Detection of Disaster-Affected Cultural Heritage Sites from Social Media Images Using Deep Learning Techniques

PAKHEE KUMAR, IMT School for Advanced Studies Lucca, Italy
FERDA OFLI and MUHAMMAD IMRAN, Qatar Computing Research Institute, HBKU, Qatar
CARLOS CASTILLO, Universitat Pompeu Fabra, Spain

This article describes a method for early detection of disaster-related damage to cultural heritage. It is based on data from social media, a timely and large-scale data source that is nevertheless quite noisy. First, we collect images posted on social media that may refer to a cultural heritage site. Then, we automatically categorize these images according to two dimensions: whether they are indeed a photo in which a cultural heritage resource is the main subject, and whether they represent damage. Both categorizations are challenging image classification tasks, given the ambiguity of these visual categories; we tackle both tasks using a convolutional neural network. We test our methodology on a large collection of thousands of images from the web and social media, which exhibit the diversity and noise that is typical of these sources, and contain buildings and other architectural elements, heritage and not-heritage, damaged by disasters as well as intact. Our results show that while the automatic classification is not perfect, it can greatly reduce the manual effort required to find photos of damaged cultural heritage by accurately detecting relevant candidates to be examined by a cultural heritage professional.

CCS Concepts: • **Computing methodologies → Scene understanding**; **Image representations**; **Object recognition**; **Neural networks**; • **Applied computing → Architecture (buildings)**;

Additional Key Words and Phrases: Cultural heritage sites, social media, damage assessment, deep learning

1 INTRODUCTION

Cultural heritage resources are finite, scarce, non-renewable, and valuable [80]. They represent our collective memory, shape our identity, and also drive the economy [49, 50, 82]. These resources are globally under

immense threat in present times due to natural and human-induced disasters. The increased frequency and severity of disasters affecting cultural heritage [85] has increased the international awareness toward protection and conservation of cultural heritage [13, 91]. It also points toward the need for an organized response in such cases by utilizing efficient tools.

Social networking sites, particularly Twitter, have been acknowledged as an efficient communication tool for disaster management due to its instantaneous nature [65]. Twitter has been used to disseminate news, support the immediate disaster response, and track efforts of relief and reconstruction. Consequently, developing efficient systems to harness and use real-time information from social media to help relief activities for humanitarian response in disasters has been a priority area for researchers [16, 64]. Researchers have developed methods for timely detection of events [7, 9, 10, 74], automatic extraction of information from postings [46, 95], and automatic classification of images [4], among many other tasks. Most works have focused on extracting urgent needs from the affected populations, while in comparison applications for detecting and evaluating damage to cultural heritage using social media data have not been studied.

This article aims to bridge this gap by describing a method to automatically detect images of cultural heritage sites, particularly images depicting damage.

The need for this automation arises from the quantity and variety of images posted on social media. Firstly, the amount of images posted on social media is enormous. According to Meeker and Wu, approximately 1.8 billion images are shared daily on social media platforms [63]. The quantity of images posted on social media during disasters is even larger [16]. Secondly, this enormous amount of images posted during disasters contain irrelevant and redundant content, including images not related to the disaster, duplicate images, and "memes," among many others [68]. In fact, the images of cultural heritage sites are a small proportion of the total images: in our datasets from social media during disasters, we estimate that less than 10% of images shared might be about heritage sites. Nevertheless, these images are an unparalleled source of information to detect in near real-time if a cultural heritage site has been affected by a disaster.

Considering the enormous amount of relevant and irrelevant images, manual annotation of each image might not be feasible. In this work, we propose to use supervised machine learning techniques, specifically deep neural networks, to automatically identify heritage sites and detect if they show any damage. The models trained on images found through Google Image Search are evaluated on a real-world disaster dataset collected from Twitter. The automatic classification methodology discussed in this article provides a helpful tool to support the work of heritage preservation professionals. By examining a relatively small set of potentially relevant candidate images extracted by automatic means from a much larger collection, professionals are able to understand the extent of damage to cultural heritage without necessarily being on site, saving time and resources. Given the immediacy of social media, the tool is particularly useful for preliminary analysis, and therefore, can help toward organizing the response by identifying priority areas.

There are four main contributions of this article:

(1) A methodology for collecting, annotating, and learning classifiers to identify heritage sites' images
(2) An evaluation of this methodology performed on a real-world dataset taken from a disaster event
(3) A corpus[1] of annotated images into heritage vs. not-heritage sites with/without damage labels
(4) A lexicon of heritage-related keywords for social media filtering tasks

The rest of this article is structured in five parts. Section 2 conceptually frames this research, particularly linking it to similar techniques used in the heritage context and beyond. Section 3 briefly describes the methodology adopted for this work. Section 4 discusses the process of data collection and annotation. Lastly, Section 5 describes the experiments and results. The article concludes with possibilities of future work in Section 6.

---

[1]https://crisisnlp.qcri.org/heritage.

2 RELATED WORK

2.1 Images of Disaster and Cultural Heritage in Social Media

There is a growing interest among researchers to study images about disasters posted on social media. Images have been studied from many different perspectives including typology analysis [41], spatial and temporal pattern analysis [39], and ethics [35, 72], among others. Faulkner et al. present a good overview of the current research on social media images by using three methodological approaches: large-scale image analysis, working with images at different scales, and in-depth qualitative analysis of images [25]. Further, they adopt the three methodological approaches to analyze the case of Alan Kurdi[2] in order to draw different insights from the same dataset. Bozdag and Smets's qualitative study using small data concluded that the images of Alan Kurdi did not cause a major shift in common discourses and representations [14]. Similarly, Kharroub and Bas's analysis of 518 images circulated during the 2011 Egyptian revolution revealed more efficacy-eliciting than emotionally arousing content posted by Egyptian users [53]. Hjorth and Burgess analyzed the 100 most retweeted images during the Queensland flood to understand the genres and resonating themes in images [38]. Vis et al.'s exploratory study of the images tweeted during the 2011 UK riots also considers different types of images posted during the event [93]. Seo identified themes and frames prominently appearing in a total of 243 Twitter images posted by the Twitter accounts of the Israel Defense and Hamas's Alqassam Brigades during the Israeli-Hamas Conflict [76]. Additionally, a few studies focus on the analysis of self-portraits (i.e., "selfies") posted on social media during disasters [36, 44].

The cultural heritage domain use the images on social media for two main purposes: (i) enable users to interact with an already existing image database, and (ii) create new databases of (heritage) images on social media. The US Library of Congress uses photo sharing platform Flicker to enable users to interact with old photographs [81]. Other cultural institutions in the US such as The Smithsonian carried out similar initiatives [48, 52]. In contrast, Terras investigated the growing trend of the creation of digital images of cultural and heritage materials by amateurs on Flickr [87]. Garduño Freeman studied the public engagement with the world heritage site Sydney Opera House on Flicker and argued that such socio-visual practices themselves constitute an intangible heritage [29]. A number of studies focus on cultural heritage institution's use of image-based social media such as Flickr and Instagram to understand the content created by the institutions, the relation between audience and institution, among other topics [47, 60]. To the best of our knowledge, no prior study deals with the analysis of images depicting cultural heritage circulated on social media during disasters.

2.2 Automated Processing of Images from Heritage Sites

Image processing techniques have been used in the cultural heritage context for various purposes. For example, Hurtut et al. introduced a method for the analysis of the pictorial content of line drawings using the geometrical information of stroke contours [43]. They showed that the proposed method could be used successfully for the indexing of line drawings in a retrieval framework. In another example, Makridis and Daras presented a technique for automatic archaeological sherd classification based on a bag-of-visual-words representation of local color and texture information and discriminative feature selection [61]. Amato et al. defined a pipeline that combined a convolutional neural network with Fisher vector features for visual recognition of ancient inscriptions. Their study suggested that these features could be effective in visual retrieval of other types of objects related to cultural heritage such as landmarks and monuments [6]. Can et al. studied visual analysis of Maya glyphs using both handcrafted and data-driven shape representations in a bag-of-words-based pipeline [15]. Similarly, Hu et al. proposed a system for automatic extraction of hieroglyph strokes from images of degraded ancient Maya codices via a region-based image segmentation framework [40]. According to their

---

[2]A Syrian boy, age 3, who drowned in the Mediterranean Sea in 2015 while his family, escaping from the Syrian war and ISIS, attempted to reach Greece from Turkey.

experimental results, automatically extracted glyph strokes achieved comparable retrieval results to those obtained using glyphs manually segmented by epigraphers.

Focusing more on architectural heritage, Shalunts et al. presented an approach based on clustering and learning of local features to classify the architectural style of facade windows [78]. Mathias et al. used features extracted by a steerable pyramid of Gabor filters to train a Support Vector Machine for automatic architectural style recognition [62]. To tackle the same problem, Chu and Tsai proposed a higher-level feature representation that takes into account spatial relationships between local features to identify repetitive subgraphs as visual patterns in an image [18]. Furthermore, Goel et al. explored the utility of mining characteristic configurations of low-level discriminative features in categorizing different architectural styles and used them for improving classification performance [31]. Alternatively, Oses and Dornaika presented a semi-automatic approach for delineation of the masonry to classify architectural style [71], whereas Zhang et al. introduced blocklets that capture the morphological characteristics of buildings and developed an architectural style recognition model based on hierarchical sparse coding of blocklets [98]. Xu et al., on the other hand, adopted Deformable Part-based Models to capture the morphological characteristics of basic architectural components and proposed Multinomial Latent Logistic Regression for architectural style classification [94]. Amato et al. combined k-nearest neighbor classification and landmark recognition techniques to tackle the problem of monument recognition in images efficiently [5]. More recently, Llamas et al. explored deep learning-based techniques, specifically convolutional neural networks, for the classification of architectural heritage images into one of the 10 types of architectural elements of heritage buildings [58]. However, their dataset consists mostly of churches and religious temples. More importantly, they do not consider images from any damage or disaster context. In contrast, in this article, our goal is to analyze the visual content of images to determine whether they show any type of cultural heritage, even when the image is taken potentially in some damage or disaster context.

2.3 Detection of Images Showing Damaged Structures

There has been a significant increase in the use of image analysis techniques for automatic damage assessment in the last couple of decades. Most of these studies can be divided into two groups based on the type of data and domain knowledge they use.

The first group of studies corresponds mainly to the *remote sensing domain* and mostly rely on the analyses of images obtained from satellites, aircrafts, and unmanned aerial vehicles (UAVs). Early examples include detection of damaged or collapsed buildings using aerial photographs collected from earthquake-hit regions [89, 90]. Similarly, Pesaresi et al. investigated rapid damage assessment of built-up structures using satellite data in tsunami-affected areas [73]. In order to produce comprehensive per-building damage scores, Fernandez Galarreta et al. studied UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning [26], whereas Attari et al. explored fine-grained segmentation of UAV imagery based on deep learning techniques for damage assessment [8]. Alternatively, Vetrivel et al. combined multiple kernel learning with 3D point cloud features derived from high resolution oblique aerial images to detect disaster damage [92]. Likewise, Cusicanqui et al. investigated the usability of aerial video footage for 3D scene reconstruction and structural damage assessment [21]. To maximize their data utilization, Kakooei and Baleghi [51] and Duarte et al. [24] explored fusion of multiple data sources such as satellite, aircraft, and UAVs for automatic disaster damage assessment.

The second group of studies includes relatively recent work in the *crisis informatics domain* and rely mostly on the analyses of ground-level images collected from online social media platforms during disasters [4, 11, 68]. Early examples specific to damage assessment tasks are presented by Lagerstrom et al. [55] and by Daly and Thom [23] where both studies analyzed social media data in a binary image classification setting for fire/not-fire detection. Later, Nguyen et al. investigated a more generic solution to classify disaster images according to damage severity using convolutional neural networks [69]. Similarly, Li et al. proposed a method based on class activation mapping to localize and quantify damage in social media images posted during disasters [57]. Taking
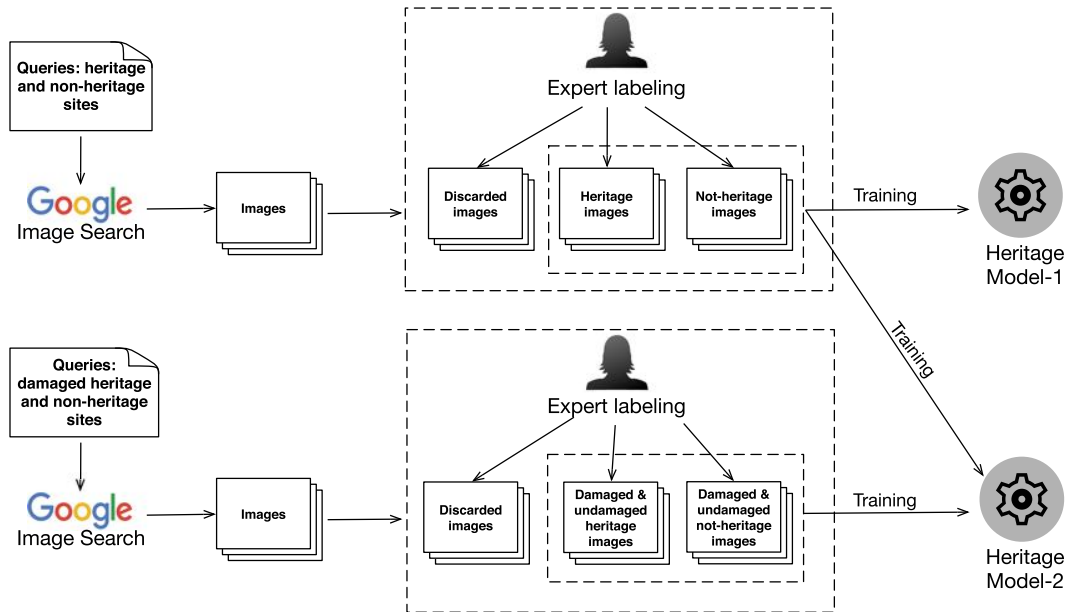
Fig. 1. Overview of methodology.

a step further, Li et al. explored a domain adaptation approach to identify disaster damage images during an emergent event when there is scarcity of labeled data [56]. To advance the state-of-the-art in this area, Alam et al. [3] and Mouzannar et al. [67] recently introduced multimodal datasets comprising both social media text messages and images. Furthermore, Mouzannar et al. defined a deep learning approach to identify damage images in their dataset [67]. Inspired by these recent advancements, Alam et al. developed an image processing pipeline to extract meaningful information from social media images during a crisis situation, including damage severity assessment [2]. In this study, we ran the images in our heritage image datasets through Alam et al.'s system to perform the damage assessment task. It is important to note that our dataset, in contrast to previous works, focuses on elements from cultural heritage sites that often look old or aged. This makes the damage assessment task more challenging than the aforementioned studies, which use all kinds of images; indeed, the vast majority of images processed in previous works to identify damaged structures are not images of heritage sites.

3   METHODOLOGY OVERVIEW

The methodology adopted for this research has the following steps:

(1) Definition of elements and categories of interest
(2) Data collection
(3) Data filtering and annotation
(4) Construction of classification models

Figure 1 outlines the overall methodology.

Definition of elements and categories of interest. The elements we want to classify are images embedded in social media postings. The category of interest corresponds to all images that show damage to a heritage site. This is the intersection of two broader categories: images depicting heritage sites and images depicting damaged structures.

Fig. 2. Images in our collection corresponding to heritage sites (left) and not-heritage sites (right).

Firstly, a balanced list containing the names of cultural heritage and not-cultural heritage sites was created. Given the inherent complexity of cultural heritage, we considered the legal protection status as the criteria for defining cultural heritage and not-cultural heritage. At the international level, the scope of the term "cultural heritage" has broadened since the adoption of the Venice Charter in 1964 [45]. It is now applicable to individual buildings, groups of buildings, historical areas, towns, sites, environments, social factors, and intangible entities (intangible heritage). It also includes artifacts, artworks, practices, and the like. At the national level, finer terminologies of *heritage* are not standardized; therefore, no uniformity exists between countries [1]. Moreover, researchers have argued that heritage is an inherently complex phenomenon and can contain conflicting meanings [33]. Acknowledging these complexities, we decided to limit our dataset to the legally protected (either by national or local governments) cultural heritage.

The cultural heritage list included archaeological sites, monuments, cultural landscapes, museums, galleries, libraries, and artifacts in urban space. We tried to create a list that was visually and geographically varied in terms of period (ancient to modern), material, and construction. The not-cultural heritage lists also included buildings and artifacts in urban space. The lists of cultural heritage and not-cultural heritage sites are provided in Tables 10 and 11, respectively, in Appendix Section A. We must acknowledge that defining heritage is always an ongoing process, depending on what is valuable to people in a given place and time. Indeed, there is even de-listing of protected heritage buildings in some countries. Therefore, the list of heritage sites used as training data for the automatic classifier needs to be updated regularly to maintain the quality of the results.

Data collection (Section 4.1). Google Image Search was used to construct two datasets of images. The first dataset corresponds to images of heritage and not-heritage sites. Figure 2 shows examples of cultural heritage and not-cultural heritage from our list. The second dataset corresponds to damaged scenes from both heritage and not-heritage sites.

Data filtering and annotation (Section 4.2). The underlying problem of online images, whether on social media or Google, is that it contains many irrelevant or unusable images. In this study, the irrelevant or unusable images were primarily the ones where heritage was not the primary subject of the image or images, which were edited to an extent that the original context was significantly altered. Figure 3 shows some irrelevant images in our dataset. Firstly, these irrelevant images were removed, as explained in Section 4.2.1 in depth. Secondly, the remaining images were annotated using the following criteria: heritage vs. not-heritage and damaged heritage (Section 4.2.2) vs. not-heritage (Section 4.2.3). Both of the tasks were carried out by the lead author of this article, as shown in Figures 1 and 6.

Construction of classification models (Section 5). We built two different heritage classifiers using the labeled data annotated by our expert. First, we used only the images collected without any damage queries to train a

Fig. 3. Images from Google that could not be used for training the classifier.

classifier as shown in Figure 1 (top). Second, we used all images collected both with and without damage queries to train another heritage classifier (Figure 1, bottom). The performance of both classifiers is evaluated on the dataset collected during the 2015 Nepal earthquake (Section 5.3).

4 DATA COLLECTION AND ANNOTATION

In this section, we discuss our data collection and annotation details.

4.1 Cultural Heritage and Not-Cultural Heritage Images

We selected 92 cultural heritage sites around the world and downloaded their images from Google. The list includes sites related to architectural heritage, archaeology, monuments, cultural landscapes, museums, galleries, libraries, and art in urban space. We sought to make the list geographically, period (ancient to modern), material and construction-wise, and visually representative. Since we treat the detection of heritage sites as a binary classification task, we also create another list containing built structures (i.e., buildings and sites), which look somewhat similar to heritage sites, but, officially, they are not designated as cultural heritage. Selecting not-cultural heritage sites is a difficult task, given the ever-expanding boundaries of cultural heritage. Keeping in mind the protection criteria, a list of 32 not-cultural heritage sites was carefully curated to be geographically and visually representative. Interestingly, some of the buildings in this list are iconic buildings that are not protected. The complete lists containing the selected sites related to heritage and not-heritage are provided in Tables 10 and 11 in Appendices Section A. Figure 4 shows all the selected sites for both *heritage* and *not-heritage* categories on a map.

We downloaded approximately 100 images of each *heritage* and *not-heritage* site from Google image search using the heritage site name as a query. The image search criteria needed to be robust to yield better results. Some of the site names had more risk of yielding bad results. For instance, image search criterion for the Walkie-Talkie building in London was *Walkie-Talkie London*, as the possibility of a bad result was higher if *London* was not included in the search query.

In addition to the images that show heritage sites that are potentially undamaged, we searched for images of the heritage sites showing some damage. For this purpose, our query consists of the heritage site name combined with two keywords (i.e., "*damage*" and "*destroyed*") separately. In total, we were able to download 13,333 images from Google.

It is important to note that the proportion of cultural heritage images in social media streams, especially during disasters, is minuscule, which makes the detection task inherently imbalanced. In such situations, it is not ideal to leave the minority class (i.e., heritage) so small to avoid training machine learning models that trivially predict the majority class (i.e., not-heritage) all the time. Therefore, we recommend that this type of imbalance reduction

Fig. 4. Map showing locations of heritage and not-heritage sites.

Table 1. Filtering and Annotation Results for *Heritage* vs. *Not-Heritage* Annotation
of Images Found Using Google Image Search

Dataset	Removed images	Labeled as Heritage	Labeled as Not-heritage
Images found using heritage/not-heritage queries	2,974	6,612	2,266
Images found using damaged heritage queries	78	836 (447)	567
Total	3,052	7,448	2,833

The number in parentheses represents the number of damaged heritage images.

is done, i.e., to over sample from the minority class, ultimately to strengthen the performance of the classifier in the minority class (see Table 1).

4.2 Data Filtering and Annotation

4.2.1 Data Filtering. Many images, which were collected from publicly available websites using Google Image Search, are not useful for training an automatic classifier and were thus removed. Specifically, images with one of the following issues were removed: images that are significantly edited, images where a heritage site is merely a backdrop and not the main subject (e.g., selfies), images that are covered almost entirely with text, 3D reconstruction or 3D models of sites, paintings of heritage sites, memes, architectural plans and sections of heritage sites, sketches, maps, images in which contextual information is missing (e.g., a close-up photograph of a stone in a building), and images of replicas unless it has a protected status. Table 1 shows the results of the filtering task. Figure 3 shows a few images that were removed as a result of manual filtering. The remaining images are used to perform two annotation tasks as described next.

4.2.2 Heritage vs. Not-Heritage Annotation. This annotation task aims to identify whether an image contains a heritage site or not. The lead author of this article (a domain expert) labeled 13,333 images as *heritage* (which depicts a heritage site) and *not-heritage* (which does not depict a heritage site) using separate folders on a shared drive. The first row in Table 1 shows the results of the filtering and the heritage annotation tasks for images that were collected without damage queries.

4.2.3 Damaged Heritage vs. Not-Heritage Annotation. This annotation task aims to determine whether an image having a heritage site shows any sign of damage to the site or not. This task was also carried out by the lead author of this article using separate folders on a shared drive. The quantification of the scale of damage is a subjective task; hence, we follow the annotation scheme described in the literature [69], which defines the damage concept in three categories: (i) SEVERE damage, (ii) MILD damage, (iii) NO damage. However, in this work, we merged the SEVERE and MILD classes to a single class named "Damage." Table 1 shows the results of the filtering and the damage heritage annotation tasks for images that were collected with damage queries. Images collected from Google using the damage queries contain both heritage sites showing some damage and sites without any damage. The number of heritage sites with some damage are shown in parentheses in the second row of Table 1.

5 EXPERIMENTAL RESULTS OF AUTOMATIC CLASSIFICATION

In this section, we describe our experiments and present our results.

5.1 Classification Approach

We considered various alternative approaches ranging from more traditional techniques such as bag-of-visual-word models to more advanced deep learning techniques such as convolutional neural networks. Eventually, we decided to use a deep learning-based solution since the state-of-the-art performance in many computer vision tasks are achieved by deep learning models [37, 54, 77, 79, 84] that leverage on large-scale datasets such as ImageNet [75] and Places [99].

In a nutshell, deep learning models, convolutional neural networks (CNNs) in particular, learn low-, medium-, and high-level features and classifiers in an end-to-end fashion to optimize on the target prediction task directly from raw data [97]. For example, the lower layers of deep CNN architectures correspond to a representation suitable for low-level vision tasks while the higher layers represent more domain-specific information [32], and, hence, eliminate the need for hand-crafted features like Scale Invariant Feature Transform (SIFT) [59] or Histogram of Oriented Gradients (HOG) [22].

More importantly, the features learned in deep convolutional networks have been shown to be transferable and quite effective when used in other visual recognition tasks [30, 96], particularly when training samples are limited and learning a successful deep model is not feasible due to over-fitting. For instance, Nguyen et al. show the success of this transfer learning approach for damage assessment tasks performed on disaster images collected from social media [69]. Considering that we also have limited training examples, we adopt a transfer learning approach for the heritage classification problem.

Our heritage classification system is composed of two stages: (i) deep feature extraction and (ii) training a heritage/not-heritage classification model, as illustrated in Figure 5. In the deep feature extraction stage, each image from the training set is simply fed as input to a deep convolutional neural network (CNN) that is pre-trained on the ImageNet dataset, which has over 1.2M images and 1,000 categories [75]. The features extracted from the penultimate layer of the network are then used to represent the input image. Then, in the second stage, these deep features are used to construct the desired heritage classification model.

In this study, we experiment with a number of well-known CNN architectures in combination with a variety of classification algorithms. The CNN architectures used in the experiments include VGG16 [79], ResNet50 [37], DenseNet121 [42], InceptionResNetV2 [83], Xception [17], and NASNetLarge [100], whereas the classification
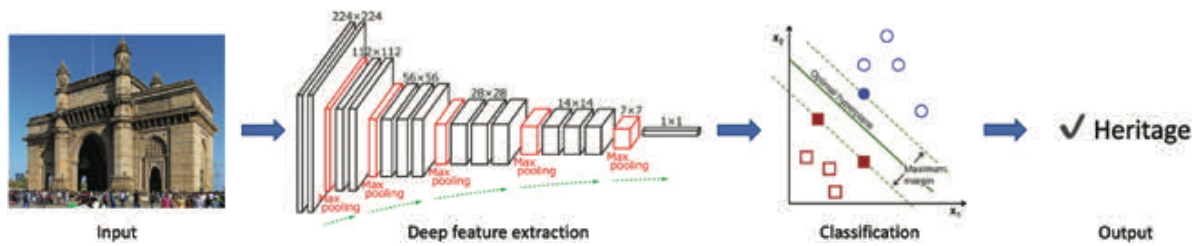
Fig. 5. Overview of the heritage classification system.

Table 2. Training/Test Set Split by Site (80:20 Ratio)

	Training Set-1		Training Set-2		Test Set	
	Sites	Images	Sites	Images	Sites	Images
Heritage	69	5,376	73	6,075	19	1,373
Not-heritage	25	1,869	25	2,380	7	453
Total	94	7,245	98	8,455	26	1,826

algorithms employed in the experiments comprise Logistic Regression [20], Support Vector Machines [19], Random Forests [88], and AdaBoost [28]. All the experimental results achieved by different network architectures and classification algorithms are presented in Tables 12–15 in Appendix Section C. Overall, DenseNet121 and NASNetLarge features seem to yield slightly better results than other feature types. And, in terms of algorithms, Logistic Regression and Support Vector Machines seem to perform better than Random Forests and AdaBoost.

For brevity, we hereinafter discuss only the results achieved by the model trained using the Logistic Regression algorithm on DenseNet121 features, as this is the best performing method in most of the experiments for our problem setup—although, in other settings, a different method may perform better.[3]

5.2 Heritage/Not-Heritage Classifier Training

In our dataset, we have 7,448 images from 92 heritage sites and 2,833 images from 32 not-heritage sites. In order to create disjoint training and test sets, we follow a site-based data split approach. That is, 80% of the heritage sites (i.e., 73 out of 92) are chosen at random and all images (i.e., 6,075) belonging to these sites are assigned to the training set (i.e., Training Set-2 in Table 2). Then, all images (i.e., 1,373) belonging to the remaining 20% of the heritage sites (i.e., 19 out of 92) are assigned to the test set (i.e., Test Set in Table 2). We follow the same approach to distribute images from not-heritage sites into the training and test sets.

To investigate the benefits of having images with damage context while training our heritage classifier, we create another training set (i.e., Training Set-1 in Table 2) where we ablate from Training Set-2 those images collected by *heritage-sites-with-damage* queries. In other words, Training Set-1 is a subset of Training Set-2 where images in Training Set-1 do not show any damage content. The resulting data split is summarized in Table 2. It is important to note that we opt for site-based data split rather than image-based data split to obtain models with better generalization capability on new images from previously-unseen sites.

Heritage Model-1. In this first scenario, we train a heritage classifier using only the images collected by heritage site queries with *no* damage keywords (i.e., Training Set-1 in Table 2).

---

[3]The DenseNet121 network consists of 121 layers and around 8 million weight parameters [42]. We choose the penultimate layer as our 1,024-dimensional deep feature extractor.

Table 3. Confusion Matrices of the Heritage Classifiers

			Classified as	
			Heritage	Not-heritage
Heritage Model-1	Actual label	Heritage	1,193	180
		Not-heritage	113	340
Heritage Model-2	Actual label	Heritage	1,174	199
		Not-heritage	74	379

Table 4. Performance Comparison of the Heritage Classifiers

	Precision		Recall		F1-score	
	Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Heritage Model-1	0.91	0.65	0.87	0.75	0.89	0.70
Heritage Model-2	0.94	0.66	0.86	0.84	0.90	0.74

Heritage Model-2. In the second scenario, we train a heritage classifier using all of the images collected by heritage sites queries both with and without damage keywords (i.e., Training Set-2 in Table 2).

In addition to a confusion matrix, which displays the number of correctly and incorrectly categorized instances on each class, we use three standard performance metrics for classification tasks. Precision (positive predictive value) is the probability that an item classified automatically into a class actually belongs to that class. Recall (or sensitivity) is the probability that an item that actually belongs to a class is classified automatically as such. The F1-score is the harmonic mean of precision and recall, and is one of several metrics that can be used to summarize them into a single number.

Results are shown in Tables 3 and 4. Confusion matrices (Table 3) for both models are dominated by the diagonal, meaning that heritage sites are more likely to be classified as such than as not-heritage. Performance in terms of precision and recall (Table 4) shows precision above 0.9 and recall above 0.8 for the heritage class. In practice, this means that an image automatically detected as a heritage will be, indeed, heritage at least 9 out of 10 times; and that images of heritage will be found by the classifier at least 8 out of 10 times. Overall, we do not observe much performance difference between the two heritage models on the Google images test set.

5.3 Case Study: 2015 Nepal Earthquake (SMERP Workshop Dataset)

We now present the results of our case study in a real-world scenario where we evaluate the performance of both of our heritage classifiers as well as an off-the-shelf damage assessment model of Alam et al. [2] on a Twitter dataset collected during the 2015 Nepal earthquake (i.e., Exploitation of Social Media for Emergency Relief and Preparedness (SMERP) Workshop Dataset [66]). As an alternative baseline, we also consider a lexicon-based model to analyze Twitter text messages for the heritage classification task. Figure 6 illustrates our case study design.

5.3.1 Data Filtering and Annotation. A dataset containing images of damaged heritage sites, extracted from social media, is essential to evaluate the proposed approach. We use images posted on Twitter during the 2015 Nepal earthquake, an event that damaged a large number of heritage sites in this country. Specifically, we use the SMERP Workshop Dataset [66], which contains 6,529 images collected after the Nepal Earthquake in 2015. The tweets in this dataset were collected using the keywords "Nepal earthquake" and "Nepal quake." It is evident from the keywords that this dataset was not curated for heritage purposes. Nevertheless, the dataset consists of information regarding heritage damaged due to the disaster. These 6,529 images are annotated manually using Nvivo, a qualitative data-analysis software by our expert for heritage and damage severity classification tasks. At
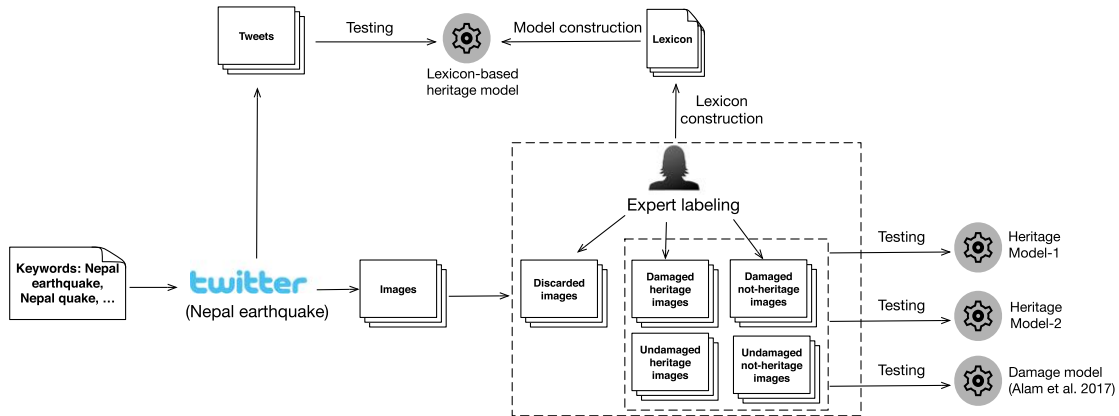
Fig. 6. Case study design and testing.

Table 5. Heritage and Damage Annotation Results for the SMERP Dataset

	Heritage	Not-heritage	Total
Damage	377	1,445	1,822
No-damage	110	4,388	4,498
Total	487	5,833	6,320

the end of this manual annotation process, there are 6,320 images labeled with heritage and damage categories, excluding the images labeled as "maybe_heritage" or "dont_know" as well as the images with multiple heritage or damage labels. All of these 6,320 images are treated as test images in our case study. Table 5 summarizes the results of both heritage and damage annotation tasks. Figure 7 shows a few images with and without damage. Moreover, the textual content associated with these 6,320 images (i.e., tweet text) is used to test our lexicon-based classifier, which we describe next in detail.

5.3.2 Baseline Construction. To set a baseline, we developed a lexicon consisting of 176 terms covering heritage-related concepts such as *museum*, *temple*, *monuments*. As a domain expert, the lead author of this study manually curated the lexicon. The full lexicon is provided in Appendix Section B. The lexicon terms were then used to categorize tweets from our case study event (i.e., 2015 Nepal earthquake), as shown in Figure 6. Specifically, we first extract uni-grams and bi-grams features from a tweet content. We then find if any of those extracted features are present in the lexicon. A tweet having at least one of the lexicon terms was categorized as heritage and not-heritage, otherwise. The categorized tweets were evaluated using the ground-truth labels. The resulting confusion matrix is presented in Table 6 and the performance reported in the first row of Table 7. Not surprisingly, the lexicon-based classifier misses many of the true heritage cases (i.e., 388 out of 487 instances), which results in a fairly low Recall= 0.20 for the heritage class. In practice, this means that only 1 out of 5 images of the heritage will be identified correctly by the lexicon-based classifier.

5.3.3 Heritage/Not-heritage Classification. First, we apply our heritage models on the SMERP dataset and compare their predictions with the ground-truth annotations.[4] Table 6 shows the resulting confusion matrices

---

[4]The results discussed in this section are obtained by the heritage models trained by Logistic Regression algorithm using `DenseNet121` features. To examine the performance of other models obtained by different combinations of CNN features and classification algorithms, we suggest that the reader inspects Tables 16–19 in Appendix Section D.
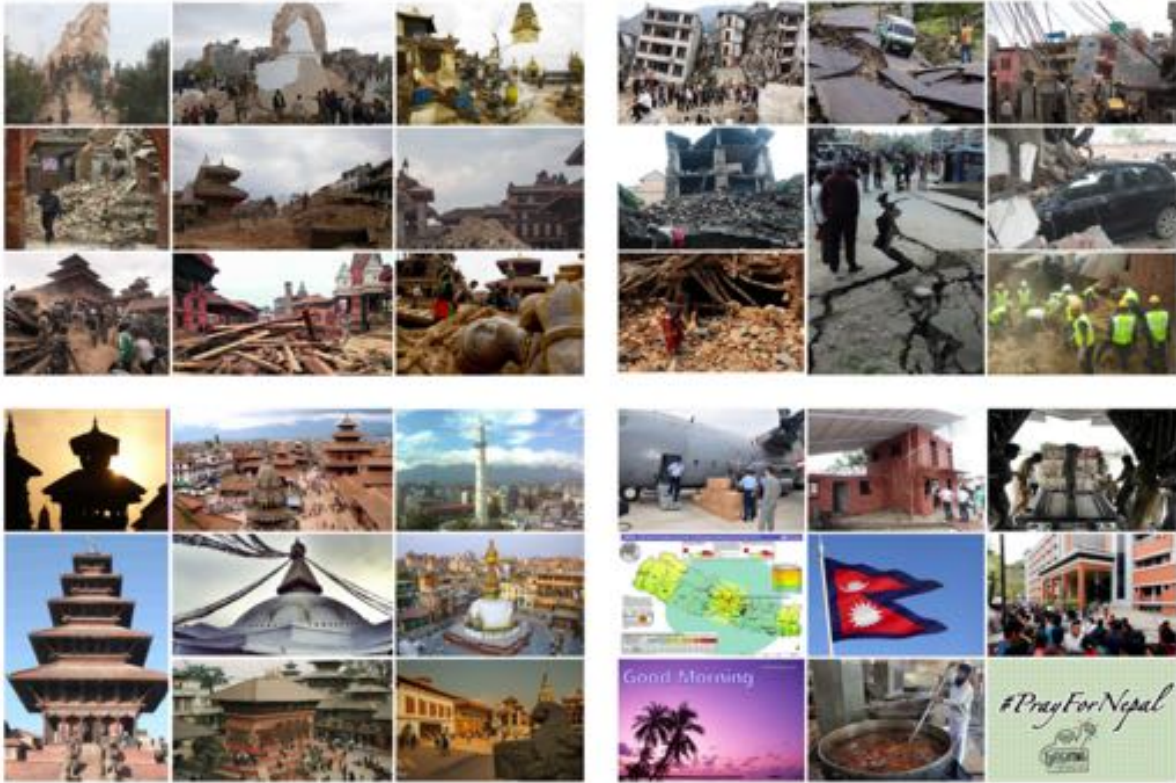
Fig. 7. Examples of annotated images from the SMERP dataset, showing heritage with damage (top left), not-heritage with damage (top right), heritage with no damage (bottom left), and not-heritage with no damage (bottom right).

Table 6. Confusion Matrices of the Heritage Classifiers on the SMERP Dataset

| | | | Classified as | |
			Heritage	Not-heritage
Lexicon-based Model	Actual label	Heritage	99	388
		Not-heritage	80	5,753
Heritage Model-1	Actual label	Heritage	423	64
		Not-heritage	3,869	1,964
Heritage Model-2	Actual label	Heritage	369	118
		Not-heritage	1,039	4,794

Table 7. Performance Comparison of the Heritage Classifiers on the SMERP Dataset

| | Precision | | Recall | | F1-score | |
	Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Lexicon-based Model	0.55	0.94	0.20	0.99	0.30	0.96
Heritage Model-1	0.10	0.97	0.87	0.34	0.18	0.50
Heritage Model-2	0.26	0.98	0.76	0.82	0.39	0.89
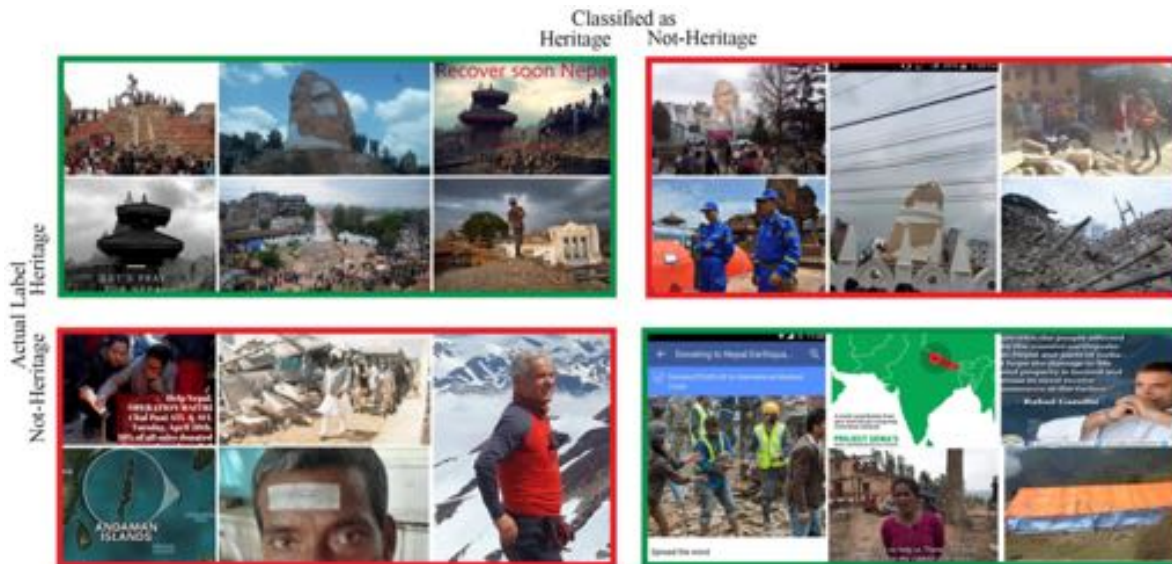
Fig. 8. Examples of images classified with Heritage Model 2.

between the predicted and ground-truth labels for both heritage models. Figure 8 illustrates the confusion matrices between the predicted and ground-truth with examples of images classified with the Heritage Model-2. It was found that the images that were particularly difficult to accurately classify included edited or altered images, aerial images, and satellite images. Images with overlapping architectural elements between the "heritage" and "not-heritage" categories were also difficult to classify. Lastly, images in which heritage was not the main subject of the image (refer top-right of Figure 8) tend to be difficult to classify. Further, Figures 10 and 11 in Appendix Section E provide examples of images classified with Heritage Model-1 and Lexicon-based Model. Moreover, Table 7 summarizes the performance of the heritage classifiers in terms of precision, recall, and F1-score. Ideally, the confusion matrix for a perfect model would have non-zero values only in the diagonal entries and zeros elsewhere (i.e., no incorrect predictions). However, this is rarely the case in real-world systems. Likewise, Heritage Model-2 does a decent job in classifying heritage images as heritage (i.e., 369 out of 487), which corresponds to a Recall= 0.76, and not-heritage images as not-heritage (i.e., 4,794 out of 5,833), which corresponds to a Recall= 0.82. However, the model makes some errors and classifies many not-heritage images also as heritage (i.e., 1,039), which results in a low score of Precision= 0.26; although, in the other direction, the model makes less errors and classifies fewer not-heritage images as heritage (i.e., 118), which leads to a high score of Precision= 0.98.

Another important observation to note is the significant difference in performance between the two heritage models on our case study dataset, although they seemed to perform on par on our Google images test set (as presented earlier in Section 5.2). First, there is a big difference in precision scores where Heritage Model-2 achieves a score of Precision= 0.26 while the Heritage Model-1 achieves only a score of Precision= 0.10. As Heritage Model-1 was *not* trained on sample images with damage context, it tends to classify many not-heritage images as heritage (i.e., 3,869 to be specific), which corresponds to a false positive rate of FPR= 0.66 based on Table 6. On the other hand, Heritage Model-2 makes a lower number of the Type-I errors (i.e., 1,039 to be specific), which brings the false positive rate down to FPR= 0.18 according to Table 6.

However, this increase in precision for Heritage Model-2 comes at the expense of a slight decrease in recall since Heritage Model-2 makes more Type-II errors than Heritage Model-1. Specifically, Heritage Model-2 predicts 118 heritage images as not-heritage (which corresponds to a false negative rate of FNR= 0.24), whereas Heritage

Table 8. Confusion Matrix for the Damage Classification

| | | Classified as | |
		Damage	No-damage
Actual label	Damage	1,580	242
	No-damage	368	4,130

Fig. 9. Examples of damage classification images.

Model-1 predicts 64 heritage images as not-heritage (which corresponds to a false negative rate of FNR= 0.13). When we compare the F1-scores of both models, which is the harmonic mean of the precision and recall scores, we see that the overall performance of Heritage Model-2 with a score of F1= 0.39 is much better than that of Heritage Model-1 with a score of F1= 0.18. In other words, Heritage Model-2 presents better generalization capabilities.

Although the lexicon-based model achieves the highest precision score (i.e., Precision= 0.55), its overall performance in terms of F1-score remains at F1= 0.30 due to its poor recall rate (i.e., Recall= 0.20) for the heritage class. Therefore, we conclude that Heritage Model-2 provides the best compromise for the heritage image classification task in practice.

5.3.4   Damage/No-Damage Classification. Then, we apply the damage assessment model of Alam et al. [2] on the SMERP dataset and compare the model's predictions with expert labels. As shown by the confusion matrix in Table 8, the model classifies most of the damage images correctly and misses only 242 damage images. Similarly, it misclassifies only 368 no-damage images. This yields a classification accuracy of 0.90. Figure 9 shows examples of damage classification images. Moreover, Table 9 summarizes the performance of the damage assessment model in terms of precision, recall, and F1-score. Based on these class-specific assessments, the model seems to perform relatively better on no-damage images than on damage images. The weighted average of these precision and recall scores tends to be closer to those for the no-damage class because of the imbalance distribution of damage and no-damage images in the SMERP dataset.

Table 9. Performance of the Damage Classifier
on the SMERP Dataset

	Precision	Recall	F1-score
Damage	0.81	0.87	0.84
No-damage	0.94	0.92	0.93

5.4 Discussion

The results from our experiments suggest that the proposed methodology to classify images from social media is helpful to understand damage to heritage sites during disasters. Its application can contribute toward a better understanding of the impact of disasters on cultural heritage and prepare a coordinated response.

We performed a comparative analysis of the precision and recall of each model to understand their relative performance in our case study. Table 7 summarizes the performances achieved by the three models. Even though the lexicon-based model yields the highest precision in our case study, its applicability can not be generalized for various reasons. First, the manually-curated lexicon contains only English terms. However, people often refer to terms in the local language when describing a heritage site. For instance, a temple is often referred to as a *mandir* in some countries. Second, the words in a lexicon can be used in a different context. For instance, heritage has been used to refer to *lineage* in many instances. Third, the lexicon-based model can result in data from undamaged or unaffected areas. We found that the term temple was also used to refer to an unaffected temple in an unaffected region. Fourth, the low recall of the lexicon-based model implies that only 20% of images from heritage sites will be found by this model. While low precision results in more manual work for the heritage professionals, low recall implies that many images simply go undetected. In a real-world scenario, it means that the overall assessment of damaged heritage may be quite incomplete with this model.

In comparison, Heritage Model-2's lower precision implies more manual labor for heritage professionals in sorting the relevant images, but its higher recall suggests that the chances of relevant images being undetected is substantially lower. Therefore, compared to the lexicon-based model, Heritage Model-2 is more likely to provide a better overall picture of the affected areas. On the other hand, Heritage Model-1's lowest precision and higher recall suggest that the manual labor of professionals is more than doubled, even though the overall picture of the affected areas may not be significantly better than the Heritage Model-2. More manual work for heritage professionals in this case would result in a delayed assessment in a real-world scenario. Therefore, we conclude that, among the three models, Heritage Model-2 is the most suitable model for heritage image classification as it will result in better assessment in less amount of time and require less manual work from the heritage professionals.

This is a challenging image classification task, as high performance would require visual features that can characterize *heritage sites* in an unambiguous manner. Overlapping spatial qualities, building form, architectural elements, and material of construction in *heritage* and *not-heritage* categories means this problem is inherently ambiguous. In addition, the fact that we try to identify heritage images in a disaster context makes the problem even more challenging. Our case study results revealed that a subtle difference in data curation and training (i.e., including *damaged* heritage and not-heritage images in the training of Heritage Model-2) can lead to significant differences in generalization capabilities and robustness of the trained models, especially when tested in a real-world scenario. To this end, our results highlight that the automatic classification of heritage images in disaster context is not an impossible task.

Many of the images depicting damaged heritage did not contain contextual information. This complicates the task further, even for a professional. However, our damage assessment model gave a high accuracy. Given the precision is above 80% for damaged heritage, a heritage professional examining the output would find false positives (images the system says are damaged heritage, but are not) of up to 20%. Given the recall is above 80% for the same class, 4 out of 5 images of damaged heritage can be found using these methods.

The automated method can significantly reduce the effort by the cultural heritage professional who would have to scan potentially hundreds of images just to find one heritage image. We are greatly reducing this effort by quickly processing and filtering thousands of images and presenting candidates, even if now the professional needs to look at four images on average to find one relevant image (i.e., precision of 26%, Table 7). Even with this level of precision, we believe these methods will substantially improve post-disaster heritage management practices. Since culture (or heritage) may not be an immediate need or priority in disaster-struck societies, the rescue of cultural heritage is rarely well-integrated with disaster management practices [34, 86]. In fact, post-disaster damage assessment in many cases is not completed until several months after the disaster [12], thereby making heritage vulnerable to further damage/decay. Further, in some cases, damage assessments are ineffective or incomplete due to lack of inventories, well-established processes, or expertise [86]. Our model, on the other hand, will give heritage professionals a tool to complete the rapid damage assessment soon after the disaster, and, therefore, the time-lag between disaster and action can be significantly reduced.

In our case study, we used a single type of disaster, i.e., an earthquake, a type of geophysical disaster. Therefore, a discussion on the classifiers' applicability in different subgroups of disasters is necessary at this point. The training of Heritage Model-2 using sample images with varied damage context increases the chances of correct predictions in different types of disasters such as geophysical, and miscellaneous accidents, as defined by EM-DAT [27]. It is also likely to detect damage in case of deliberate destruction of heritage during wars. However, the Heritage Model-2 will not perform in other scenarios such as hydrological disasters (e.g., floods), as the training dataset included only the above-mentioned sub-groups of disasters. Indeed, the characteristics of images produced in different types of disasters may vary in various aspects. Further, the characteristics of images produced on social media during two similar events may also vary in attributes. Therefore, further training of Heritage Model-2 with larger datasets from different scenarios will increase the wider applicability of the classifier.

Moreover, we have used a single data source, Twitter. The model may not work in different social media platforms (e.g., Instagram or Facebook), in which users may post other types of photos. Indeed, different platforms might be used by different users for different purposes [70]. Further training of Heritage Model-2 with datasets from different platforms can increase its applicability across platforms.

The results from our extensive experiments using various network architectures as feature extractors together with several classification algorithms showed that there can be variations in performance across different configurations. Although these variations are usually not dramatic, it is possible to obtain further performance improvements in precision and recall via some further engineering and parameter fine-tuning efforts. However, such engineered configurations may not translate from one setup to another, and should be part of the work done when deploying and maintaining these systems in practice.

6 CONCLUSIONS

The process we have described requires many elements: a careful delimitation of the images to be processed, a comprehensive data collection strategy that ensures diversity, a careful annotation of data points that can avoid ambiguities in the training set, a state-of-the-art deep learning method to learn to classify images, and an in-depth evaluation to understand the performance of different classifiers.

The results, however, are, in our opinion, worth the effort. Social media provides a nearly instantaneous view of cultural heritage sites affected by a disaster, including many ground-level photos that cannot be replaced by the bird's-eye perspective provided by UAVs and satellite images. However, photos of heritage sites are a tiny minority of all the images that are posted, and images depicting a damaged heritage site as the main subject are rare. Finding them manually in an avalanche of unrelated images from social media is simply impractical. Our methods can greatly reduce the number of images to be examined by a cultural heritage professional.

Future work. The quality of our classifier can be improved by a larger, more diverse training set. However, annotating images selected at random from a social media stream during a disaster is impractical considering the relatively low frequency of damaged heritage photos. Hence, we envision using the classifier we have created

to find candidate images for further annotation. Moreover, we can maximize our utilization of multimedia content on social media platforms by formulating the heritage classification problem in a more sophisticated way as a multimodal learning problem where the goal would be to combine features extracted from various modalities (e.g., text, image, video) to train a heritage classification model. That being said, unlike Twitter, such aligned multimodal data are not prominent on most other social media platforms (e.g., Instagram and Flickr). Therefore, a technology based only on images would still be desirable in such cases. An additional area for further work is the identification of different types of damage, such as mild and severe damage, which may help in the prioritization of efforts. Dealing with images from an earthquake may be easier than dealing with images from a more localized disaster, such as an explosion (intentional or accidental), because after an earthquake, there is a large number of people distributed over a large area who can directly witness the consequences of the event. It might also be the case that during natural disasters, there is less misleading information than during a human-made disaster such as a war; in any case, further experimentation with other types of disasters would help improve and fine-tune these methods. Ultimately, joint modeling of heritage classification and damage assessment tasks in a unified framework bears great potential to provide better understanding of heritage images in a disaster context.

APPENDICES

A LIST OF HERITAGE AND NOT-HERITAGE SITES

Table 10. List of Heritage Sites

Class	Name of the Site	Location	Country
Architectural	Hagia Sophia	Istanbul	Turkey
	Jaisalmer Fort	Jailsalmer	India
	City of Bath	Bath	UK
	Historic City of Ahmedabad	Ahmedabad	India
	Roskilde Cathedral	Roskilde	Denmark
	Tamshing Monastery	Bumthang	Bhutan
	Notre-Dame Cathedral	Paris	France
	Santa Maria Novella	Florence	Italy
	Alhambra, Generalife, and Albayzín	Granada	Spain
	Red Fort	New Delhi	India
	Sydney Opera House	Sydney	Australia
	Summer Palace	Beijing	China
	Borobudur Temple	Jawa Tengah	Indonesia
	Chinque Terre	Chinque Terre	Italy
	Edinburgh Castle	Edinburgh	UK
	Capitol Complex	Chandigarh	India
	Ellora Caves	Aurangabad	India
	Wellington Arch	London	UK
	Taj Mahal	Agra	India
	Kings Cross St. Pancras Station	London	UK
	Trafalgar Square	London	UK
	Chhatrapati Shivaji Terminus	Mumbai	India
	India Gate	Delhi	India
	The Taj Mahal Palace	Mumbai	India
	Adalaj ni Vav	Ahmedabad	India
	Fatehpur Sikri	Agra	India

	Sarnath Stupa	Sarnath	India
	Sun Temple Modhera	Ahmedabad	India
	Gadisar Lake	Jaisalmer	India
	Mehrangarh Fort	Jodhpur	India
	US Capitol Building	Washington, D.C.	USA
	Notre-Dame Cathedral Basilica	Saigon	Vietnam
	Parthenon	Nashville	USA
	Colosseum	Rome	Italy
	Jama Masjid	Delhi	India
	Dochula Temple	Hungtso	Bhutan
	Punakha Dzong	Punakha	Bhutan
	Tiger Nest Monastery	Taktsang trail	Bhutan
	Arc de Triomphe du Carrousel	Paris	France
	Pyathatgyi Temple	Minnanthu Region	Myanmar
	Bamiyan Buddha	Bamyan	Afghanistan
	Palmyra	Tadmur	Syria
	Aleppo's Umayyad Mosque	Aleppo	Syria
	Sanaa Old City	Sanaa	Yemen
	Windsor Castle	Windsor	UK
Gallery/Library/Museum	Kensington Palace Museum	London	UK
	British Museum	London	UK
	Victoria and Albert Museum	London	UK
	The Louvre	Paris	France
	Uffizi Gallery	Florence	Italy
	British Library	London	UK
	Museum Orsay	Paris	France
	Solomon R. Guggenheim Museum	New York	USA
	Rijksmuseum	Amsterdam	Netherlands
	National Museum of Cinema	Turin	Italy
	Camposanto	Pisa	Italy
	The São Paulo Museum of Art	São Paulo	Brazil
	National War Museum Malta	Valletta	Malta
	Library of Parliament Ottawa	Ottawa	Canada
	Metropolitan Museum of Art	New York	USA
	National Museum	Paro	Bhutan
Archaeological	Machu Pichu	Urubamba River Valley	Peru
	Stonehenge	Salisbury	UK
	Mohenjo Daro	Sindh	Pakistan
	Teotihuacan	Teotihuacan	Mexico
	Hagar Qim	Qrendi	Malta
	Palmyra	Palmyra	Syria
	Ajanta Caves	Aurangabad	India
	Pyramids of giza	Giza	Egypt
	Golden Temple of Dambulla	Dambulla	Sri Lanka
	Rani ki vav	Ahmedabad	India
	Petra	Petra	Jordan

	Pompeii	Campania	Italy
	Delphi	Phocis	Greece
	Parthenon	Athens	Greece
	Angkor Wat	Siem Reap	Cambodia
Artifact in Urban Space	Christ the Redeemer	Rio de Janeiro	Brazil
	Terracotta Warriors of Shaanxi	Shaanxi	China
	Statue of Liberty	New York	USA
	The Little Mermaid Statue	Copenhagen	Denmark
	Telephone Booth London	London	UK
	Stroke Fountain	Copenhagen	Denmark
	Lincoln Memorial	Washington, D.C.	USA
	Gateway of India	Mumbai	India
	The Porcellino	Florence	Italy
	Statue of Hans Christian Andersen	Copenhagen	Denmark
	Open Hand Monument	Chandigarh	India
	Flaminio Obelisk	Rome	Italy
	Christopher Columbus Statue	New York	USA
	Dandi March Sculpture	New Delhi	India
	Statue of Mahatma Gandhi	London	UK
	Marble Arch	London	UK
	Sphinx	Giza	Egypt

Table 11. List of Not-heritage Sites

Class	Name of the Site	Location	Country
Architectural	India Habitat Center	New Delhi	India
	The Shard	London	UK
	IIM	Ahmadabad	India
	Walkie Talkie London	London	UK
	Kanchanganga Apartment	Mumbai	India
	Dharavi	Mumbai	India
	New Delhi Railway Station	New Delhi	India
	Lucca Railway Station	Lucca	Italy
	IT University of Copenhagen	Copenhagen	Denmark
	Pittsburgh Airport	Pittsburgh	USA
	Northlake Mall	Charlotte	USA
	Wembley Stadium	London	UK
	Radisson Blu Hotel	Copenhagen	Denmark
	Tiaa Cref Office Charlotte	Charlotte	USA
	University College Hospital	London	UK
	Danish Opera House	Copenhagen	Denmark
	Hall of Nations	New Delhi	India
	Turning Torso	Malmo	Sweden

	Tata Steel Industry Building	Jamshedpur	India
	Volkswagen Factory Building	Salzgitter	Germany
	Bella Sky Hotel Copenhagen	Copenhagen	Denmark
	CSV Building Wardha	Wardha	India
	Railway Office Bilaspur	Bilaspur	India
	Navi Mumbai Railway Station	Mumbai	India
	Belapur Housing Building	Mumbai	India
	8 House	Copenhagen	Denmark
	Munich Airport	Munich	Germany
	Gherkin Building	London	UK
	UNCC	Charlotte	USA
	Fisketorvet	Copenhagen	Denmark
	Great India Place Mall	Noida	India
Gallery Library Museum	Copenhagen Main Library	Copenhagen	Denmark
	Lalit Kala Akademi	Delhi	India
	Husain Doshi Gufa	Ahmadabad	India
	Crafts Museum Delhi	New Delhi	India
	The Blue Planet	Copenhagen	Denmark
	The Black Diamond	Copenhagen	Denmark
	Wax Museum London Building	London	UK
	Sanskriti Kendra	Delhi	India
	Jawahar Kala Kendra	Jaipur	India
	World Trade Center Museum	New York	USA
	New Jewish Museum	Berlin	Germany
	5 Pointz	New York	USA
	Niterói Contemporary Art Museum	Rio de Janeiro	Brazil
	Petrie Museum building	London	UK
	UNCC Library	Charlotte	USA
Artifact in Urban Space	Tilted Arc by Richard Serra	New York	USA
	Penis Christmas Tree	Paris	France
	Brown Nosing Sculpture	Prague	Czech
	Milan Stock Exchange Sculpture	Milan	Italy
	Anish Kapoor Versailles	Versailles	France
	Les Deux Plateaux, Colonnes de Buren	Paris	France
	The Vigeland Park	Oslo	Norway
	Sun Dial New Delhi Barahpulla	New Delhi	India
	Sunbather Sculpture	Long Island	USA
	Fearless Girl	New York	USA
	MGR Memorial	Chennai	India
	Rooster National Gallery	London	UK
	Calgary Sculpture Controversy	Calgary	Canada
	Chicago and Milwaukee Eyeball	Chicago	USA

B LEXICON

All terms are case-insensitive. For presentation purposes, we divide them into five groups of comma-separated terms; we do not make any difference between the groups for the purposes of matching within a text.

1. Generic words: heritage, heritages, cultural, culturally, culture, cultured, cultures, historically, historic, historical, ancient, ancients, architecture, architectural, architecturally, architectures, archaeology, archaeological, archaeologically, civilizations, civilization.

2. Bi-grams: traditional building, traditional architecture, cultural center, cultural complex, cultural ensemble, cultural landscape, cultural masterpiece, historic building, historic town, historic city, historic site, historic architecture, historic center, historic settlement, historic settings, historic civilization, historic ensemble, historic built, historic settlement, historic environment, old city, old town, old buildings, sacred building, ancient architecture, ancient building, ancient settlement, heritage building, heritage city, heritage property, heritage site, ceremonial architecture, ceremonial buildings, landmark building, iconic site, iconic building.

3. Site types: churches, church, palaces, palace, temple, temples, monuments, monumental, monumentality, monastery, monasteries, towers, tower, towered, towering, castles, castle, cathedral, cathedrals, tombs, tomb, caves, cave, mosque, mosques, fortresses, fortress, fortified, fortify, fortifying, chapels, chapel, fortifications, fortification, forts, fort, forte, museum, museums, basilicas, basilica, sculptures, sculptural, sculpture, sculptured, sculpturing, monastic, citadels, citadel, mausoleum, mausoleums, abbey, abbeys, pyramids, pyramid, pyramidal, memorial, memorials, memories, memory.

4. Styles and periods: romans, roman, romane, romanization, romanized, medieval, empires, empire, dynasty dynasties, kingdom, kingdoms, gothicized, gothic, gothicism, gothicizing, baroque, renaissance, imperial, imperialism, classical, classic, classically, classicism, classics, buddhist, buddhists, byzantine, byzantines, romanesque, prehistoric, prehistorical, neolithic, ottoman, ottomans, hellenistic, neoclassical, 1st century, 2nd century, 3rd century, 4th century, 5th century, 6th century, 7th century, 8th century, 9th century, 10th century, 11th century, 12th century, 13th century, 14th century, 15th century, 16th century, 17th century, 18th century, 19th century.

5. Organization: unesco, #unesco, @unesco.

C ALL EXPERIMENTAL RESULTS ON GOOGLE IMAGES

Table 12. Performance Comparison of Various CNN Features with Logistic Regression Classifier

	Architecture	Precision		Recall		F1-score	
		Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Heritage Model-1	VGG16	0.93	0.64	0.85	0.81	0.89	0.72
	ResNet50	0.94	0.65	0.86	0.82	0.89	0.73
	DenseNet121	0.91	0.65	0.87	0.75	0.89	0.70
	InceptionResNetV2	0.91	0.63	0.86	0.75	0.88	0.69
	Xception	0.92	0.66	0.87	0.78	0.89	0.72
	NASNetLarge	0.92	0.71	0.90	0.78	0.91	0.75
Heritage Model-2	VGG16	0.94	0.63	0.84	0.84	0.88	0.72
	ResNet50	0.93	0.63	0.84	0.81	0.88	0.71
	DenseNet121	0.94	0.66	0.86	0.84	0.90	0.74
	InceptionResNetV2	0.93	0.62	0.84	0.81	0.88	0.70
	Xception	0.94	0.62	0.83	0.84	0.88	0.72
	NASNetLarge	0.94	0.67	0.86	0.84	0.90	0.74

Table 13. Performance Comparison of Various CNN Features with Support Vector Machine Classifier

	Architecture	Precision		Recall		F1-score	
		Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Heritage Model-1	VGG16	0.93	0.63	0.84	0.82	0.88	0.71
	ResNet50	0.93	0.64	0.85	0.81	0.89	0.72
	DenseNet121	0.92	0.63	0.86	0.76	0.88	0.69
	InceptionResNetV2	0.91	0.61	0.84	0.74	0.88	0.67
	Xception	0.92	0.64	0.86	0.77	0.89	0.70
	NASNetLarge	0.92	0.70	0.89	0.76	0.90	0.73
Heritage Model-2	VGG16	0.93	0.62	0.83	0.82	0.88	0.71
	ResNet50	0.93	0.62	0.84	0.81	0.88	0.70
	DenseNet121	0.94	0.66	0.86	0.83	0.90	0.73
	InceptionResNetV2	0.93	0.59	0.81	0.80	0.87	0.68
	Xception	0.93	0.58	0.80	0.82	0.86	0.68
	NASNetLarge	0.93	0.64	0.85	0.82	0.89	0.72

Table 14. Performance Comparison of Various CNN Features with Random Forest Classifier

	Architecture	Precision		Recall		F1-score	
		Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Heritage Model-1	VGG16	0.89	0.72	0.91	0.66	0.90	0.69
	ResNet50	0.90	0.71	0.91	0.69	0.90	0.70
	DenseNet121	0.89	0.72	0.92	0.67	0.90	0.70
	InceptionResNetV2	0.90	0.68	0.89	0.69	0.90	0.69
	Xception	0.88	0.66	0.89	0.64	0.89	0.65
	NASNetLarge	0.91	0.73	0.91	0.71	0.91	0.72
Heritage Model-2	VGG16	0.91	0.72	0.90	0.74	0.91	0.73
	ResNet50	0.93	0.73	0.90	0.78	0.91	0.75
	DenseNet121	0.91	0.71	0.90	0.74	0.91	0.73
	InceptionResNetV2	0.92	0.68	0.88	0.75	0.90	0.71
	Xception	0.90	0.67	0.89	0.70	0.89	0.69
	NASNetLarge	0.93	0.72	0.90	0.81	0.92	0.76

Table 15. Performance Comparison of Various CNN Features with AdaBoost Classifier

Architecture	Precision		Recall		F1-score	
	Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Heritage Model-1						
VGG16	0.91	0.64	0.87	0.74	0.89	0.69
ResNet50	0.92	0.62	0.85	0.77	0.88	0.69
DenseNet121	0.93	0.67	0.87	0.79	0.90	0.72
InceptionResNetV2	0.91	0.65	0.87	0.72	0.89	0.68
Xception	0.90	0.60	0.84	0.73	0.87	0.66
NASNetLarge	0.91	0.70	0.90	0.74	0.90	0.72
Heritage Model-2						
VGG16	0.92	0.63	0.85	0.78	0.88	0.70
ResNet50	0.94	0.65	0.86	0.82	0.89	0.73
DenseNet121	0.93	0.63	0.84	0.81	0.88	0.71
InceptionResNetV2	0.92	0.61	0.84	0.77	0.88	0.68
Xception	0.91	0.62	0.85	0.74	0.88	0.68
NASNetLarge	0.93	0.64	0.85	0.81	0.89	0.72

D ALL EXPERIMENTAL RESULTS ON SMERP IMAGES

Table 16. Performance Comparison of Various CNN Features with Logistic Regression Classifier

Architecture	Precision		Recall		F1-score	
	Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Heritage Model-1						
VGG16	0.11	0.97	0.81	0.46	0.20	0.63
ResNet50	0.10	0.96	0.85	0.34	0.17	0.50
DenseNet121	0.10	0.97	0.87	0.34	0.18	0.50
InceptionResNetV2	0.12	0.98	0.85	0.50	0.22	0.66
Xception	0.10	0.97	0.85	0.38	0.18	0.55
NASNetLarge	0.10	0.97	0.91	0.28	0.17	0.44
Heritage Model-2						
VGG16	0.24	0.97	0.73	0.81	0.36	0.88
ResNet50	0.24	0.97	0.74	0.81	0.37	0.88
DenseNet121	0.26	0.98	0.76	0.82	0.39	0.89
InceptionResNetV2	0.24	0.97	0.74	0.81	0.37	0.88
Xception	0.23	0.97	0.76	0.79	0.35	0.87
NASNetLarge	0.25	0.98	0.79	0.81	0.38	0.88

Table 17. Performance Comparison of Various CNN Features with Support Vector Machine Classifier

	Architecture	Precision		Recall		F1-score	
		Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Heritage Model-1	VGG16	0.11	0.97	0.81	0.48	0.20	0.64
	ResNet50	0.10	0.96	0.83	0.35	0.17	0.51
	DenseNet121	0.09	0.97	0.87	0.31	0.17	0.47
	InceptionResNetV2	0.12	0.97	0.84	0.48	0.21	0.64
	Xception	0.10	0.97	0.85	0.38	0.18	0.54
	NASNetLarge	0.10	0.98	0.91	0.32	0.18	0.48
Heritage Model-2	VGG16	0.22	0.97	0.72	0.79	0.34	0.87
	ResNet50	0.23	0.97	0.73	0.79	0.35	0.87
	DenseNet121	0.25	0.97	0.74	0.82	0.38	0.89
	InceptionResNetV2	0.23	0.98	0.77	0.78	0.35	0.87
	Xception	0.20	0.97	0.73	0.76	0.32	0.85
	NASNetLarge	0.22	0.98	0.77	0.77	0.34	0.86

Table 18. Performance Comparison of Various CNN Features with Random Forest Classifier

	Architecture	Precision		Recall		F1-score	
		Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Heritage Model-1	VGG16	0.10	0.98	0.94	0.26	0.17	0.41
	ResNet50	0.09	0.98	0.96	0.15	0.16	0.26
	DenseNet121	0.09	0.98	0.97	0.13	0.16	0.24
	InceptionResNetV2	0.09	0.97	0.93	0.23	0.17	0.38
	Xception	0.09	0.97	0.93	0.18	0.16	0.30
	NASNetLarge	0.09	0.97	0.95	0.15	0.16	0.27
Heritage Model-2	VGG16	0.19	0.98	0.86	0.69	0.31	0.81
	ResNet50	0.17	0.99	0.90	0.64	0.29	0.78
	DenseNet121	0.19	0.99	0.89	0.69	0.32	0.81
	InceptionResNetV2	0.20	0.98	0.85	0.72	0.33	0.83
	Xception	0.16	0.98	0.84	0.62	0.26	0.76
	NASNetLarge	0.18	0.99	0.88	0.67	0.30	0.80

Table 19. Performance Comparison of Various CNN Features with AdaBoost Classifier

	Architecture	Precision		Recall		F1-score	
		Heritage	Not-heritage	Heritage	Not-heritage	Heritage	Not-heritage
Heritage Model-1	VGG16	0.10	0.97	0.86	0.35	0.18	0.52
	ResNet50	0.11	0.97	0.86	0.41	0.19	0.58
	DenseNet121	0.10	0.97	0.88	0.30	0.17	0.46
	InceptionResNetV2	0.10	0.97	0.88	0.31	0.17	0.47
	Xception	0.09	0.97	0.89	0.28	0.17	0.43
	NASNetLarge	0.10	0.97	0.86	0.38	0.19	0.55
Heritage Model-2	VGG16	0.20	0.98	0.78	0.74	0.32	0.84
	ResNet50	0.24	0.98	0.78	0.80	0.37	0.88
	DenseNet121	0.24	0.98	0.76	0.80	0.37	0.88
	InceptionResNetV2	0.20	0.98	0.80	0.73	0.32	0.84
	Xception	0.20	0.98	0.81	0.73	0.32	0.84
	NASNetLarge	0.20	0.98	0.82	0.73	0.32	0.84

E   IMAGES

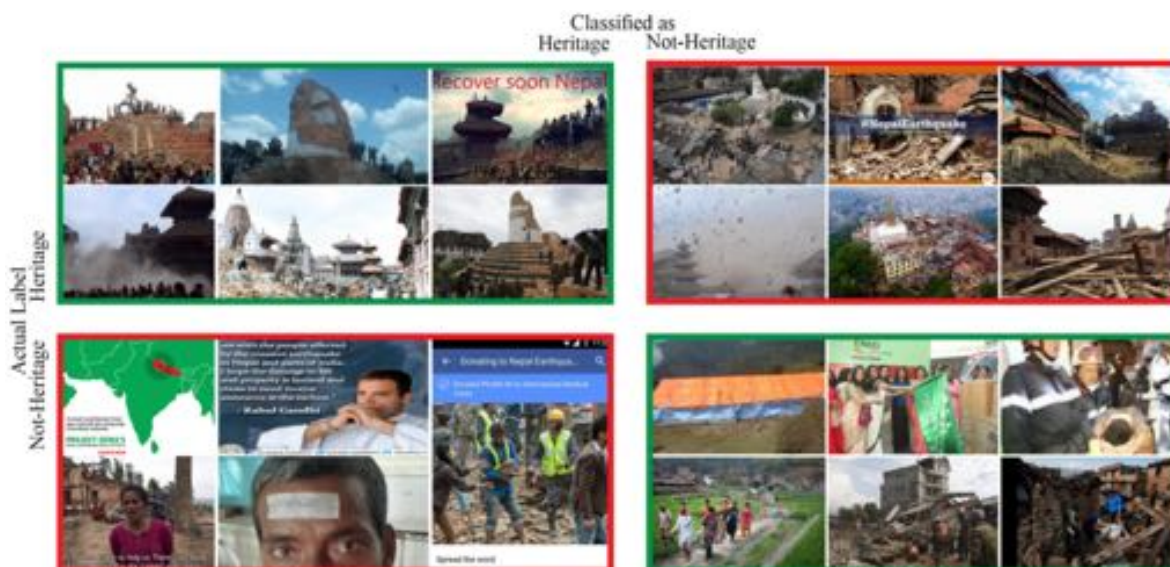Fig. 10.   Examples of images classified with Lexicon-based Model.

Fig. 11. Examples of images classified with Heritage Model 1.

REFERENCES

[1] Yahaya Ahmad. 2006. The scope and definitions of heritage: from tangible to intangible. *Int. J. Heritage Stud.* 12, 3 (2006), 292–300.

[2] Firoj Alam, Muhammad Imran, and Ferda Ofli. 2017. Image4Act: Online social media image processing for disaster response. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE/ACM, Sydney, Australia, 1–4.

[3] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. AAAI, Stanford, CA, 465–473.

[4] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Processing social media images by combining human and machine computing during crises. *Int. J. Hum. Comput. Interact.* 34, 4 (2018), 311–327.

[5] Giuseppe Amato, Fabrizio Falchi, and Claudio Gennaro. 2015. Fast image classification for monument recognition. *J. Comput. Cult. Herit.* 8, 4, Article 18 (August 2015), 25 pages. DOI : https://doi.org/10.1145/2724727

[6] Giuseppe Amato, Fabrizio Falchi, and Lucia Vadicamo. 2016. Visual recognition of ancient inscriptions using convolutional neural network and Fisher vector. *J. Comput. Cult. Herit.* 9, 4, Article 21 (December 2016), 24 pages. DOI : https://doi.org/10.1145/2964911

[7] Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. 2014. Tweedr: Mining twitter to inform disaster response. In *ISCRAM*.

[8] Nazia Attari, Ferda Ofli, Mohammad Awad, Ji Lucas, and Sanjay Chawla. 2017. Nazr-CNN: Fine-grained classification of UAV imagery for damage assessment. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, Tokyo, Japan, 1–10.

[9] Marco Avvenuti, Stefano Cresci, Mariantonietta N La Polla, Andrea Marchetti, and Maurizio Tesconi. 2014. Earthquake emergency management by social sensing. In *Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*. IEEE, 587–592.

[10] Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, and Maurizio Tesconi. 2014. EARS (Earthquake Alert and Report System): A real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1749–1758.

[11] Melissa Bica, Leysia Palen, and Chris Bopp. 2017. Visual representations of disaster. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'17)*. ACM, New York, NY, 1262–1276. DOI : https://doi.org/10.1145/2998181.2998212

[12] Luigia Binda, C Modena, F. Casarin, F. Lorenzoni, Lorenzo Cantini, and Stefano Munda. 2011. Emergency actions and investigations on cultural heritage after the LâAquila earthquake: The case of the spanish fortress. *Bull. Earthquake Eng.* 9, 1 (2011), 105–138.

[13] Alessandra Bonazza, Ingval Maxwell, Miloš Drdácký, Ellizabeth Vintzileou, and Christian Hanus. 2018. *Safeguarding Cultural Heritage from Natural and Man-made Disasters – A Comparative Analysis of Risk Management in the EU*. Publications Office, Luxembourg. Retrieved from http://publications.europa.eu/publication/manifestation_identifier/PUB_NC0517059ENN.

[14] Cigdem Bozdag and Kevin Smets. 2017. Understanding the images of alan kurdi with "small data": A qualitative, comparative analysis of tweets about refugees in turkey and flanders (belgium). *Int. J. Comm.* 11, 0 (2017). https://ijoc.org/index.php/ijoc/article/view/7252.

[15] Gülcan Can, Jean-Marc Odobez, and Daniel Gatica-Perez. 2016. Evaluating shape representations for Maya glyph classification. *J. Comput. Cult. Herit.* 9, 3, Article 14 (September 2016), 26 pages. DOI: https://doi.org/10.1145/2905369

[16] Carlos Castillo. 2016. *Big Crisis Data: Social Media in Disasters and Time-critical Situations.* Cambridge University Press.

[17] Francois Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 1800–1807.

[18] Wei-Ta Chu and Ming-Hung Tsai. 2012. Visual pattern discovery for architecture image classification and product image search. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR'12).* ACM, New York, NY, Article 27, 8 pages. DOI: https://doi.org/10.1145/2324796.2324831

[19] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.* 20, 3 (Sept. 1995), 273–297. DOI: https://doi.org/10.1007/BF00994018

[20] David R. Cox. 1958. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B (Methodol.)* 20, 2 (1958), 215–242. http://www.jstor.org/stable/2983890.

[21] Johnny Cusicanqui, Norman Kerle, and Francesco Nex. 2018. Usability of aerial video footage for 3-D scene reconstruction and structural damage assessment. *Natural Hazards and Earth System Science* 18, 6 (2018), 1583–1598.

[22] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 886–893.

[23] Shannon Daly and James A. Thom. 2016. Mining and classifying image posts on social media to analyse fires. In *International Conference on Information Systems for Crisis Response and Management (ISCRAM).* ISCRAM, Rio de Janeiro, Brazil, 1–14.

[24] D. Duarte, F. Nex, N. Kerle, and G. Vosselman. 2018. Satellite image classificaiton of building damages using airborne and satellite image samples in a deep learning approach. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2 (2018), 89–96.

[25] Simon Faulkner, Farida Vis, and Francesco D'Orazio. 2018. Analysing social media images. In *The SAGE Handbook of Social Media*, J. Burgess, A. Marwick, and T. Poell (Eds.). SAGE Publications, London, UK, 160–178.

[26] J. Fernandez Galarreta, N. Kerle, and M. Gerke. 2015. UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. *Nat. Hazards Earth Syst. Sci.* 15, 6 (2015), 1087–1101.

[27] Centre for Research on the Epidemiology of Disasters. 2019. General Classification. Retrieved from https://www.emdat.be/classification.

[28] Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* 55, 1 (Aug. 1997), 119–139.

[29] Cristina Garduño Freeman. 2010. Photosharing on flickr: Intangible heritage and emergent publics. *Int. J. Heritage Stud.* 16, 4–5 (2010), 352–368.

[30] Ross Girshick et al. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 580–587.

[31] Abhinav Goel, Mayank Juneja, and C. V. Jawahar. 2012. Are buildings only instances?: Exploration in architectural style categories. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'12).* ACM, New York, NY, Article 1, 8 pages. DOI: https://doi.org/10.1145/2425333.2425334

[32] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press. http://www.deeplearningbook.org.

[33] Brian Graham, Greg Ashworth, and John Tunbridge. 2016. *A Geography of Heritage: Power, Culture and Economy.* Routledge.

[34] Kristy Graham and Dirk H. R. Spennemann. 2006. Heritage managers and their attitudes towards disaster management for cultural heritage resources in New South Wales, Australia. *Int. J. Emergency Manage.* 3, 2–3 (2006), 215–237.

[35] Larry P. Gross, John Stuart Katz, and Jay Ruby. 2003. *Image Ethics in the Digital Age.* University of Minnesota Press, Minneapolis, MN.

[36] Catherine Hartung. 2017. *Selfies for/of Nepal: Acts of Global Citizenship and Bearing Witness.* Springer International Publishing, Cham. 39–47 pages. DOI: https://doi.org/10.1007/978-3-319-45270-8_5

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90

[38] Larissa Hjorth and Jean Burgess. 2014. Intimate banalities: The emotional currency of shared camera phone images during the Queensland flood disaster. (2014).

[39] Nadav Hochman and Lev Manovich. 2013. Zooming into an Instagram city: Reading the local through social media. *First Monday* 18, 7 (2013).

[40] Rui Hu, Jean-Marc Odobez, and Daniel Gatica-Perez. 2017. Extracting Maya glyphs from degraded ancient documents via image segmentation. *J. Comput. Cult. Herit.* 10, 2, Article 10 (April 2017), 23 pages. DOI: https://doi.org/10.1145/2996859

[41] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What we Instagram: A first analysis of instagram photo content and user types. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8118/8087.

[42] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2261–2269. DOI : https://doi.org/10.1109/CVPR.2017.243

[43] Thomas Hurtut, Yann Gousseau, Farida Cheriet, and Francis Schmitt. 2011. Artistic line-drawings retrieval based on the pictorial content. *J. Comput. Cult. Herit.* 4, 1, Article 3 (August 2011), 23 pages. DOI : https://doi.org/10.1145/2001416.2001419

[44] Yasmin Ibrahim. 2015. Self-representation and the disaster event: Self-imaging, morality, and immortality. *J. Media Practice* 16, 3 (2015), 211–227. DOI : https://doi.org/10.1080/14682753.2015.1116755

[45] ICOMOS. 1964. The Venice Charter: International for the conservation and restoration of monuments and sites. (1964).

[46] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Extracting information nuggets from disaster-related messages in social media. In *ISCRAM*.

[47] Bente Jensen. 2013. Instagram as cultural heritage: User participation, historical documentation, and curating in museums and archives through social media. In *Proceedings of the 2013 Digital Heritage International Congress (DigitalHeritage)*, Vol. 2. IEEE, 311–314.

[48] Jacob Jett, Megan Senseney, and Carole L. Palmer. 2012. Enhancing cultural heritage collections by supporting and analyzing participation in Flickr. *Proceedings of the American Society for Information Science and Technology* 49, 1 (2012), 1–4.

[49] Rohit Jigyasu. 2016. Reducing disaster risks to urban cultural heritage: Global challenges and opportunities. *Journal of Heritage Management* 1, 1 (2016), 59–67.

[50] Rohit Jigyasu, Manas Murthy, Giovanni Boccardi, Christopher Marrion, Diane Douglas, Joseph King, Geoff O'Brien, Glenn Dolcemascolo, Yongkyun Kim, Paola Albrito, et al. 2013. Heritage and resilience: Issues and opportunities for reducing disaster risks. (2013).

[51] Mohammad Kakooei and Yasser Baleghi. 2017. Fusion of satellite, aircraft, and UAV data for automatic disaster damage assessment. *Int. J. Remote Sens.* 38, 8–10 (March 2017), 2511–2534.

[52] Martin R. Kalfatovic, Effie Kapsalis, Katherine P. Spiess, Anne Van Camp, and Michael Edson. 2008. Smithsonian team Flickr: A library, archives, and museums collaboration in Web 2.0 space. *Arch. Sci.* 8, 4 (2008), 267–277.

[53] Tamara Kharroub and Ozen Bas. 2016. Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution. *New Media Soc.* 18, 9 (2016), 1973–1992.

[54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. 1097–1105.

[55] Ryan Lagerstrom, Yulia Arzhaeva, Piotr Szul, Oliver Obst, Robert Power, Bella Robinson, and Tomasz Bednarz. 2016. Image classification to support emergency situation awareness. *Front. Rob. AI* 3 (2016), 54. DOI : https://doi.org/10.3389/frobt.2016.00054

[56] Xukun Li, Doina Caragea, Cornelia Caragea, Muhammad Imran, and Ferda Ofli. 2019. Identifying disaster damage images using a domain adaptation approach. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. ISCRAM, Valencia, Spain, 1–13.

[57] Xukun Li, Huaiyu Zhang, Doina Caragea, and Muhammad Imran. 2018. Localizing and quantifying damage in social media images. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE/ACM, Barcelona, Spain, 194–201.

[58] Jose Llamas, Pedro M. Lerones, Roberto Medina, Eduardo Zalama, and Jaime Gómez-García-Bermejo. 2017. Classification of architectural heritage images using deep learning techniques. *Appl. Sci.* 7, 10 (2017), 1–25. DOI : https://doi.org/10.3390/app7100992

[59] David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60, 2 (November 2004), 91–110. DOI : https://doi.org/10.1023/B:VISI.0000029664.99615.94

[60] D-Lib Magazine. 2014. Participatory cultural heritage: A tale of two institutions' use of social media. *D-lib Magazine* 20, 3/4 (2014).

[61] Michael Makridis and Petros Daras. 2013. Automatic classification of archaeological pottery sherds. *J. Comput. Cult. Herit.* 5, 4, Article 15 (January 2013), 21 pages. DOI : https://doi.org/10.1145/2399180.2399183

[62] M. Mathias, A. Martinovic, J. Weissenberg, S. Haegler, and L. Van Gool. 2011. Automatic architectural style recognition. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 3816 (September 2011), 171–176. DOI : https://doi.org/10.5194/isprsarchives-XXXVIII-5-W16-171-2011

[63] Mary Meeker and Liang Wu. 2016. Internet trends report 2016. *Kleiner Perkins Caufield Byers* (2016).

[64] Yelena Mejova, Ingmar Weber, and Michael W Macy. 2015. *Twitter: A Digital Socioscope*. Cambridge University Press.

[65] Alexander Mills, Rui Chen, JinKyu Lee, and H. Raghav Rao. 2009. Web 2.0 emergency applications: How useful can Twitter be for emergency response? *J. Inf. Privacy Secur.* 5, 3 (2009), 3–26. DOI : https://doi.org/10.1080/15536548.2009.10855867 arXiv:https://doi.org/10.1080/15536548.2009.10855867

[66] Marie-Francine Moens, Gareth J. F. Jones, Saptarshi Ghosh, Debasis Ganguly, Tanmoy Chakraborty, and Kripabandhu Ghosh. 2018. WWW'18 Workshop on Exploitation of Social Media for Emergency Relief and Preparedness: Chairs' Welcome and Organization. In *Companion Proceedings of the Web Conference 2018 (WWW'18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1609–1611. DOI : https://doi.org/10.1145/3184558.3192321

[67] H. Mouzannar, Y. Rizk, and M. Awad. 2018. Damage identification in social media posts using multimodal deep learning. In *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. ISCRAM, Rochester, NY, 529–543.

[68] Dat Tien Nguyen, Firoj Alam, Ferda Ofli, and Muhammad Imran. 2017. Automatic image filtering on social networks using deep learning and perceptual hashing during crises. In *Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management*. ISCRAM, Albi, France, 499–511.

[69] Dat Tien Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. 2017. Damage assessment from social media imagery data during disasters. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE/ACM, Sydney, Australia, 569–576.

[70] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. *SSRN Pre-print DOI:10.2139/ssrn.2886526.*

[71] Noelia Oses and Fadi Dornaika. 2013. Image-based delineation of built heritage masonry for automatic classification. In *Image Analysis and Recognition*, Mohamed Kamel and Aurélio Campilho (Eds.). Springer, Berlin, 782–789.

[72] Sharrona Pearl. 2015. *Images, Ethics, Technology*. Routledge.

[73] M. Pesaresi, A. Gerhardinger, and F. Haag. 2007. Rapid damage assessment of built-up structures using VHR satellite data in tsunami-affected areas. *Int. J. Remote Sens.* 28, 13–14 (July 2007), 3013–3036. DOI : https://doi.org/10.1080/01431160601094492

[74] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2015. Extracting situational information from microblogs during disaster events: A classification-summarization approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 583–592.

[75] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. DOI : https://doi.org/10.1007/s11263-015-0816-y

[76] Hyunjin Seo. 2014. Visual propaganda in the age of social media: An empirical analysis of Twitter images during the 2012 Israeli–Hamas conflict. *Visual Commun. Q.* 21, 3 (2014), 150–161.

[77] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. 2014. OverFeat: Integrated recognition, localization and detection using convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. CBLS, 16. http://openreview.net/document/d332e77d-459a-4af8-b3ed-55ba.

[78] Gayane Shalunts, Yll Haxhimusa, and Robert Sablatnig. 2011. In *Advances in Visual Computing*, George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Song Wang, Kim Kyungnam, Bedrich Benes, Kenneth Moreland, Christoph Borst, Stephen DiVerdi, Chiang Yi-Jen, and Jiang Ming (Eds.). Springer, Berlin, 280–289.

[79] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR*. 14. http://arxiv.org/abs/1409.1556.

[80] Dirk H. R. Spennemann. 1999. Cultural heritage conservation during emergency management: Luxury or necessity? *International J. Public Admin.* 22, 5 (1999), 745–804.

[81] Michelle Springer, Beth Dulabahn, Phil Michel, Barbara Natanson, David W. Reser, Nicole B. Ellison, Helena Zinkham, and David Woodward. 2008. For the common good: The Library of Congress Flickr pilot project. Library of Congress, Prints and Photographs Division, 55.

[82] Zuzana Stanton-Geddes and Salman Anees Soz. 2017. *Promoting Disaster Resilient Cultural Heritage*. Technical Report. World Bank, Washington, DC.

[83] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence (AAAI)*. 4278–4284.

[84] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9. DOI : https://doi.org/10.1109/CVPR.2015.7298594

[85] June Taboroff. 2003. Natural disasters and urban cultural heritage: A reassessment. In *Building Safer Cities – the Future of Disaster Risk*, Alcira Kreimer, Margaret Arnold, and Anne Carlin (Eds.). World Bank, Washington, DC, Chapter 16, 233–240.

[86] Aparna Tandon. 2017. Post-disaster damage assessment of cultural heritage: Are we prepared? In *Proceedings of the ICOM-CC 18th Triennial Conference*.

[87] Melissa Terras. 2011. The digital wunderkammer: Flickr as a platform for amateur cultural and heritage content. *Library Trends* 59, 4 (2011), 686–706.

[88] Tin Kam Ho. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 8 (Aug. 1998), 832–844. DOI : https://doi.org/10.1109/34.709601

[89] M. Turker and B. T. San. 2004. Detection of collapsed buildings caused by the 1999 Izmit, Turkey earthquake through digital analysis of post-event aerial photographs. *Int. J. Remote Sens.* 25, 21 (2004), 4701–4714. DOI : https://doi.org/10.1080/01431160410001709976

[90] Mustafa Turker and Emre Sumer. 2008. Building-based damage detection due to earthquake using the watershed segmentation of the post-event aerial images. *Int. J. Remote Sens.* 29, 11 (June 2008), 3073–3089.

[91] UNISDR. 2015. Sendai framework for disaster risk reduction 2015–2030. (2015).

[92] Anand Vetrivel, Markus Gerke, Norman Kerle, Francesco Nex, and George Vosselman. 2018. 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* 140 (June 2018), 45–59.

[93] Farida Vis, Simon Faulkner, Katy Parry, Yana Manyukhina, and Lisa Evans. 2013. Twitpic-ing the riots: Analysing images shared on Twitter during the 2011 UK riots. In *Twitter and Society*, K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann (Eds.). Vol. 89. New York:Peter Lang, 385–398. http://eprints.whiterose.ac.uk/79098/.

[94] Zhe Xu, Dacheng Tao, Ya Zhang, Junjie Wu, and Ah Chung Tsoi. 2014. Architectural style classification using multinomial latent logistic regression. In *Computer Vision − ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 600–615.

[95] Jie Yin, Sarvnaz Karimi, Bella Robinson, and Mark Cameron. 2012. ESA: Emergency situation awareness via microbloggers. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. ACM, New York, NY, 2701–2703. DOI : https://doi.org/10.1145/2396761.2398732

[96] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *NIPS*. 3320–3328.

[97] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the 13th European Conference Computer Vision (ECCV)*. 818–833. DOI : https://doi.org/10.1007/978-3-319-10590-1_53

[98] Luming Zhang, Mingli Song, Xiao Liu, Li Sun, Chun Chen, and Jiajun Bu. 2014. Recognizing architecture styles by hierarchical sparse coding of blocklets. *Inf. Sci.* 254 (2014), 141–154. DOI : https://doi.org/10.1016/j.ins.2013.08.020

[99] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (June 2018), 1452–1464. DOI : https://doi.org/10.1109/TPAMI.2017.2723009

[100] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2018. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 8697–8710.