets). These are presented at both an image pair/sextuple and at a façade level.

	Prec.	Rec.	Acc.	F1
Image/image-pair/image-sextuple level				
MN-trd	0.49 (0.48–0.58)	0.66 (0.55–0.73)	0.50 (0.47–0.55)	0.55 (0.52–0.64)
MN-scr.	0.58 (0.45–0.63)	0.85 (0.65–1.00)	0.64 (0.61–0.68)	0.72 (0.53–0.73)
MN-ft	0.64 (0.57–0.64)	0.84 (0.47–0.96)	0.63 (0.57–0.64)	0.67 (0.54–0.76)
MTa-1str	0.73 (0.65–0.77)	0.60 (0.60–0.69)	0.72 (0.64–0.74)	0.67 (0.62–0.71)
MTa-2str	0.83 (0.79–0.85)	0.76 (0.57–0.80)	0.81 (0.72–0.83)	0.80 (0.66–0.82)
MTa-2str-ws	0.76 (0.66–0.81)	0.60 (0.56–0.89)	0.73 (0.65–0.83)	0.69 (0.61–0.82)
MTb-1str	0.76 (0.66–0.78)	0.64 (0.52–0.64)	0.73 (0.64–0.80)	0.70 (0.58–0.81)
MTb-2str	0.71 (0.66–0.77)	0.64 (0.52–0.87)	0.76 (0.64–0.71)	0.70 (0.58–0.70)
MTb-2str-sw	0.77 (0.64–0.77)	0.75 (0.55–0.78)	0.76 (0.63–0.78)	0.75 (0.59–0.78)
Façade level				
MN-trd	0.52 (0.43–0.63)	0.67 (0.38–0.70)	0.55 (0.52–0.60)	0.60 (0.40–0.65)
MN-scr.	0.55 (0.50–0.60)	0.92 (0.67–1.00)	0.67 (0.52–0.74)	0.70 (0.57–0.72)
MN-ft	0.65 (0.54–0.70)	0.61 (0.38–0.67)	0.63 (0.59–0.70)	0.60 (0.49–0.63)
MTa-1str	0.69 (0.67–0.9)	0.82 (0.56–0.83)	0.72 (0.62–0.87)	0.74 (0.62–0.86)
MTa-2str	0.88 (0.88–0.92)	0.73 (0.64–0.79)	0.82 (0.80–0.86)	0.80 (0.74–0.85)
MTa-2str-ws	0.75 (0.70–0.81)	0.58 (0.38–0.81)	0.68 (0.59–0.83)	0.64 (0.51–0.81)
MTb-1str	0.78 (0.67–0.81)	0.58 (0.46–0.93)	0.70 (0.58–0.86)	0.67 (0.55–0.87)
MTb-2str	0.72 (0.67–0.80)	0.73 (0.56–0.73)	0.78 (0.59–0.78)	0.76 (0.56–0.76)
MTb-2str-sw	0.82 (0.75–0.85)	0.73 (0.50–0.82)	0.79 (0.60–0.82)	0.79 (0.67–0.83)

considered as damaged if the majority of the images (mono-temporal) or image pairs/sextuples (multi-temporal) of a given façade were classified as damaged.

Overall, the multi-temporal approaches clearly outperformed the mono-temporal ones. This is seen in both the MTa and MTb experiments, and also when using either the original façade image patches or the rectified ones.

In general, using an epoch-specific set of convolutions per epoch was preferable in all the multi-temporal experiments. However, the results differ when considering different inputs, original or rectified façade image patches. While having similar results, the MTb-2str-sw-r performed slightly better than MTa-2str. Hence, the use of shared convolutions (in a Siamese setting) is most valuable when considering the image sextuples using as input the rectified façade image patches. On the other hand, when using the original façade image patches, the network cannot take advantage of the simultaneous consideration of several views per façade.

The results of MTa-2str and MTb-2str-sw-r were considerably different when compared at an image pair/sextuple level, where the difference was bridged when evaluated at a façade level. While having less correctly predicted image pairs/sextuples, MTb-2str-sw-r (82% f1-score) outperformed MTa-2str (80% f1-score) at a façade level. Hence, the better results at an image pair/sextuple by MTa-2str were more distributed among the façades, not being enough to change the prediction at a façade level. On the other hand, MTb-2str-sw-r, improved the results at a façade level while having less correctly predicted image pairs/sextuples. Hence, in this case the correct predictions were more distributed among the façades, which in turn, through the majority vote, improved the results at this level.

The overall statistical measures range between the three different data splits was also smaller when using the rectified image patches. In some of the experiments (e.g., MTa-2str-r and MTa-2srt-sw-r) recall and precision achieve 1.0 at least in one of the data splits, where the approach struggled to differentiate between the two classes. However, their non-rectified counterpart did not present this behavior, indicating that the combined use of rectified façade image patches and image pairs may not be optimal.

The mono-temporal approaches presented the worst results. The traditional approach trained on rubble piles and debris was the worst performing approach, especially using rectified façade image patches.

Figure 8 presents activations (right) considering a given façade (pre- and post-rectified façade image patches) (left). These activations were extracted from the last set of activations of each of the experiments predicting on an image sample that was present in training. This aimed at understanding where the approaches were focusing their attention on a given façade image patch, to derive a given class prediction. Figure 8 B, C, D, E and F were predicted as damaged. The only clear activation focusing on the damaged area is present in E. In the B and D cases, in spite of also considering the correct damaged area, these are not so clear and often consider other areas of the image. For example in D, post-event, the balconies area was relevant for the approach to derive the damaged class (also close to the damaged portion of the façade, see red indication in Figure 8). Figure 8-A presents damaged areas which were predicted as not damaged. In C, the attention is focused on the area of vegetation occluding the façade. In both cases, A and C, the multi-temporal approaches failed to correctly classify the façade image patches.

In Figure 9 several correct (on the left) and wrong (on the right) predictions are shown considering the best performing approach. This approach correctly identifies several degrees of spalling, building segment collapses and larger cracks. However, it is not able to detect areas with small spalling when these are too small when compared with the size of the façade. Façades which only presented cracks were often missed by the approach, probably because of the limited resolution of the images.

Regarding overall processing times there are several steps that need to be taken into account, from the façade extraction from the images to the classification for damage with the convolutional neural network. The times reported here are based on a computer with an i7 processor (4^{th} generation), 32 GB RAM and a NVIDIA GeForce GTX 1060. The façade extraction process took around 41 minutes for the façades considered in the study and using mostly mono-core processes. The damage prediction part took around 21 minutes when considering the image sextuples and 11 minutes when considering the image pairs (assuming we need to predict in all the dataset, see Table 1). The remainder of the tasks can be performed preemptively and only using the pre-event imagery.

6. Discussion

All the multi-temporal approaches outperformed the mono-temporal ones. This confirms the commonly reported results on multi-temporal approaches in remote sensing, where the use of multi-temporal data is often translated into an improvement in the quality of a given task (Hussain et al. 2013; Lu et al. 2004; Singh 1989; Tewkesbury et al. 2015). The best performing approach can identify partially and totally collapsed

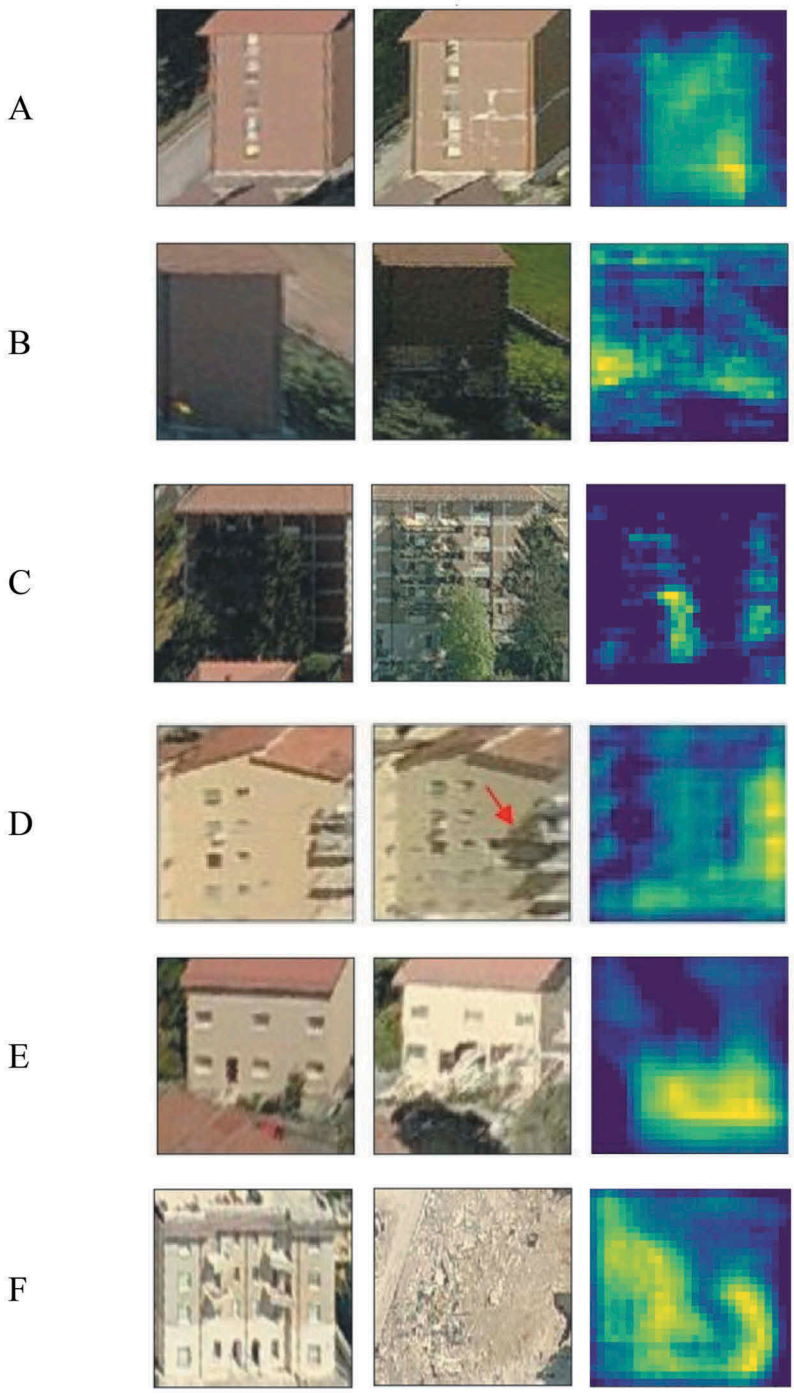

Figure 8. Activations extracted from the last activation layer of the network (training) MTb-2str-sw-r (right). Left (pre-event) and middle (post-event) facade image patches. A predicted as not damaged, while B, C, D, E and F were predicted as damaged. The red arrow present in D, indicates the damaged portion close to the balconies.

buildings, and façades with large areas of spalling and cracks. However, the overall results (best performing network with 82% f1-score) reflect some difficulties in the detection of damage to the facades, from manned aerial oblique imagery, even when also using pre-event images. This can be mainly explained by the low resolution of the used data (GSD ~0.14 m) that hinders the reliable detection of smaller signs of damage such as small cracks.

From all the mono-temporal approaches, fine-tuning or training from scratch did not lead to considerable differences as reported in other works focused on building damage detection (Duarte et al. 2018a; Vetrivel et al. 2017). The mono-temporal traditional approach using

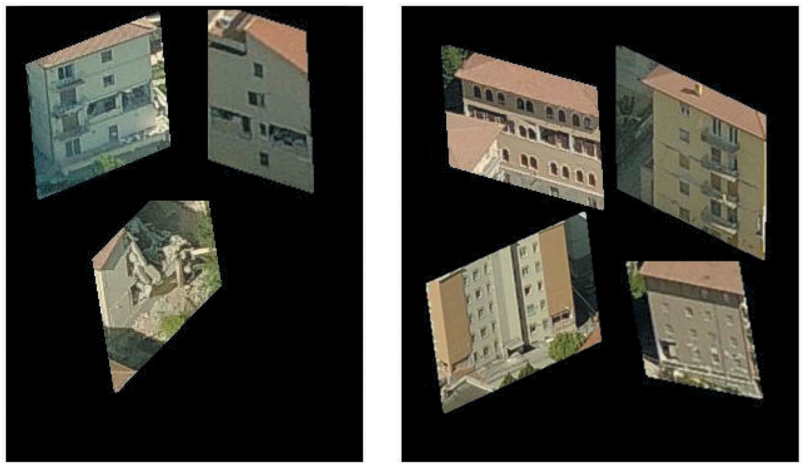

Figure 9. Left, correctly classified as damaged. Right, incorrectly classified as not-damage. Both using the best performing approach MTb-2str-sw-r, when these façades were not present in training.

a model trained with image samples depicting rubble piles and debris was the worst approach: as expected the model was not able to identify lower levels of damage present in the façades. This resulted in a high rate of both false negatives and false positives as in Duarte et al. (2017). This problem was more accentuated when using the rectified façade image patches, as the traditional approach was trained on non-rectified image patches.

Regarding the multi-temporal approaches the relevance of intra-epoch features must be noticed, which are merged later in the network. This can be observed in the results, where the single stream approaches were always outperformed by the 2str approaches, independently of the use of original or rectified image patches or the input data (i.e. image pairs or sextuples). This was also the case in Duarte et al. (2019), where approaches relying on intra-epoch features performed better, even if only considering image pairs and non-rectified façade image patches. However, recent literature in remote sensing that made use of multi-branch networks reported different results in this regard. For example, Daudt et al. (2018) reported that for the specific case of satellite imagery change detection, the concatenation of the images before being fed to the network would be preferable, as images share the features within a single convolutional set. This was also the case when localizing street view images using overhead images (Vo and Hays 2016). However, there are also studies in which the merging of the feature information later in the network, instead of considering as input a merged layer of both epochs, is preferable (Chen et al. 2018). The merging of the pre- and post-event information seems to be application dependent, where for the case of the image classification of façade damages it is preferred to merge the feature information at a later stage in the network. Also, the differences between the results at an image pair/sextuple and at a façade level are noteworthy. Since in most of the approaches the façade level results were worse than their image pair/sextuple counterpart, it seems that in such situations there was no considerable variation of the predictions within the same façade.

In the case where the façade image patches are rectified, the approach using the sextuples as input outperforms all the other approaches (MTb-2str-sw-r) at a façade level. Besides using image sextuples and rectified façade image patches the approach also shared the convolutions between the two streams. Given the rectification and registration of the façades, this approach aimed at taking advantage of considering different viewpoints of the same façade simultaneously. In this way it was expected that the networks would leverage the multi-view information and learn to distinguish between differences due to illumination, for example, and differences due to damage; extracting features across both the different epochs and the different views. However, at a façade level the approach considering only image pairs (MTa-2str) performed comparatively well (difference of 2% f1-score) while using the original image patches. In this regard, although the rectification/registration procedure is preferable, it may at the same time smooth out the often-small damage evidences present in the façades.

This work was an extension of a previous study by Duarte et al. (2019). In the present work it was introduced both the rectification of the façades before being fed to the networks and a whole new set of experiments where image sextuples were considered besides the image pairs presented in Duarte et al. (2019). This allows to compare some of the results of Table 2 (experiments

regarding the image pairs only) with the previously reported work. Namely the multi-temporal experiments 1-str and 2-str and the mono-temporal scr and trd. Overall, the quality metrics regarding these experiments were improved, when comparing with the previous study; this could be due to the use of a different architecture since the merging approaches remained similar. Dense connections were adopted in this study, while residual connections were used in the previous one. Such improvements when considering the use of dense over residual connections have been reported before (Huang et al. 2017) for image recognition tasks.

In this study only 88 damaged façades were extracted, where the high overlap of manned aerial systems allowed to derive several image pairs and image sextuples per façade, per epoch. In this way it was possible to perform the experiments reported in this paper. This is an understudied subject, where usually the redundancy of aerial surveys is not fully used.

Although the image coverage of the area is relatively high, another limitation was given by the occlusions in dense urban areas: several buildings or part of them were almost invisible in the images. This is an intrinsic limitation of aerial-manned platforms with pre-defined flight patterns not tailored to decrease such occlusions. In this sense more careful flight plans and using UAV could attenuate this problem.

7. Conclusions

This paper assessed the image classification of façade damages using multi-temporal aerial oblique imagery. Six multi-temporal and three mono-temporal approaches were tested, following a binary classification approach using CNN. For this purpose, the only dataset (to the best of the authors' knowledge) available with pre- and post-event data was used for this analysis. Although the dataset is not optimal in terms of number of images and resolution, it has shown very encouraging results and good indications for the wide adoption of multi-temporal data in the assessment of catastrophic event damages.

The objective of this study was twofold: 1) determine the optimal framework to combine the multi-temporal image data within a CNN approach, and 2) investigate the improvement introduced by the use of the multi-view characteristics of aerial (manned) systems (extracting several image patches per façade and per epoch) in the image classification of façade damages. In this regard two main approaches were tested: 1) using image pairs by pairing every pre-event façade image patch to the corresponding post-event façade image patches, and 2) using image sextuples where three views per façade per epoch were considered.

An important element tested in this paper was the use of rectified façades instead of the original façade image patches. Regarding the original façade image patches, the best approach was to use image pairs and a 2-stream network (no shared convolutions) while using rectified façade image patches, the use of the image sextuples and shared convolutions was more advantageous. Given the rectification and registration of the façade image patches, considering three views per epoch only slightly improved over the approach considering image pairs. A study considering more data would need to be performed to assess if the network can learn not only inter epoch dependencies, but also to cope with different views of a given façade.

The multi-temporal approaches generally outperformed the mono-temporal ones. Large differences in the multi-temporal results were, however, visible according to the used network. The use of epoch-specific convolutions was preferable to single stream architectures, where both epochs inputs are concatenated together before being fed to the network. Epoch-specific feature information is in this way valuable for the image classification of façade damages. This was the case regardless of the use of original or rectified image patches as input, and regardless of the use of image pairs or image sextuples. However, while the best performing network using the original image pairs considered a 2-stream network without shared convolutions, this was not the case when using the rectified façade image patches where the 2-stream network sharing the convolutions (i.e. Siamese) was preferable.

While the multi-temporal approaches showed better performance, these would also need a specific framework to be prepared preemptively. These would include the aerial oblique flight covering the region, its processing to generate the 3D information, labeling of a portion of the façade images in damaged and non-damaged groups and the image classification of building façade damages. The aerial flight, human resources and material needed to have such framework established, need to be considered. Traditionally such façade damage mapping task is performed using lengthy and costly ground observation campaigns. The screening approach presented in this paper may be considered to attenuate both the length and cost of such ground campaigns. Given the processing times reported in this paper it can also be of use to first responders in the case post-event imagery is readily available.

Regarding the mono-temporal approaches, the network trained on image samples depicting debris and rubble piles was often not able to have a better score

than random guess (i.e. 50% accuracy). Hence, such networks trained with only rubble piles and debris from mostly nadir viewing imagery, are not transferable for façade cases where damage evidences are often different in image content but also in extent (e.g., small signs of spalling or cracks). The mono-temporal approach using damaged and non-damaged façade image patches performed better when trained from scratch; however, overall it behaved poorly.

A notable limitation of the approach presented here is its binary nature that precludes more nuanced damage assessment. In the disaster response phase, the location of partially and totally collapsed buildings is a priority. Hence, in such case the binary nature of the approach is not sufficient, since it considers several typologies of damage (from spalling to completely destroyed façades). More work is needed, based on more oblique multi-temporal image datasets, to move toward the classification of the different types of façade damages and their localization within the façade. Nonetheless, given the focus of this work on the specific façade damage detection, this approach could be performed in parallel with the already extensively reported methods in the literature to detect rubble piles and debris.

The dataset used was extremely challenging not only for the limited number of images (and facades) and the low resolution, but for the urban typology (historical city center) that introduced additional challenges. Several façades in the test area were impossible to extract given the often-narrow streets. This was further exacerbated by the use of an aerial (manned) system and its pre-defined flight pattern, which limited the data completeness in some narrow streets.

Disclosure statement

No potential conflict of interest was reported by the authors.

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.

ORCID

Diogo Duarte http://orcid.org/0000-0002-1599-8956
Francesco Nex http://orcid.org/0000-0002-5712-6902
Norman Kerle http://orcid.org/0000-0002-4513-4681
George Vosselman http://orcid.org/0000-0001-8813-8028

References

Armesto-González, J., B. Riveiro-Rodríguez, D. González-Aguilera, and M. T. Rivas-Brea. 2010. "Terrestrial Laser Scanning Intensity Data Applied to Damage Detection for Historical Buildings." *Journal of Archaeological Science* 37 (12): 3037–3047. doi:10.1016/j.jas.2010.06.031.

Audebert, N., B. Le Saux, and S. Lefèvre. 2017. "Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks." In *Computer Vision – ACCV 2016*, edited by S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, 180–196. Springer International Publishing: Cham. doi:10.1007/978-3-319-54181-5_12.

Audebert, N., B. Le Saux, and S. Lefèvre. 2018. "Beyond RGB: Very High Resolution Urban Remote Sensing with Multimodal Deep Networks." *ISPRS Journal of Photogrammetry and Remote Sensing* 140: 20–32. doi:10.1016/j.isprsjprs.2017.11.011.

Axelsson, P., 2000. "DEM Generation from Laser Scanning Data Using Adaptive TIN Models." International Archives of Photogrammetry and Remote Sensing. Presented at the XIX ISPRS Congress, Amsterdam, pp. 111–118.

Brunner, D., K. Schulz, and T. Brehm. 2011. "Building Damage Assessment in Decimeter Resolution SAR Imagery: A Future Perspective, In: Joint Urban Remote Sensing Event." *IEEE* 217–220. doi:10.1109/JURSE.2011.5764759.

Chen, Z., Y. Zhang, C. Ouyang, F. Zhang, and J. Ma. 2018. "Automated Landslides Detection for Mountain Cities Using Multi-temporal Remote Sensing Imagery." *Sensors* 18: 821. doi:10.3390/s18030821.

Clevert, D.-A., T. Unterthiner, and S. Hochreiter, 2016. "Fast and Accurate Deep Network Learning by Exponential Linear Units (Elus)." ICLR at San Juan, Puerto Rico

Curtis, A., and W. F. Fagan. 2013. "Capturing Damage Assessment with a Spatial Video: An Example of a Building and Street-scale Analysis of Tornado-related Mortality in Joplin, Missouri, 2011." *Annals of the Association of American Geographers* 103 (6): 1522–1538. doi:10.1080/00045608.2013.784098.

Cusicanqui, J., N. Kerle, and F. Nex. 2018. "Usability of Aerial Video Footage for 3D-scene Reconstruction and Structural Damage Assessment." *Natural Hazards and Earth System Sciences Discussions 1–23* 18 (6): 1583–1598. doi:10.5194/nhess-2017-409.

Daudt, R. C., B. Le Saux, A. Boulch, and Y. Gousseau. 2018. "Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks." Presented at the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Valencia, pp. 2115–2118. 10.1109/IGARSS.2018.8518015

Dell'Acqua, F., and P. Gamba. 2012. "Remote Sensing and Earthquake Damage Assessment: Experiences, Limits, and Perspectives." Proceedings of the IEEE, 100, 2876–2890. 10.1109/JPROC.2012.2196404

Dong, L., and J. Shan. 2013. "A Comprehensive Review of Earthquake-induced Building Damage Detection with Remote Sensing Techniques." *ISPRS Journal of Photogrammetry and Remote Sensing* 84: 85–99. doi:10.1016/j.isprsjprs.2013.06.011.

Duarte, D., 2020. "Debris, Rubble Piles and Façade Damage Detection Using Multi-resolution Optical Remote Sensing

Imagery (Phd)." University of Twente, Enschede, The Netherlands. 10.3990/1.9789036549400

Duarte, D., F. Nex, N. Kerle, and G. Vosselman. 2017. "Towards a More Efficient Detection of Earthquake Induced Facade Damages Using Oblique UAV Imagery." ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 93–100. 10.5194/isprs-archives-XLII-2-W6-93-2017at Bonn, Germany.

Duarte, D., F. Nex, N. Kerle, and G. Vosselman. 2018a. "Multi-resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks." *Remote Sensing* 10 (10): 1636. doi:10.3390/rs10101636.

Duarte, D., F. Nex, N. Kerle, and G. Vosselman, 2018b. "Satellite Image Classification of Building Damages Using Airborne and Satellite Image Samples in a Deep Learning Approach." ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 89–96. 10.5194/isprs-annals-IV-2-89-2018 at Riva del Garda, Italy.

Duarte, D., F. Nex, N. Kerle, and G. Vosselman. 2019. "Damage Detection on Building Façades Using Multi-temporal Aerial Oblique Imagery." *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2 (W5): 29–36. doi:10.5194/isprs-annals-IV-2-W5-29-2019.

Dubois, D., and R. Lepage. 2014. "Fast and Efficient Evaluation of Building Damage from Very High Resolution Optical Satellite Images." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 (10): 4167–4176. doi:10.1109/JSTARS.2014.2336236.

Fernandez Galarreta, J., N. Kerle, and M. Gerke. 2015. "UAV-based Urban Structural Damage Assessment Using Object-based Image Analysis and Semantic Reasoning." *Natural Hazards and Earth System Science* 15 (6): 1087–1101. doi:10.5194/nhess-15-1087-2015.

Freeman, H., and R. Shapira. 1975. "Determining the Minimum-area Encasing Rectangle for an Arbitrary Closed Curve." *Communications of the ACM* 18 (7): 409–413. doi:10.1145/360881.360919.

Gerke, M., and N. Kerle. 2011. "Automatic Structural Seismic Damage Assessment with Airborne Oblique Pictometry© Imagery." *Photogrammetric Engineering and Remote Sensing* 77 (9): 885–898. doi:10.14358/PERS.77.9.885.

Hamaguchi, R., A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka. 2017. *Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Images*. IEEE winter conference on applications of computer vision (WACV) at South Lake Tahoe, United States

He, K., X. Zhang, S. Ren, and J. Sun, 2016. "Deep Residual Learning for Image Recognition." CVPR at Las Vegas, United States

Hu, F., G.-S. Xia, J. Hu, and L. Zhang. 2015a. "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-resolution Remote Sensing Imagery." *Remote Sensing* 7 (11): 14680–14707. doi:10.3390/rs71114680.

Hu, W., Y. Huang, L. Wei, F. Zhang, and H. Li. 2015b. "Deep Convolutional Neural Networks for Hyperspectral Image Classification." *Journal of Sensors* 2015: 1–12. doi:10.1155/2015/258619.

Huang, G., Z. Liu, L. van der Maaten, and K. Q. Weinberger., 2017. "Densely Connected Convolutional Networks." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, pp. 2261–2269. 10.1109/CVPR.2017.243

Hussain, M., D. Chen, A. Cheng, H. Wei, and D. Stanley. 2013. "Change Detection from Remotely Sensed Images: From Pixel-based to Object-based Approaches." *ISPRS Journal of Photogrammetry and Remote Sensing* 80: 91–106. doi:10.1016/j.isprsjprs.2013.03.006.

Jiang, H., and N. Lu. 2018. "Multi-scale Residual Convolutional Neural Network for Haze Removal of Remote Sensing Images." *Remote Sensing* 10 (6): 945. doi:10.3390/rs10060945.

Jung, J., S.-H. Yun, D. Kim, and M. Lavalle. 2018. "Damage-mapping Algorithm Based on Coherence Model Using Multitemporal Polarimetric–interferometric SAR Data." *IEEE Transactions on Geoscience and Remote Sensing* 56 (3): 1520–1532. doi:10.1109/TGRS.2017.2764748.

Kampffmeyer, M., A. Salberg, and R. Jenssen, 2016. "Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks." CVPR at Las Vegas, United States.

Kerle, N., and R. R. Hoffman. 2013. "Collaborative Damage Mapping for Emergency Response: The Role of Cognitive Systems Engineering." *Natural Hazards and Earth System Science* 13 (1): 97–113. doi:10.5194/nhess-13-97-2013.

Khoshelham, K., S. Oude Elberink, and X. Sudan. 2013. "Segment-based Classification of Damaged Building Roofs in Aerial Laser Scanning Data." *IEEE Geoscience and Remote Sensing Letters* 10 (5): 1258–1262. doi:10.1109/LGRS.2013.2257676.

Koukal, T., C. Atzberger, and W. Schneider. 2014. "Evaluation of Semi-empirical BRDF Models Inverted against Multi-angle Data from a Digital Airborne Frame Camera for Enhancing Forest Type Classification." *Remote Sensing of Environment* 151: 27–43. doi:10.1016/j.rse.2013.12.014.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012. "ImageNet Classification with Deep Convolutional Neural Networks." NIPS at South Lake Tahoe, United States pp. 1105–1907.

Längkvist, M., A. Kiselev, M. Alirezaie, and A. Loutfi. 2016. "Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks." *Remote Sensing* 8 (4): 329. doi:10.3390/rs8040329.

Li, X., X. Chen, L. Liang, X. Chen, and L. Liang. 2012. "A New Approach to Collapsed Building Extraction Using RADARSAT-2 Polarimetric SAR Imagery." *IEEE Geoscience and Remote Sensing Letters* 9 (4): 677–681. doi:10.1109/LGRS.2011.2178392.

Liu, T., and A. Abd-Elrahman. 2018a. "Deep Convolutional Neural Network Training Enrichment Using Multi-view Object-based Analysis of Unmanned Aerial Systems Imagery for Wetlands Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 139: 154–170. doi:10.1016/j.isprsjprs.2018.03.006.

Liu, T., and A. Abd-Elrahman. 2018b. "Multi-view Object-based Classification of Wetland Land Covers Using Unmanned Aircraft System Images." *Remote Sensing of Environment* 216: 122–138. doi:10.1016/j.rse.2018.06.043.

Liu, T., A. Abd-Elrahman, A. Zare, B. A. Dewitt, L. Flory, and S. E. Smith. 2018. "A Fully Learnable Context-driven Object-based Model for Mapping Land Cover Using Multi-view Data from Unmanned Aircraft Systems." *Remote*

Sensing of Environment 216: 328–344. doi:10.1016/j.rse.2018.06.031.

Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation, In: CVPR." *IEEE* 3431–3440. doi:10.1109/CVPR.2015.7298965.

Lu, D., P. Mausel, E. Brondízio, and E. Moran. 2004. "Change Detection Techniques." *International Journal of Remote Sensing* 25 (12): 2365–2401. doi:10.1080/0143116031000139863.

Maggiori, E., Y. Tarabalka, G. Charpiat, and P. Alliez. 2017. "Convolutional Neural Networks for Large-scale Remote-sensing Image Classification." *IEEE Transactions on Geoscience and Remote Sensing* 55 (2): 645–657. doi:10.1109/TGRS.2016.2612821.

Murtiyoso, A., F. Remondino, E. Rupnik, F. Nex, and P. Grussenmeyer, 2014. "Oblique Aerial Photography Tool for Building Inspection and Damage Assessment." ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 309–313. 10.5194/isprsarchives-XL-1-309-2014. Denver, United States

Nex, F., D. Duarte, A. Steenbeek, and N. Kerle. 2019. "Towards Real-Time Building Damage Mapping with Low-Cost UAV Solutions." *Remote Sensing* 11 (3): 287. doi:10.3390/rs11030287.

Nogueira, K., O. A. B. Penatti, and J. A. Dos Santos. 2017. "Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification." *Pattern Recognition* 61: 539–556. doi:10.1016/j.patcog.2016.07.001.

Persello, C., and A. Stein. 2017. "Deep Fully Convolutional Networks for the Detection of Informal Settlements in VHR Images." *IEEE Geoscience and Remote Sensing Letters* 14 (12): 2325–2329. doi:10.1109/LGRS.2017.2763738.

Singh, A. 1989. "Review Article Digital Change Detection Techniques Using Remotely-sensed Data." *International Journal of Remote Sensing* 10 (6): 989–1003. doi:10.1080/01431168908903939.

Sui, H., J. Tu, Z. Song, G. Chen, and Q. Li. 2014. "A Novel 3D Building Damage Detection Method Using Multiple Overlapping UAV Images." *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL–7* XL-7: 173–179. doi:10.5194/isprsarchives-XL-7-173-2014.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. *Going Deeper with Convolutions*. CVPR at Boston, United States

Tewkesbury, A. P., A. J. Comber, N. J. Tate, A. Lamb, and P. F. Fisher. 2015. "A Critical Synthesis of Remotely Sensed Optical Image Change Detection Techniques." *Remote Sensing of Environment* 160: 1–14. doi:10.1016/j.rse.2015.01.006.

Tu, J., H. Sui, W. Feng, K. Sun, C. Xu, and Q. Han. 2017. "Detecting Building Façade Damage from Oblique Aerial Images Using Local Symmetry Feature and the Gini Index." *Remote Sensing Letters* 8 (7): 676–685. doi:10.1080/2150704X.2017.1312027.

United Nations. 2009. *2009 UNISDR Terminology on Disaster Risk Reduction*. Geneva, Switzerland: United Nations International Strategy for Disaster Reduction.

United Nations. 2015. *INSARAG Guidelines, Volume II: Preparedness and Response, Manual B: Operations*. United Nations Office for the Coordination of Humanitarian Affairs.

Vetrivel, A., D. Duarte, F. Nex, M. Gerke, N. Kerle, and G. Vosselman. 2016. "Potential of Multi-temporal Oblique Airborne Imagery for Structural Damage Assessment." *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences III–3* III-3: 355–362. doi:10.5194/isprsannals-III-3-355-2016.

Vetrivel, A., M. Gerke, N. Kerle, F. Nex, and G. Vosselman. 2017. "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 Journal of Photogrammetry and Remote Sensing*. doi:10.1016/j.isprsjprs.2017.03.001.

Vo, N. N., and J. Hays. 2016. "Localizing and Orienting Street Views Using Overhead Imagery." In *Computer Vision – ECCV 2016*, edited by B. Leibe, J. Matas, N. Sebe, and M. Welling, 494–509. Amsterdam: Springer International Publishing. doi:10.1007/978-3-319-46448-0_30.

Volpi, M., and D. Tuia. 2017. "Dense Semantic Labeling of Subdecimeter Resolution Images with Convolutional Neural Networks." *IEEE Transactions on Geoscience and Remote Sensing* 55 (2): 881–893. doi:10.1109/TGRS.2016.2616585.

Vosselman, G. 2012. "Automated Planimetric Quality Control in High Accuracy Airborne Laser Scanning Surveys." *ISPRS Journal of Photogrammetry and Remote Sensing* 74: 90–100. doi:10.1016/j.isprsjprs.2012.09.002.

Wallemacq, P., and R. House. 2018. "*Economic Losses, Poverty & Disasters: 1998-2017*" Centre for Research on the Epidemiology of Disasters, United Nations Office for Disaster Risk Reduction.

Wang, Q., X. Zhang, G. Chen, F. Dai, Y. Gong, and K. Zhu. 2018. "Change Detection Based on Faster R-CNN for High-resolution Remote Sensing Images." *Remote Sensing Letters* 9 (10): 923–932. doi:10.1080/2150704X.2018.1492172.

Xia, G.-S., Z. Wang, C. Xiong, and L. Zhang. 2015. "Accurate Annotation of Remote Sensing Images via Active Spectral Clustering with Little Expert Knowledge." *Remote Sensing* 7 (11): 15014–15045. doi:10.3390/rs71115014.

Yu, F., and V. Koltun, 2016. "Multi-scale Context Aggregation by Dilated Convolutions." ICLR at San Juan, Puerto Rico.

Yu, F., V. Koltun, and T. Funkhouser, 2017. "Dilated Residual Networks." CVPR at Honolulu, Hawaii, United States.

Zhang, C., S. Wei, S. Ji, and M. Lu. 2019. "Detecting Large-scale Urban Land Cover Changes from Very High Resolution Remote Sensing Images Using CNN-nased Classification." *ISPRS International Journal of Geo-Information* 8 (4): 189. doi:10.3390/ijgi8040189.

Zhao, J., X. Xie, X. Xu, and S. Sun. 2017. "Multi-view Learning Overview: Recent Progress and New Challenges." *Information Fusion* 38: 43–54. doi:10.1016/j.inffus.2017.02.007.

Zhuo, X., F. Fraundorfer, F. Kurz, and P. Reinartz. 2019. "Automatic Annotation of Airborne Images by Label Propagation Based on a bayesian-CRF Model." *Remote Sensing* 11 (2): 145. doi:10.3390/rs11020145.