Deep Learning Methods Towards Generalized Change Detection on Planetary Surfaces

Hannah R. Kerner, Kiri L. Wagstaff, Brian Bue, Patrick C. Gray, James F. Bell III, and Heni Ben Amor

***Abstract*—Ongoing Mars exploration missions are returning large volumes of image data. Identifying surface changes in these images, *e.g.*, new impact craters, is critical for investigating many scientific hypotheses. Traditional approaches to change detection rely on image differencing and manual feature engineering. These methods can be sensitive to irrelevant variations in illumination or image quality and typically require before and after images to be co-registered, which itself is a major challenge. To overcome these limitations, we propose novel deep learning approaches to change detection that rely on transfer learning, convolutional autoencoders, and Siamese networks. Our experiments show that these approaches can detect meaningful changes with high accuracy despite significant differences in illumination, image quality, imaging sensors, and alignment between before and after images. We show that the latent representations learned by a convolutional autoencoder yield the most general representations for detecting change across surface feature types, scales, sensors, and planetary bodies.**

I. Introduction

ONGOING Mars imaging investigations such as the High Resolution Imaging Science Experiment (HiRISE) [1] and ConTeXt Camera (CTX) [2] on the Mars Reconnaissance Orbiter are returning large volumes of image data that continue to grow faster than scientists can analyze and categorize them. There is a need for systems that can rapidly and intelligently analyze these data and prioritize observations of interest to scientists. Change detection, the process of automatically identifying changes in surface features between two images collected over the same location at different points in time, is a critical tool for analyzing these data. For example, recurring slope lineae (RSL) are narrow, low-albedo features that extend from bedrock to incrementally lengthen down steep slopes and are thought to be formed by shallow subsurface liquid water flows [3]. RSL appear to lengthen on timescales of several months and appear/disappear on timescales close to one year [3]. Because of their implications for the past and present history of water on Mars, RSL are key features that scientists are actively monitoring for changes as they develop and evaluate theories on RSL formation and growth mechanisms (*e.g.,* Figure 1A). A system that automatically detects RSL in images from HiRISE or other high-resolution imaging systems could help scientists better understand where RSL occur, how they evolve, and how or why they form in the first place.

New meteorite impacts are another key surface feature that scientists monitor for change. Daubar *et al.*, 2013 [4] reported the discovery of 248 new impact sites that formed on Mars within the last few decades (*e.g.*, Figure 1B). The landscape of Mars and other planets continues to be altered by impact events. Speyerer *et al.*, 2016 reported the discovery of 222 new impact sites on the Moon, which was 30% more impacts than was expected based on current estimates of the current cratering rate on the Moon [5] (*e.g.*, Figure 1C). Documenting when and where new impacts occur helps scientists to refine estimates of the past and present cratering rates in the solar system, which in turn enables improved age estimates of important events in the solar system's history [4], [5].

High spatial and temporal resolution imaging of the Earth by government and commercial satellites enables observation of countless surface features that change due to natural or human-induced processes. Detecting changes in surface features on the Earth—*e.g.*, new construction, fires, or volcanic eruptions (*e.g.*, Figure 1D)—is important for our scientific understanding of the Earth as well as for many commercial, humanitarian, and defense applications.

Many successful approaches exist for detecting changes in specific surface features or land cover types [5]–[9]. The most popular approaches are post-classification and difference-based methods. The main drawback of post-classification methods is that they require pre-trained classifiers with well-defined classes for each type of surface feature being monitored for change. Difference-based approaches require images to be precisely co-registered and are sensitive to irrelevant variations caused by illumination and processing artifacts [10].

The goal of our approach is to learn general representations of bi-temporal image pairs (*i.e.*, pairs of images acquired on two different dates) that are useful for identifying when changes in surface features have occurred on a planetary body. We introduce state-of-the-art deep learning methods that require a relatively small number of labeled training examples for learning these representations and for detecting if a surface feature change has occurred based on these representations. We test our approaches on Mars, the Moon, and Earth to evaluate the capacity for generalization to new surface features, imaging sensors, resolutions, level of co-registration, and planetary body, and show that difference-based approaches

H. R. Kerner is with the School of Earth and Space Exploration, Arizona State University, Tempe, AZ, 85282 USA e-mail: hkerner@asu.edu.

K. L. Wagstaff and B. Blue are with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, 91109.

P. C. Gray is with the Nicholas School of the Environment, Duke University, Durham, NC, 27710.

J. F. Bell III is with the School of Earth and Space Exploration, Arizona State University, Tempe, AZ, 85282 USA.

H. Ben Amor is with the School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, 85282.

This research was carried out (in part) at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration and funded through the Internal Strategic University Research Partnerships (SURP) program.

Manuscript submitted on 21 December 2018.

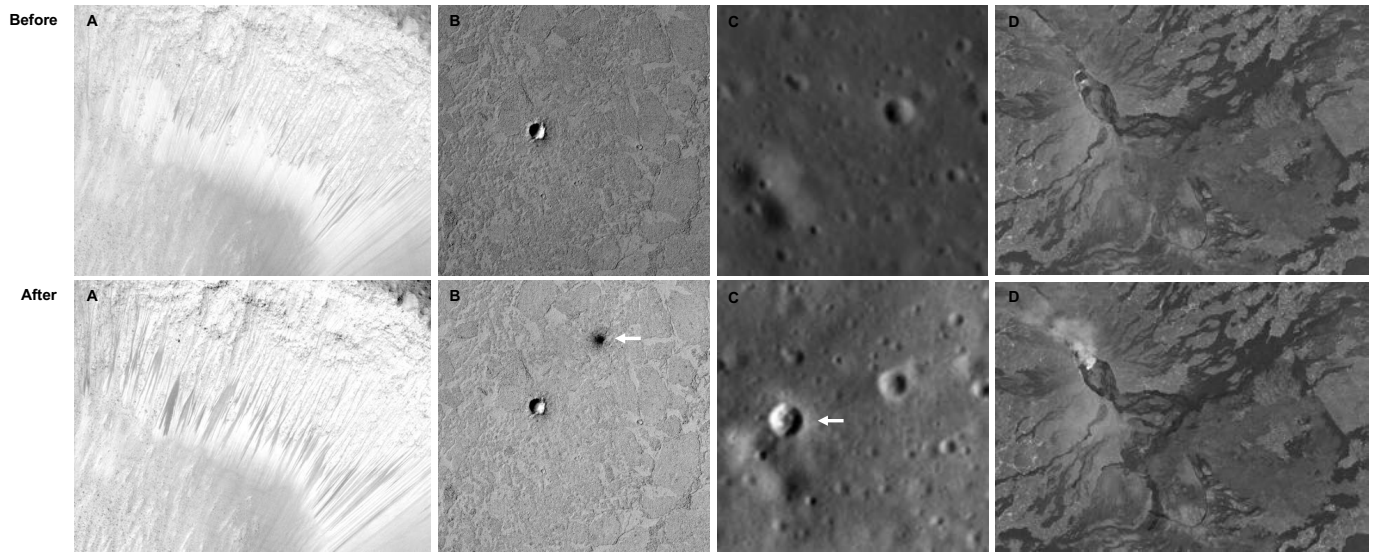

Fig. 1. Example before and after images from the (A) HiRISE RSL, (B) CTX Impacts, (C) LROC Impacts, and (D) Planet Earth datasets. Image IDs from Table I ESP_030769_1685, ESP_031059_1685, P06_003451_2035_XN_23N171W, P13_006286_2073_XN_27N171W, M181330922L, and M1139065512L, ertaale-20170116-before, and ertaale-20170123-after.

may be limited in their ability to generalize to new datasets.

II. Previous Work

Tewkesbury *et al.*, 2015 [11] proposes organizing change detection literature in terms of the *comparison method*, describing the method used to determine if a change in surface features has occurred between two images, and the *unit of analysis*, describing the image representation that will be analyzed by the comparison methods. Raw pixel intensities, obtained from images that have had little to no pre-processing applied, are the most common units of analysis and have been used widely since the beginning of remote sensing change detection research (*e.g.,* [11]–[16]). Difference images, in which each pixel represents the difference between corresponding pixel intensities in a pair of images, are also a common unit of analysis [11]. Thresholding is a common change detection method in approaches where the unit of analysis is the difference in pixel values (*e.g.,* [17]). If the difference between two images for one pixel location is greater than a threshold t, that pixel is classified as a *change* pixel. If the difference is less than or equal to t, the pixel is classified as a *no-change* pixel. This threshold can be derived empirically as one (or more) standard deviations from the mean pixel difference in a distribution of pixel differences across each band of an image [18]. When comparing pixel values between before and after images directly, it is common to use a threshold on the ratio between the corresponding pixel values from the before and after image to classify pixels that constitute change in surface features as in [5]. While these approaches are computationally fast, they are sensitive to changes caused by mis-registration, illumination, and image artifacts that are not relevant for assessing surface feature change [10].

As the spatial resolution of satellite cameras improved and pixels sampled smaller ground distances, researchers began developing higher-level representations from pixels that were more suitable for detecting land cover classes, made comparison methods less sensitive to mis-registration, and reduced noise that contributed to false detections [19]. These representations include kernel filters or windows, image objects, change vectors, and classifier labels (*e.g.,* [11], [18]–[25]). Land-cover classifiers have been used widely to produce labels that can be compared between before and after images, *e.g.*, [6]–[9], [26]. Comparing the land cover classes predicted for the before and after image by a classifier is known as post-classification comparison and is widely used in the remote sensing change detection literature [11], [18]. Post-classification approaches that produce a single label for each image being compared may not suffer from mis-registration issues in the post-classification comparison, but the change detection accuracy depends directly on the classifier accuracy. Post-classification approaches that produce pixel-wise labels *do* suffer from mis-registration issues in post-classification comparison because the labels are at the same resolution of the input image. The unique advantage of this approach is that the semantics of the change are provided with the detection. The primary disadvantages are that performance scales with the accuracy of the land-cover classifier and it is difficult and computationally expensive to learn a new classifier to predict land-cover classes for new applications and image datasets.

III. Datasets

A. HiRISE: Recurring Slope Lineae (RSL)

Recurring slope lineae (RSL) are dark, narrow features (typically 0.5-5 m) that incrementally lengthen down steep slopes and fade/recur throughout the year. They are thought to be formed by shallow subsurface liquid water flows [3]. Scientists are actively monitoring RSL for changes as they develop and evaluate theories on RSL formation and growth mechanisms (*e.g.*, Figure 1A). We created a dataset for change detection of RSL using repeat observations of a well studied site with

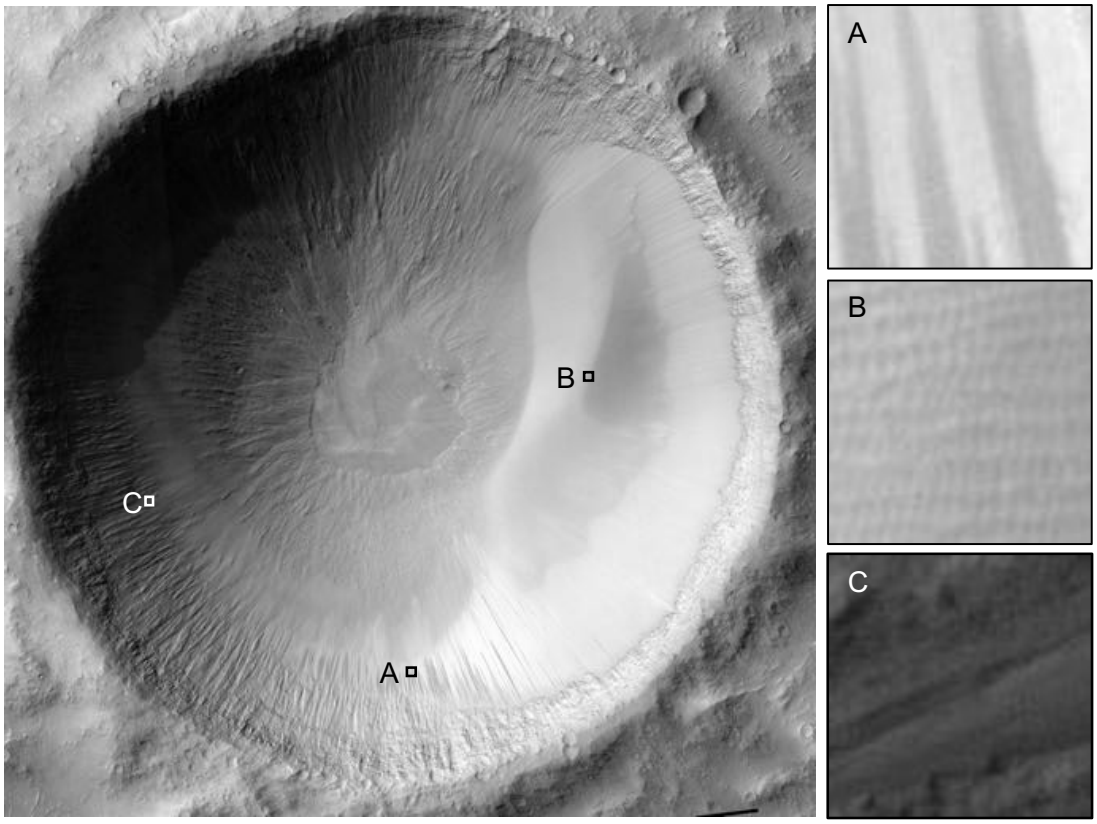

Fig. 2. Example of 100×100-pixel tiles sampled from HiRISE image of Garni crater (ESP_027802_1685) where RSL occur (A), where RSL do not occur (B), and where it is difficult to distinguish RSL from other shadowed topography (C).

known RSL activity called Garni Crater in Valles Marineris [27]. These observations were made by the High Resolution Imaging Science Experiment (HiRISE) camera onboard the Mars Reconnaissance Orbiter. HiRISE has a spatial resolution of ∼30 cm/pixel and three color channels: red (550-850 nm), blue-green (400-600 nm), and near infrared (800-1000 nm) [1]. We used the HiRISE red channel orthorectified products in Table I corresponding to 12 different acquisition dates. We used the red band because it has the highest spectral coverage and signal-to-noise ratio [1]. We cropped each image to the $10,000 \times 10,000$-pixel boundary of Garni crater and converted to grayscale. Since RSL are relatively small, we sampled 100×100-pixel tiles from the cropped images of Garni crater for our experiments (*e.g.*, Figure 2). We sampled using a sliding window with a stride size of 50 pixels. This resulted in a dataset of 39,601 tiles for each of the 12 images (475,212 total).

We selected three bi-temporal pairs that exhibited changes in RSL for labeling: ESP_027802_1685-ESP_028501_1685, ESP_030347_1685-ESP_030769_1685, and ESP_029213_1685-ESP_029780_1685. We selected bounding boxes surrounding the changed regions and labeled bi-temporal tiles as having *change* or *no-change* within these bounds by inspecting animated GIFs transitioning between before and after tiles. We labeled 203 *change* and 3,333 *no-change* tiles for ESP_027802_1685-ESP_028501_1685, 73 *change* and 613 *no-change* tiles for ESP_030347_1685-ESP_030769_1685, and 75 *change* and 179 *no-change* tiles for ESP_029213_1685-ESP_029780_1685 (Table II).

These sets of un-labeled and labeled image tiles were used to train the convolutional autoencoder and change detection classification models, as we describe in more detail in Section IV. In Table I, Column 4 indicates which (un-labeled) images were used to train the autoencoder and which (labeled) images were used to train the classification models. Columns 5-6 indicate which (labeled) images were used for validation and testing with respect to the classification models.

B. CTX: Meteorite Impacts

The landscape of Mars and other planets is continually altered by meteorite impacts (Figure 1B). Scientists use data about when and where these impacts occur to refine estimates of the current cratering rate in our solar system and constrain the impact production function over time [4]. In one study of contemporary impact cratering, Daubar *et al.* 2013 [4] reported 248 new impact sites that formed on Mars within the last few decades discovered by comparing images taken over the same location at different times from multiple instrument datasets. We selected eight image pairs (16 images) from Daubar *et al.*'s study in which both the before and after images were taken by the ConTeXt Camera (CTX) onboard the Mars Reconnaissance Orbiter (Table VIII). CTX has a spatial resolution of ∼6 m/pixel and a single channel (500–700 nm) [2]. We map projected and co-registered each image pair

Instrument	Image ID	Date Acquired	Training		Validation	Testing
			Autoencoder	Classifiers		
HiRISE	ESP_027802_1685	07/01/2012		•		
HiRISE	ESP_028501_1685	08/25/2012		•		
HiRISE	ESP_029213_1685	10/19/2012	•			•
HiRISE	ESP_029780_1685	12/02/2012				•
HiRISE	ESP_030347_1685	01/16/2013			•	
HiRISE	ESP_030769_1685	02/18/2013			•	
HiRISE	ESP_031059_1685	03/12/2013	•			
HiRISE	ESP_031771_1685	05/07/2013	•			
HiRISE	ESP_032048_1685	05/28/2013	•			
HiRISE	ESP_032615_1685	07/11/2013	•			
HiRISE	ESP_034184_1685	11/11/2013	•			

TABLE I
IMAGE PRODUCTS USED IN THIS STUDY. BULLETS INDICATE WHICH PART OF THE STUDY THE IMAGE WAS USED FOR.

Before Image	After Image	*Change*	*No-Change*
ESP_027802_1685	ESP_028501_1685	203	3,333
ESP_030347_1685	ESP_030769_1685	73	613
ESP_029213_1685	ESP_029780_1685	75	179

TABLE II
NUMBER OF EXAMPLES LABELED IN EACH CLASS FOR THREE IMAGE PAIRS FROM HIRISE RSL DATASET.

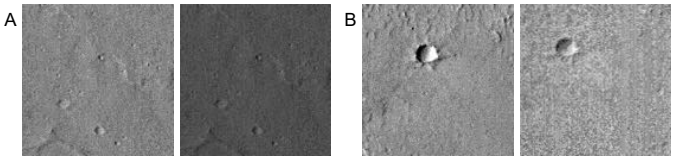

Fig. 3. Example CTX tiles where before and after images are co-registered perfectly (A) and not co-registered perfectly (B). B: The large crater in the first image is shifted up and to the left in the second image, and many small details in the first image do not appear in the second image). Both tiles were cropped from the same two images: P09_004477_1906_XN_10N100W and P13_006178_1907_XN_10N100W (Table VIII).

using the PIXL Visual Precision Targeting (VPT) algorithm [28]. While this algorithm produces good co-registration for most regions in the image, some regions are not perfectly co-registered (*e.g.*, Figure 3). For each image pair, we cropped one 150×150-pixel tile around the approximate center of the impact and six 150×150-pixel tiles at random locations in the non-impact region of the image, then resized the tiles to 100×100 pixels. We augmented the eight *change* tiles with horizontally flipped, vertically flipped, and 90-, 180-, and 270-degree rotated versions of each tile. We did not augment the *no-change* tiles. This resulted in 48 positive (*change*) and 48 negative (*no-change*) 100×100-pixel image pairs in this dataset. We used this dataset to assess generalization to surface features, image sensors, and resolution that were different than in the training set.

C. LROC: Impacts

Speyerer *et al.* [5] found that the current lunar cratering rate is significantly higher than previously thought. These new estimates, informed by updated images of the lunar surface and automatic classification methods, improved models of current cratering rates and surface regolith turnover. These models are used to constrain the ages of surface features on the Moon and other planetary bodies. To test the transfer performance of the proposed change detection approaches, we created a dataset of lunar impacts captured using the Lunar Reconnaissance Orbiter Camera's (LROC) Narrow Angle Camera (NAC) [29]. LROC has two identical NACs (for stereo imaging) that collect $\sim$0.5 m/pixel panchromatic images. Since the LROC mission began, the NACs have collected a repository of over 1.7 million 500-megapixel images[1]. A subset of these images were captured over the same region under similar illumination conditions at different points in time. We selected five of these image pairs in which an impact occurred due to a meteorite or the Chang'e lander [30] as reported in Speyerer *et al.* [5] (*e.g.*, Figure 1C). In Table VIII we give the NAC ID for each image as well as the name used by the LROC mission for each pair in parentheses. These images were map-projected, radiometrically calibrated, and co-registered to within 20 m ($\sim$40 pixels). For each pair of images, we cropped one 100×100-pixel tile around the approximate center of the impact and one 100×100-pixel tile at a random location in the non-impact region of the image. This resulted in five *change* and five *no-change* 100×100-pixel image pairs in this dataset. We used these images to assess generalization to a different planetary body (Experiment 3).

D. Planet: Earth

The Earth's surface is continually undergoing change from natural geologic processes as well as human activity. Planet, Inc. is a commercial remote sensing company that operates a constellation of small satellites ("cubesats") that acquire daily color images of the Earth's surface using their PlanetScope camera [31]. These images have $\sim$3-m spatial resolution and four bands: red, green, blue, and near-infrared. The sun-synchronous orbit of the satellites enables an equatorial overpass time between 9:30 and 11:30 a.m. local time [32]. We manually identified before and after images of four different locations taken at different points in time that showed changes in surface features on the Earth (*e.g.*, Figure 1D); we give the names of these images in Table VIII.[2] We chose desert regions where the landscape would be most similar to the Moon and Mars (Nevada, Saudi Arabia,

[1] http://wms.lroc.asu.edu/lroc/thumbnails

[2] We selected images using the Planet Gallery tool: https://planet.com/gallery

Ethiopia, and the Gobi desert). These images exhibit surface feature change due to fire, oil production, volcanic eruptions, and solar array construction. We cropped one 200×200-pixel tile around a *change* feature in each before/after pair and one 200×200-pixel tile in a region of the image pair without changes in surface features, then resized the tiles to 100×100 pixels.[3] We converted all images to grayscale. This resulted in five *change* and five *no-change* 100×100-pixel image pairs, which we used to assess generalization to a different planetary body (Experiment 3).

The source images for these four datasets are publicly available and we provide instructions for accessing them in Appendix A. We will make the datasets publicly available at 10.5281/zenodo.2373798.

IV. Approach

A. Change Detection Methods

We approach change detection as a binary classification problem. Given an input containing information about a bi-temporal pair of images, a binary classifier should predict 1 if there is a change in surface features, and predict 0 if there are no changes in surface features. Surface changes are changes to morphologies on the surface that are present in the image, *e.g.*, RSL or impact craters. There might be other differences between a pair of images, *e.g.*, illumination or resolution differences, that are not considered surface changes and should not result in a change detection. We propose two methods for change detection via binary classification: a fine-tuned deep neural network and a Siamese neural network.

1) Fine-tuned Neural Network: Deep neural networks are very effective at building hierarchical representations of complex data. Previous work has shown that the features learned by a neural network for one task can be useful for another, unrelated task (*e.g.*, [33], [34]). We initialized the Inception-V3 network [35] with weights learned from pre-training with the ImageNet database [36] and replaced the final softmax layer of Inception-V3 with a new softmax layer with two outputs (for *change* and *no-change*). We fine-tuned the network for our change detection task by optimizing weights for this new softmax layer using our change detection training and validation datasets.

2) Siamese Network: A Siamese network is a type of neural network that learns to distinguish between pairs of inputs that are similar and pairs of inputs that are dissimilar [37]. Siamese network approaches have been proposed for signature verification [37], face verification [38], dimensionality reduction [39], and one-shot learning [40]. A Siamese network consists of two identical ("Siamese") networks that each extract feature vectors for one image in the input pair. These feature extraction networks are then "conjoined" by a layer that computes the distance between the two feature vectors (*e.g.*, L1 distance) followed by a sigmoid (or other activation) layer that learns to separate distances between similar and dissimilar pairs.

[3]We cropped two tile pairs with and without change from the Erta Ale, Ethiopia image pair, hence we refer to these as "Erta Ale 1" and "Erta Ale 2."

Network	AUC
Xception	0.99
VGG-16	0.96
Inception-ResNet-V3	0.88
VGG-19	0.78
Inception-V3	0.75

TABLE III
AUC scores for pre-trained networks we compared for feature extraction in the Siamese network approach.

We propose that bi-temporal pairs in which a surface change has occurred are analogous to "dissimilar" image pairs in previous work and bi-temporal pairs without surface change are analagous to "similar" pairs. We compared several network architectures, initialized with weights learned from pre-training on the ImageNet database [36], for the feature extraction component of the Siamese network: Inception-V3 [35], Xception [41], Inception-ResNet-V2 [42], VGG-16 [43], and VGG-19 [43]. In each network, we replaced the final softmax layer with a global average pooling layer and a dense layer of 1,024 neurons. We used the L1 distance at the distance layer of the Siamese network followed by a sigmoid (single neuron) layer for binary classification of *change* or *no-change*. We trained the network by optimizing the weights connecting Xception to the dense (feature vector) layer and the L1 distance layer to the sigmoid layer. We trained each model until the validation loss converged and found that using Xception for feature extraction yielded the best change detection performance. Thus, we used Xception for feature extraction in subsequent Siamese network experiments.

B. Image Representations

The classification methods we propose for change detection each require 3-channel input images, which are typically RGB color images. In the fine-tuned neural network approach, our goal is to represent a pair of images as a single three-channel image that can be classified directly. We consider four representations of before/after image pairs to constitute these three channels: grayscale, absolute difference, signed difference, and autoencoder bottleneck maps (Figure 4). These representations are analogous to the "units of analysis" discussed in Section II.

1) Grayscale: The datasets described in Section III contain 100×100-pixel grayscale images, and we can use this representation directly. Since the input to Inception-V3 is a single 3-channel image, this representation is a composite of the before and after grayscale images where the blue channel contains the before image, the green channel contains the after image, and the red channel contains all zeros. We refer to this representation as "composite grayscale." Since the input to the Siamese network is a before and after pair of 3-channel images, each 3-channel image contains a copy of one grayscale image in each channel.

2) Absolute Difference: In this approach, we compute the absolute value of the difference between each pixel in the grayscale before and after images. The resulting single-channel image is replicated in each of the three input channels. This

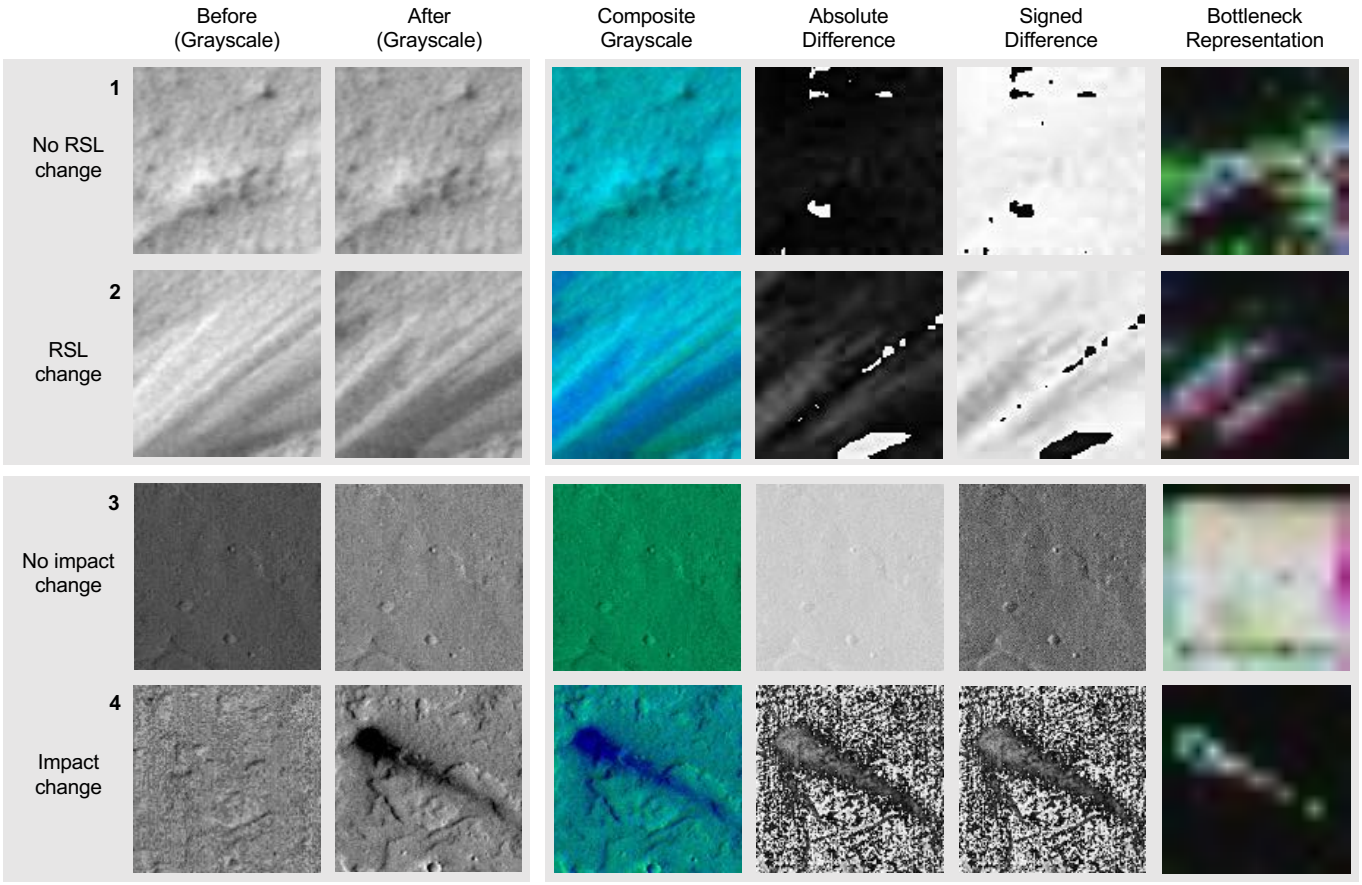

Fig. 4. Top section: Four representations of two example image pairs from the HiRISE RSL dataset, one *without* RSL change (row 1) and one *with* RSL change (row 2). Bottom section: Four representations of two example image pairs from the CTX Fresh Impacts dataset, one *without* impact change (row 3) and one *with* impact change (row 4).

representation is not used for the Siamese network, which requires a pair of images.

3) Signed Difference: As in the previous approach, we compute the difference between each pixel in the grayscale before and after images. Because image formats do not allow negative values, we re-scale the difference values (which nominally range from -255 to $+255$) into the range $[0, 255]$. The resulting single-channel image is replicated in each of the three input channels. This representation is not used for the Siamese network.

4) Autoencoder Bottleneck Maps: A convolutional autoencoder (CAE) is a type of self-supervised neural network that learns a low-dimensional representation or "code" capturing the most salient features in a dataset by optimizing its reconstruction of the original input from the learned encoding, using convolutional layers for feature extraction [44]. Autoencoders are sometimes called "encoder-decoders" because they consist of an encoder network that projects the input into the low-dimensional "bottleneck" representation and a decoder network that projects from the bottleneck representation back up to the input space (Figure 5). Once an autoencoder is trained to reconstruct all examples in a dataset well, the encoder network can be used for dimensionality reduction. This is similar to dimensionality reduction through projection into the eigenspace in principal component analysis [45]. By training an autoencoder to produce representations at the bottleneck layer that capture the salient features in the image, we can take advantage of the large dataset of *un-labeled* tiles in the HiRISE RSL dataset that to produce a potentially more refined representation for the downstream change detection classifier.

We trained a convolutional autoencoder with images indicated in Column 4 of Table I. The encoder part of the network consists of three sequences of a 3×3 convolution, batch normalization [46], and 2×2 max pooling. In the decoder part of the network, we use the same three sequences but instead of max pooling we upsample using nearest-neighbor interpolation. We generated bottleneck map representations by applying only the encoder function of the trained network. Table IV shows the size of image representations following each layer in the autoencoder. The dimension at the bottleneck layer is $13 \times 13 \times 8$, which is $\sim$7.4 times smaller than the input dimension. We selected the three (of eight) bottleneck maps that were most discriminative for change in surface features to populate the three input channels to Inception-V3 and the Siamese network. To determine which maps were most discriminative, for each of the eight bottleneck maps we computed two distributions: one of the mean squared error between the before and after image maps for the *no-change* examples in the validation dataset, and one for the *change*

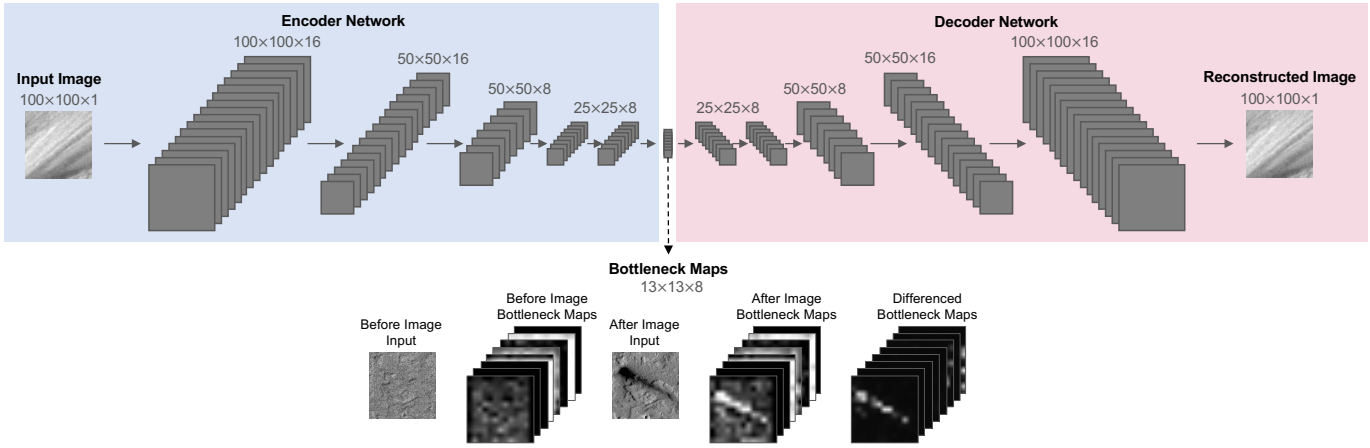

Fig. 5. Convolutional autoencoder architecture.

Layer Name	Layer Output Size
Input	$100 \times 100 \times 1$
3×3 Convolution (1)	$100 \times 100 \times 16$
Batch Normalization (1)	$100 \times 100 \times 16$
Max Pool (1)	$50 \times 50 \times 16$
3×3 Convolution (2)	$50 \times 50 \times 8$
Batch Normalization (2)	$50 \times 50 \times 8$
Max Pool (2)	$25 \times 25 \times 8$
3×3 Convolution (3)	$25 \times 25 \times 8$
Batch Normalization (3)	$25 \times 25 \times 8$
Max Pool (3)	$13 \times 13 \times 8$*
3×3 Convolution (4)	$13 \times 13 \times 8$
Batch Normalization (4)	$13 \times 13 \times 8$
Up Sample (1)	$25 \times 25 \times 8$
3×3 Convolution (5)	$25 \times 25 \times 8$
Batch Normalization (5)	$25 \times 25 \times 8$
Up Sample (2)	$50 \times 50 \times 8$
3×3 Convolution (6)	$50 \times 50 \times 16$
Batch Normalization (6)	$50 \times 50 \times 16$
Up Sample (3)	$100 \times 100 \times 8$
3×3 Convolution (7)	$100 \times 100 \times 1$

TABLE IV
OUTPUT SIZE OF EACH LAYER IN CONVOLUTIONAL AUTOENCODER. ASTERISK INDICATES BOTTLENECK REPRESENTATION.

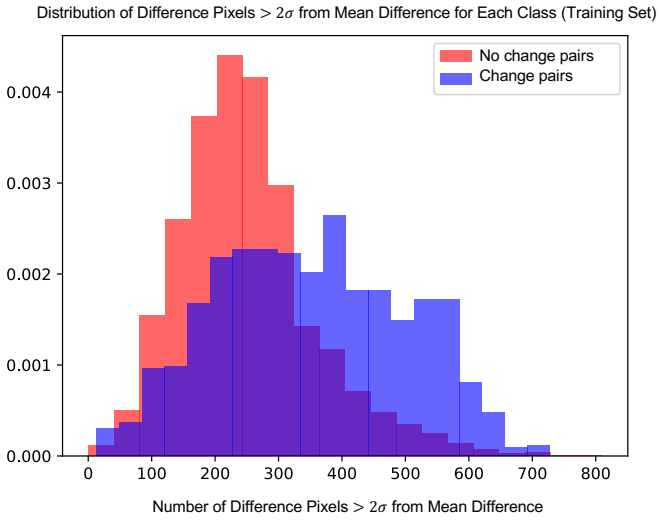

Fig. 6. Distribution of difference pixels greater than two standard deviations from the mean in each image pair for *change* and *no-change* examples.

examples in the validation dataset. We computed the Kullback-Leibler (KL) divergence between these two distributions for each of the eight maps and selected the three maps with the largest KL divergence. We up-sampled each of these three feature maps to 100×100 pixels and populated the three input channels with these three maps.

C. Baseline Change Detection

Approaches based on image differencing are the most widely used change detection approaches [18]. Since there is no widely accepted approach for difference-based change detection, we designed the following approach based on common methods:

1) Apply local contrast normalization to before and after images using the Contrast Limited Adaptive Histogram Equalization algorithm [47]
2) Compute pixelwise difference between before and after images in training set
3) Compute mean and standard deviation of distribution of pixelwise differences for each image pair
4) Compute number of pixels n in which the pixelwise difference is greater than two standard deviations from the mean pixel difference for each image pair
5) Fit conditional Gaussian probability distributions $p(x = n|y = \textit{change})$ and $p(x = n|y = \textit{no-change})$ (Figure 6)
6) Use Naive Bayes [48] to predict class label for test examples using posterior probabilities $p(y = \{\textit{change}, \textit{no-change}\}|x = n)$

We perform each experiment in Section V using this baseline method *with* local contrast normalization as above and *without* local contrast normalization ("No LCN").

V. EXPERIMENTS

We performed three experiments to evaluate the performance of the change detection approaches described in Section IV. We trained all models using the TensorFlow [49] and

Classification Method	Representation	FPR at 5% FNR	Accuracy	AUC
Inception-V3	**Absolute Difference**	**7.8%**	**94.5%**	**0.984**
Inception-V3	Bottleneck Representation	93.3%	70.1%	0.464
Inception-V3	Composite Grayscale	40.8%	89.8%	0.898
Inception-V3	Signed Difference	10.6%	91.7%	0.974
Siamese Network	Bottleneck Representation	91.1%	66.1%	0.570
Siamese Network	Composite Grayscale	81.0%	86.6%	0.827
Naive Bayes	**Pixel Difference $> 2\sigma$**	**2.2%**	**94.5%**	**0.991**
Naive Bayes	Pixel Difference $> 2\sigma$ - No LCN	29.6%	92.9%	0.958

TABLE V
PERFORMANCE METRICS FOR EXPERIMENT 1: TRAIN AND TEST ON HIRISE RSL DATASET.

Keras[4] deep learning frameworks to minimize validation loss. We describe the details of training, including number of training steps and hyperparameter settings, in Appendix B. Area under the curve (AUC) for receiver operating characteristics (ROC) curves gives a better indication of model performance than accuracy, since accuracy depends on a chosen threshold on the posterior probability. Because we want to detect as many tiles that contain true surface changes as possible, even at the expense of more false positives, a low false negative rate (FNR) is more important than a low false positive rate (FPR), or even an overall high classification accuracy or AUC. For this reason, we also report the FPR at 5% FNR to assess which approach will yield the fewest false positives given a 5% maximum FNR.

A. Experiment 1: Generalization from Train to Test Set

In this experiment, we evaluated the ability of each approach described in Section IV to generalize from detecting changes in one type of surface feature in a training dataset to detecting changes in that feature in a held-out test set. We used the labeled image pairs from the HiRISE RSL dataset described in Section III. We augmented the positive-labeled (*change*) image pairs from the training and validation sets with horizontal flips, vertical flips, and 90-, 180-, and 270-degree rotations, resulting in 1,218 positive training examples and 438 positive validation examples. We used the labeled bi-temporal image pairs ESP_027802_1685 and ESP_028501_1685 for training, ESP_030347_1685 and ESP_030769_1685 for validation, and ESP_029213_1685 and ESP_029780_1685 for testing (Table I). We chose to partition the training, validation, and test sets by image pair rather than a random sample from all available images for two reasons. First, the RSL-forming region is located on the southern wall of Garni crater in the training pair, but on the northeastern wall of Garni crater in the validation and test pairs. By sampling the training set and validation/test sets from different geographic locations within Garni crater, we avoid overestimating the classifier's performance with spatial overlap between the training and validation/test sets. Second, separating the datasets by acquisition date represents how our change detection approach would be used in practice during a science mission. Table V gives performance metrics and Figure 7 shows the ROC curves for each classification approach in this experiment.

[4]F. Chollet, "Keras," 2015. Available: https://keras.io.

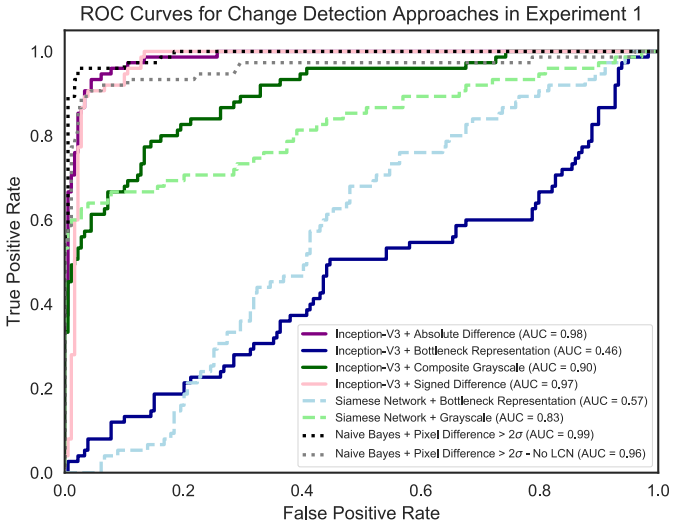

Fig. 7. ROC curves for Experiment 1: train and test on HiRISE RSL dataset.

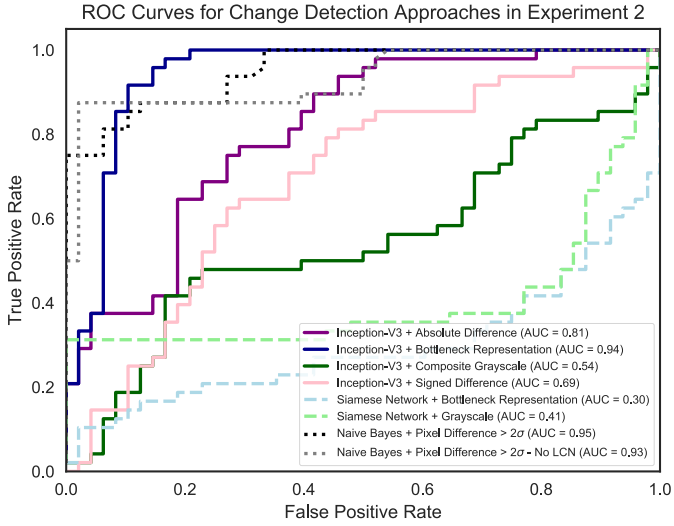

Fig. 8. ROC curves for Experiment 2: train on HiRISE RSL dataset, test on CTX Meteorite Impacts dataset.

B. Experiment 2: Generalization to New Change Type, Instrument, and Mis-Registration

In this experiment, we evaluated the ability of each proposed approach to generalize to new geographic locations on Mars (outside of Garni crater), a new type of surface feature change (meteorite impacts), a new instrument (CTX), and realistic

Classification Method	Representation	FPR at 5% FNR	Accuracy	AUC
Inception-V3	Absolute Difference	50.0%	71.9%	0.807
Inception-V3	**Bottleneck Representation**	**14.6%**	**75.0%**	**0.942**
Inception-V3	Composite Grayscale	97.9%	55.2%	0.545
Inception-V3	Signed Difference	85.4%	54.2%	0.691
Siamese Network	Bottleneck Representation	100.0%	52.1%	0.298
Siamese Network	Composite Grayscale	97.9%	64.6%	0.406
Naive Bayes	**Pixel Difference $> 2\sigma$**	**33.3%**	**52.08%**	**0.952**
Naive Bayes	Pixel Difference $> 2\sigma$ - No LCN	50.0%	55.21%	0.931

TABLE VI
PERFORMANCE METRICS FOR EXPERIMENT 2: TRAIN ON HIRISE RSL DATASET, TEST ON CTX METEORITE IMPACTS DATASET.

mis-registration compared to the training dataset. We used the same models that were trained for Experiment 1. We used the CTX Meteorite Impacts dataset of 96 images (48 *change*, 48 *no-change*) as the test set. Table VI gives performance metrics and Figure 8 shows the ROC curves for each classification approach in this experiment.

C. Experiment 3: Generalization to New Body

In this experiment, we evaluated the ability of the proposed approaches to generalize to an entirely new planetary body. Specifically, we wanted to test how general the representations learned by the autoencoder were for change detection. We used the same training set as in Experiments 1-2 of HiRISE RSL image pairs. We tested images of surface changes on the Moon captured by the Lunar Reconnaissance Orbiter Camera (LROC) and on the Earth captured by Planet satellites as described in Section III. For each dataset, we evaluated five image pairs *with* surface changes and five image pairs without. In the LROC dataset, surface changes are the result of meteorite impacts and a spacecraft landing. These images are map-projected but not co-registered, so matching features are mis-registered by as many as 40 pixels. In the Planet dataset, surface changes are the result of natural geologic processes (*e.g.*, lava flows) as well as some anthropogenic processes (*e.g.*, solar array construction). These image pairs are precisely co-registered and have very similar lighting conditions between before and after images. Figure 11 and Figure 12 show the images we evaluated in this experiment for the LROC and Planet datasets respectively. The third row of each section in these figures shows the difference between the autoencoder bottleneck representation of the before image and the after image for each tested pair, and the outline of each image indicates if it was correctly classified (green) or mis-classified (red).

VI. DISCUSSION

A. Summary of Findings

We found in Experiment 1 that the Inception-V3 network fine-tuned using absolute difference image representations of HiRISE RSL images and the baseline method (Naive Bayes) exhibited the best change detection performance on the test set of HiRISE RSL images from a spatially distinct region of Garni crater. The baseline method achieved slightly higher AUC and lower FPR at 5% FNR values than Inception-V3 with absolute difference representations, and these scores were closely followed by fine-tuned Inception-V3 using signed difference image representations (Figure 7, Table V). However, when local contrast normalization (LCN in Figure 7 and Table V) is *not* applied prior to processing with Naive Bayes, our methods outperform the baseline by a significant margin. This suggests that our deep learning approaches are able to *automatically* learn the appropriate pre-processing of the input on their own and could be used successfully without the need for pre-processing in change detection pipelines.

The purpose of Experiment 2 was to test how sensitive the input image representations and the features learned by each classification method were to the type of surface feature undergoing change, the sensor that collected the images, and realistic mis-registration. Using a test set where these properties were different from the training set, we found that the Inception-V3 network fine-tuned using bottleneck representations yielded the best change detection performance overall in this scenario. While the baseline method achieved a slightly higher AUC score (by 0.01), the bottleneck approach yielded the lowest FPR at 5% FNR (14.6%) as well as the highest accuracy of the tested approaches (Figure 8, Table VI).

In Experiment 3, we evaluated how general the latent representations learned by the autoencoder were for different types of features, sensors, scales, and planetary bodies. The surface features in both the LROC Impacts and Planet Earth datasets were different than those in training, especially for the Earth examples, as were the imaging sensor and level of mis-registration. We found that the Inception-V3 network fine-tuned with differenced latent representations of HiRISE RSL image pairs correctly classified all Planet Earth images and three out of five LROC Impacts images tested (Figure 12 and Figure 11).

B. Autoencoder Bottleneck Representations

It is interesting that the bottleneck representations yielded the lowest performance when training and testing on similar examples (Experiment 1), but the highest performance when the train and test examples were significantly different (Experiments 2-3). Surface changes in the CTX Impacts, LROC Impacts, and Planet Earth datasets mostly follow a pattern in which the feature is completely absent from one image in the pair and then appears in the other image of the pair (see Figure 4, 11, and 12). In the HiRISE RSL dataset, changes tend to be more gradual and often manifest as *growing* or *receding* rather than appearing on a blank slate (see Figures 1 and 4).

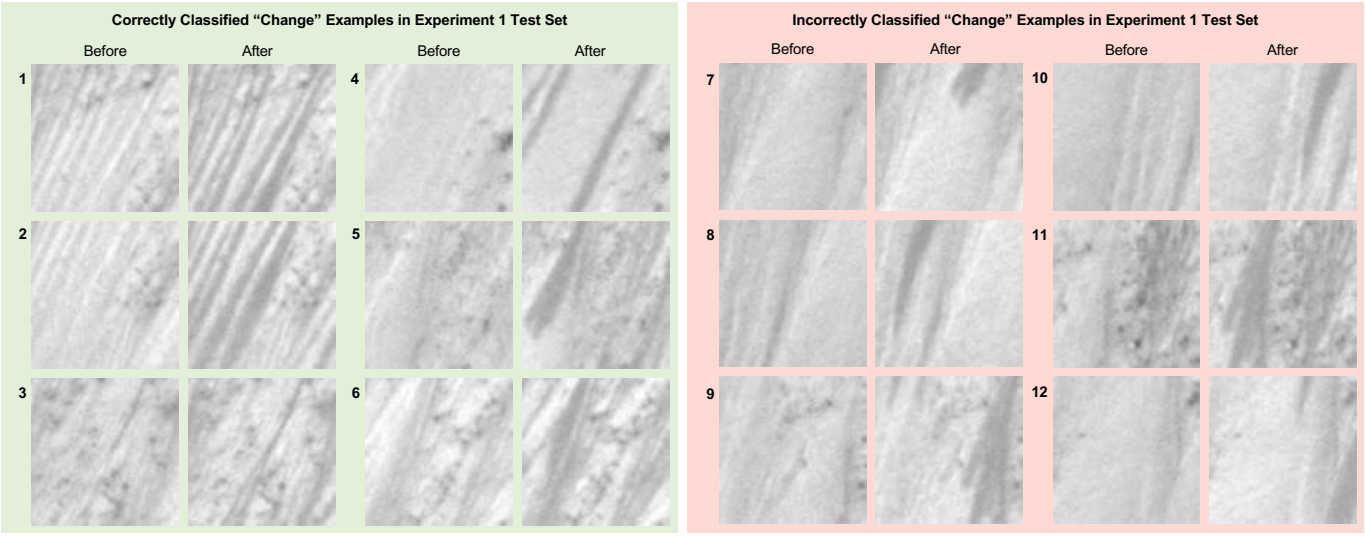

Fig. 9. Examples from Experiment 1 test set that our Inception-V3 with bottleneck representations approach assigned the *highest* probability of *change* (left) and *lowest* probability of *change* (right). The true label for all examples shown is *change*, thus examples on the right were misclassified.

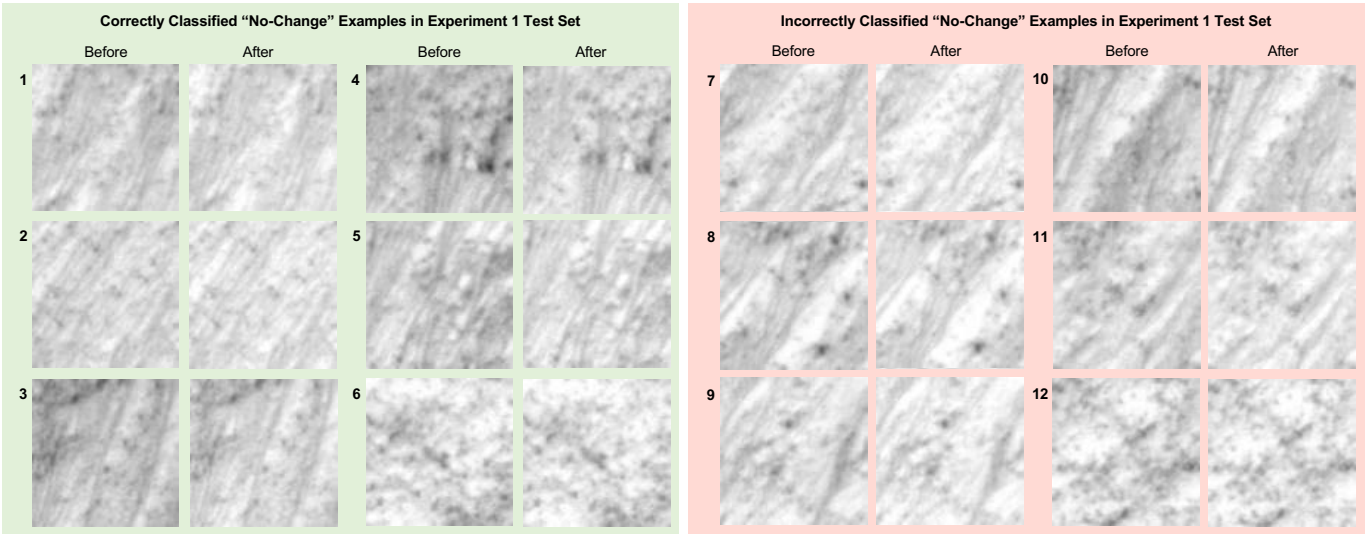

Fig. 10. Examples from Experiment 1 test set that our Inception-V3 with bottleneck representations approach assigned the *lowest* probability of *change* (left) and *highest* probability of *change* (right). The true label for all examples shown is *no-change*, thus examples on the right were misclassified.

Thus, one hypothesis is that the autoencoder is learning similar representations for both the before and after images in these cases such that the *differenced* bottleneck maps are not as suitable for change detection as they are in the other datasets. In Figure 9, we show six images from the *change* class in the test set in Experiment 1 that our Inception-V3 with bottleneck representations approach assigned the *highest* probability of *change* (left) and the *lowest* probability of *change* (right). In Figure 10, we show six images from the *no-change* class that our approach correctly (left) and incorrectly (right) classified. While this "gradual change" hypothesis could explain some of the misclassifications, it does not explain all or even most of them. In some of the misclassified *change* examples, the RSL change is on the edge and nearly out of the frame, whereas features in the other three datasets are (by design) located in the center of the frame. This would suggest that the autoencoder encoding function is less effective in representing features that are only partially in the frame.

C. Generalization to Different Planetary Bodies

In Experiment 3, we tested how sensitive the change detection approaches were to the planetary body being studied with test images of the lunar surface and the Earth's surface instead of the Martian surface seen during training. The lunar dataset also had the additional challenges from Experiment 2 in that the feature type undergoing change (meteorite and spacecraft impacts), sensor type (LROC), and level of mis-registration (up to 40 pixels) were different than in the training images. The images in the Earth dataset were precisely co-registered as in the training dataset, but the feature type and sensor type differed significantly. Comparing the bottleneck representations of the mis-classified LROC images (Figure 11,

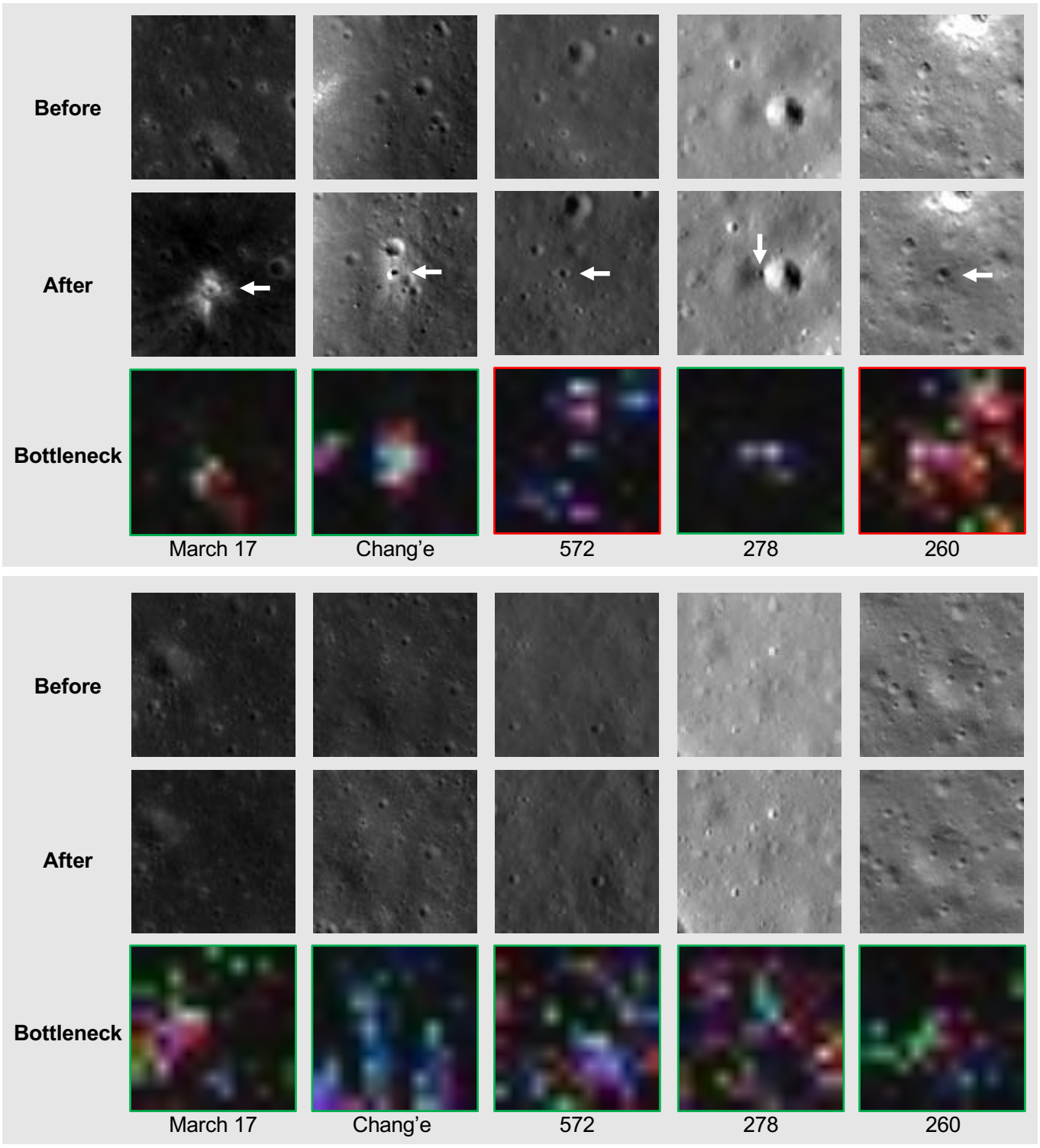

Fig. 11. *Change* (top section) and *no-change* (bottom section) image pairs from the LROC Impacts dataset used for Experiment 3. White arrows point to the new feature in the after images. The third row of each section shows the difference between bottleneck representations of before and after images. Correctly classified images are outlined in green and mis-classified images in red.

row 3) with the bottleneck representations of the other *change* image representations in Figures 11 and 12, the mis-classified representations do not appear as localized to the surface feature change. In the case of image 572, the representation seems to have picked up on the actual new impact crater as well as an impact crater that is present in the before image but not the after image due to mis-registration. The representation for image 260 also appears to have several false detections in addition to the actual new impact crater, perhaps due to features that appear in one image but not in the other due

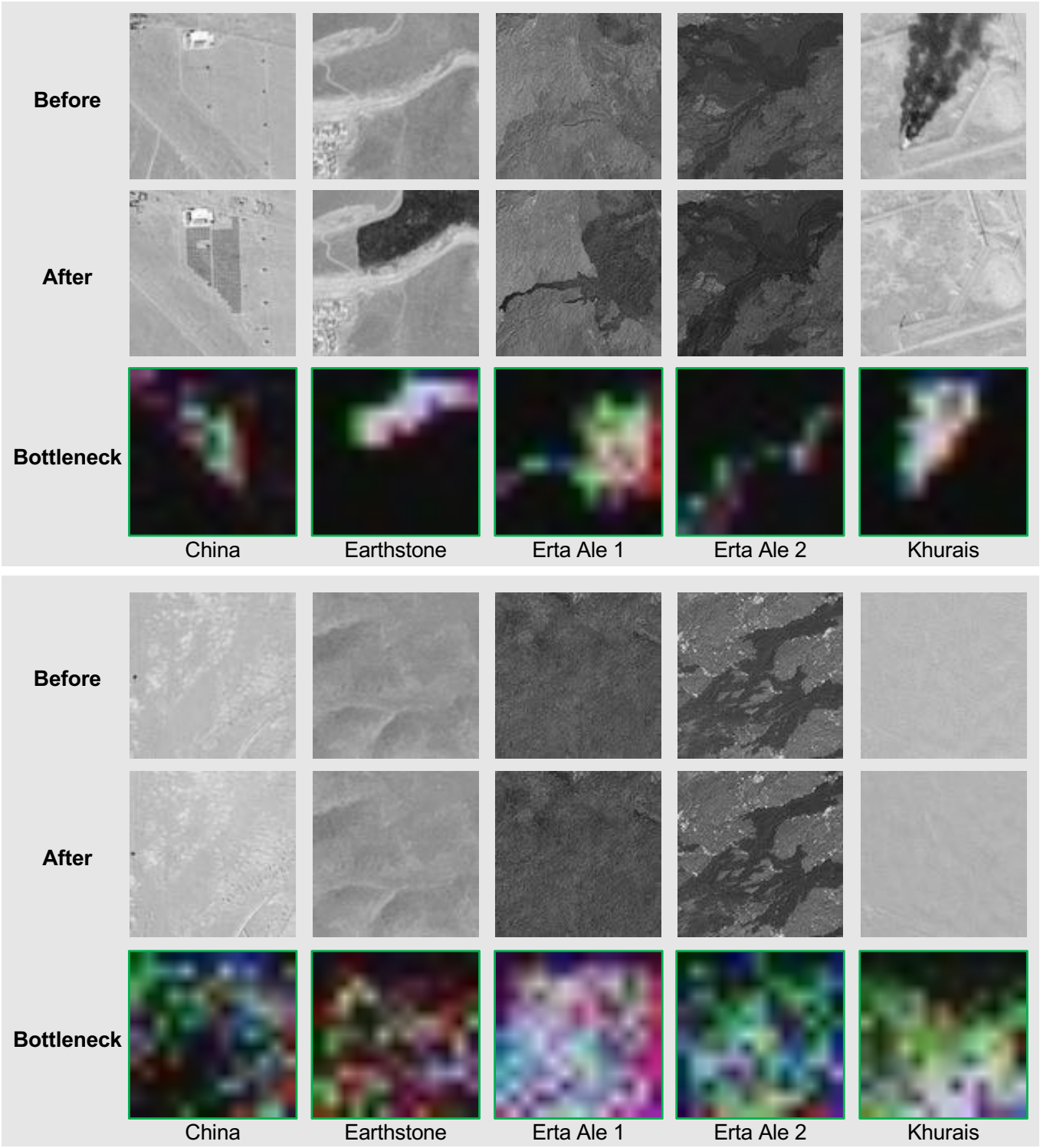

Fig. 12. *Change* (top section) and *no-change* (bottom section) image pairs from the Planet Earth dataset used for Experiment 3. The third row of each section shows the difference between bottleneck representations of before and after images. The green outlines indicate that all images were correctly classified.

to mis-registration. This suggests a limit to the level of mis-registration this approach can successfully tolerate.

Comparing the bottleneck representations in Figure 12 (row 3) to the corresponding before and after images in Figure 12 (rows 1-2), we see that despite never having seen images of Earth or the types of features in the Planet dataset during training, the autoencoder is still able to encode useful representations of the surface features for change detection. This result further supports our hypothesis that the autoencoder bottleneck representations are the most general of those we studied and might be the most suitable for general purpose change detection for surface features on planetary bodies.

Dataset	Mean/Std of n (Percent of Total Pixels)
HiRISE RSL train	$0.028\% \pm 0.013$
HiRISE RSL test (Exp. 1)	$0.032\% \pm 0.019$
CTX Impacts test (Exp. 2)	$0.079\% \pm 0.020$
LROC Impacts test (Exp. 3)	$0.026\% \pm 0.007$
Planet Earth test (Exp. 3)	$0.035\% \pm 0.019$

TABLE VII
MEAN AND STANDARD DEVIATION OF THE NUMBER OF DIFFERENCE PIXELS $n > 2\sigma$ (NAIVE BAYES INPUT) AS A FRACTION OF THE TOTAL (100×100) PIXELS FOR EACH DATASET USED FOR TRAINING OR TESTING.

These representations can also enable a more interpretable approach than the other compared methods. The representations in Figure 12 suggest that the difference between the encodings of before and after images corresponds to the regions where features changed within the image. In future work, we will explore the use of these encodings to classify change at the region or pixel level given only image-level labels.

D. Baseline Performance

We found that the baseline method exhibited good performance for the HiRISE RSL, CTX Impacts, and Planet Earth test examples, but not for the LROC Impacts examples. In these three datasets, the surface feature exhibiting change occupies a large portion of the 100×100 image (see Figure 4 and Figure 12). The scale of the impacts in the LROC Impacts dataset is much smaller compared to the features in the other three datasets. The baseline method depends directly on the number n of differenced pixels in an image pair that are beyond two standard deviations from the mean difference in pixels. If n represented only the pixels where surface feature changes occurred, we would expect this number to be larger for changes in *larger* features (as in the HiRISE RSL, CTX Impacts, and Planet Earth images) and smaller for changes in *smaller* features (as in the LROC Impacts images). In Table VII, we show the mean and standard deviation of n as a percentage of the total pixels in each image $(100 \times 100 = 10,000)$ for all datasets used for training or testing. Given the distributions of n in the test datasets shown in this table, we would expect Naive Bayes to perform well for the LROC Impacts dataset, since its distribution of n is similar to the training set. Further, we would expect Naive Bayes to have lower performance on the CTX Impacts dataset, since its distribution of n is the most different from the training set. Yet, this was not the case. This implies that pixels where a surface feature change occurred are not always represented by pixels with the highest difference in intensity, and that difference-based approaches to change detection will not reliably detect changes in such examples.

Overall, we were surprised that the Siamese network using grayscale image pairs did not perform as well as other methods. It is possible that this method would achieve better performance using a distance measure other than L1 distance or even a learned similarity metric as in Chopra *et al.* [38]. While it has been shown that features learned by deep neural networks via training on one dataset are often transferable to very different datasets and tasks [33], [34], we would not expect this to be true for the bottleneck representations which are very different from natural images (e.g., last column of Figure 4). Thus, we did not find the the low performance of the Siamese network (which used Xception without fine-tuning for feature extraction) using bottleneck representations surprising. This is a direction for future work. Towards the goal of general purpose change detection for planetary surface features, another direction for future work is to create a more diverse dataset of Martian surface feature changes (including, *e.g.*, dust devils, ice cover, and slope streaks) for both training and testing the proposed change detection approaches.

VII. CHANGE MAP VISUALIZATION

We trained our classifiers on 100×100-pixel tiles sampled from full-resolution images. In practice, temporal image pairs that scientists wish to assess for surface feature change might be in tile form, *e.g.*, if they are intermediate products in a map tile server. Most often, image pairs will be two map-projected, co-registered images that are much larger than the tile size. In this case, it is useful to produce a change map across the image pairs using our change detection classifier. To produce change maps, we convolved the classifiers over the full-resolution image pair, stored the prediction made for each tile, and averaged the predictions that were computed over each pixel to produce a likelihood estimate for each pixel. The stride size controls the resolution of the change map since pixels will be visited more frequently (and thus more predictions will be averaged for each pixel) for smaller stride sizes. Figure 13 shows the change map for a region in the southern wall of Garni crater from the HiRISE RSL dataset in which RSL that were present in the before image have faded away in the after image (before: ESP_028501_1685, after: ESP_029213_1685). We used the Absolute Difference representation and fine-tuned Inception-V3 classifier since this was the best performing approach for this dataset and a stride size of 2 pixels.

VIII. CONCLUSION

We present new change detection approaches for surface features in remote sensing images. Our deep learning approaches leverage transfer learning and image embeddings to identify changes in surface features such as recurring slope lineae (RSL), meteorite impact craters, and human-made structures using a relatively small number of labeled training examples. Our experiments showed that our change detection methods outperformed the difference-based baseline method with equal pre-processing for the HiRISE RSL dataset, and regardless of pre-processing on all other datasets. These experiments revealed a key insight that changed features may not always be represented by pixels with the highest difference in intensity between before and after images, which limits their generalization ability. We showed that latent ("bottleneck") representations learned by a convolutional autoencoder are the most general representation for surface feature change detection in our study, and that Inception-V3 fine-tuned with bottleneck representations could detect surface feature changes

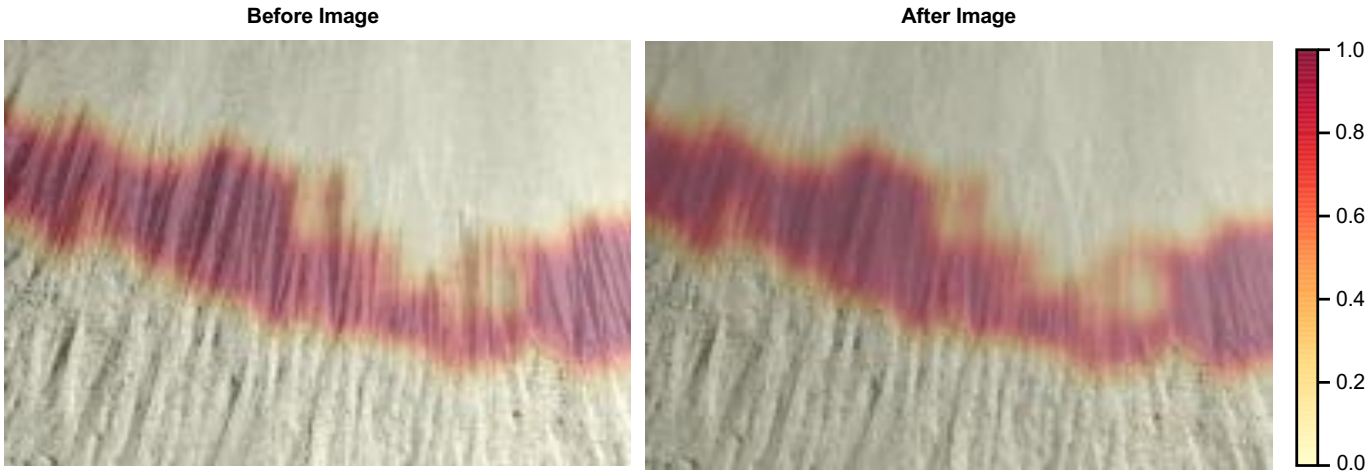

Fig. 13. Change map visualization for example HiRISE RSL image showing fading RSL, computed by convolving the Inception-V3 with absolute difference representations change classifier over image. Scale bar represents the average likelihood of change estimated in each pixel.

even when the feature type, imaging sensor, level of misregistration, feature scale, and planetary body are different than in the training dataset.

Future Work. In this study, our experimental approach was to isolate our change detection datasets by feature type, instrument, planet, and other properties in order to reveal the strengths and weaknesses of each approach. The success of our change detection approaches, in particular the fine-tuned neural network using autoencoder bottleneck representations, on examples that deviated significantly from the training set suggests that variants of these methods hold promise for general-purpose change detection for surface features on planetary bodies that share similar background characteristics. We plan to explore this hypothesis in future work. In this work, our goal was to predict image-level labels of *change* or *no-change*. In future work, we will investigate how to leverage bottleneck representations or class activation maps to predict pixel-level or region-level labels of *change* or *no-change* given only image-level labels.

ACKNOWLEDGMENTS

The authors would like to thank Dr. David Stillman of the Southwest Research Institute (SwRI) and Dr. Ingrid Daubar of the Jet Propulsion Laboratory for their expert knowledge and the Planetary Data System for supporting the development of this work, as well as Dr. Gary Doran of the Jet Propulsion Laboratory for his assistance with processing CTX image pairs. This research was carried out (in part) at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration and funded through the Internal Strategic University Research Partnerships (SURP) program.

APPENDIX A
ADDITIONAL DATASET DETAILS

In Table I, we provided identifiers of images used for the HiRISE RSL Dataset and indicated how each image was used in our experiments. Three more datasets—CTX Meteorite Impacts, LROC Impacts, and Planet Earth—were used for testing generalization of the change detection classifiers in Section V. Table VIII gives the identifiers for the images used in these datasets. We provide instructions for accessing theses images in the following subsections.

Instrument	Image ID	Date Acquired
CTX	P02_001790_1871_XN_07N182W	12/13/2006
CTX	P03_002169_1937_XI_13N091W	01/12/2007
CTX	P06_003451_2035_XN_23N171W	04/22/2007
CTX	P09_004477_1906_XN_10N100W	07/11/2007
CTX	P13_005954_1927_XI_12N105W	11/03/2007
CTX	P13_006178_1907_XN_10N100W	11/20/2007
CTX	P13_006286_2073_XN_27N171W	11/29/2007
CTX	P14_006560_1936_XN_13N091W	12/20/2007
CTX	B01_009923_1790_XN_01S113W	09/07/2008
CTX	B01_010213_1790_XN_01S113W	09/30/2008
CTX	B02_010424_1849_XI_04N113W	10/16/2008
CTX	B06_012006_1912_XI_11N104W	02/17/2009
CTX	B06_012022_1845_XI_04N182W	02/18/2009
CTX	B09_012981_1844_XI_04N084W	05/03/2009
CTX	G14_023886_1819_XN_01N083W	08/31/2011
CTX	G17_024942_1813_XI_01N112W	11/22/2011
LROC	M104318871L (572)	08/07/2009
LROC	M142531289R (260)	10/24/2010
LROC	M161489808R (278)	05/31/2011
LROC	M176811566R (278)	11/24/2011
LROC	M181330922L (March 17)	01/16/2012
LROC	M1111614190R (572)	12/31/2012
LROC	M1119134763 (260)	03/28/2013
LROC	M1129602407L (Chang'e)	07/27/2013
LROC	M1139065512L (March 17)	11/14/2013
LROC	M1144922100L (Chang'e)	01/21/2014
PlanetScope	china-solar-20151217-before	12/17/2015
PlanetScope	khurais-20151218-after	12/18/2015
PlanetScope	khurais-20160212-before	02/12/2016
PlanetScope	china-solar-20160227-after	02/27/2016
PlanetScope	ertaale-20170116-before	01/16/2017
PlanetScope	ertaale-20170123-after	01/23/2017
PlanetScope	earthstone-20170702-before	07/02/2017
PlanetScope	earthstone-20170704-after	07/04/2017

TABLE VIII
IMAGE PRODUCTS USED FOR TEST DATASETS.

a) HiRISE RSL Dataset: The HiRISE images we used for this study can be accessed at https://www.uahirise.org/

⟨ImageID⟩, where ⟨ImageID⟩ is the Image ID from Table I. We used the JP2 black and white (red channel) map-projected products. To crop the images to the $10,000 \times 10,000$-pixel region of Garni crater, we used the following ImageMagick command: convert image.jp2 -crop 10000x10000+3300+12000 cropped_image.jp2.

b) CTX Meteorite Impacts Dataset: The CTX images we used for this dataset can be found using the search tool on the Planetary Data System (PDS) Imaging Node: https://pds-imaging.jpl.nasa.gov/search. On the filter menu, select "mars reconnaissance orbiter" under Mission, "ctx" under Instrument, and type the Image ID from Table VIII in the Search bar.

c) LROC Impacts Dataset: The LROC images used for this study can be found on the Arizona State University School of Earth and Space Exploration's LROC portal at the following URLs:
http://lroc.sese.asu.edu/featured_sites/lroc_features/17%20March%202013%20Event/feature_highlights/481
http://lroc.sese.asu.edu/featured_sites/lroc_features/17%20March%202013%20Event/feature_highlights/490
http://lroc.sese.asu.edu/featured_sites/lroc_features/New%20Crater%20572/feature_highlights/504
http://lroc.sese.asu.edu/featured_sites/lroc_features/New%20Crater%20572/feature_highlights/509
http://lroc.sese.asu.edu/featured_sites/lroc_features/New%20Crater%20260/feature_highlights/491
http://lroc.sese.asu.edu/featured_sites/lroc_features/New%20Crater%20260/feature_highlights/492
http://lroc.sese.asu.edu/featured_sites/lroc_features/New%20Crater%20278/feature_highlights/496
http://lroc.sese.asu.edu/featured_sites/lroc_features/New%20Crater%20278/feature_highlights/497
http://lroc.sese.asu.edu/featured_sites/lroc_features/Chang'e%203%20Landing%20Site/feature_highlights/553
http://lroc.sese.asu.edu/featured_sites/lroc_features/Chang'e%203%20Landing%20Site/feature_highlights/558

d) Planet Earth Dataset: The images used for this dataset are browse products from the Gallery on the Planet website. They can be accessed at the following URLs:
https://www.planet.com/gallery/china-solar-20170130/
https://www.planet.com/gallery/khurais/
https://www.planet.com/gallery/ertaale-20170123/
https://www.planet.com/gallery/earthstone-fire-20170705/

APPENDIX B
DETAILS OF EXPERIMENTAL SETUP

To help reproduce the experiments in this paper, we provide details for implementing and training all models in the following subsections.

A. Fine-tuned Inception-V3

We implemented the Inception-V3 approaches using TensorFlow. Details of how to fine-tune Inception-V3 for new categories can be found on the TensorFlow website[5]. We used Stochastic Gradient Descent optimization and the Sparse Softmax Cross-Entropy loss function. We used a training batch size of 100 and learning rate of 0.001. We trained a different model for each image representation until validation loss was minimized, which resulted in different training times for each representation. We fine-tuned the absolute difference model for 2,220 steps (∼48 epochs), the autoencoder bottleneck model for 3,160 steps (∼69 epochs), the composite grayscale model for 790 steps (∼17 epochs), and the signed difference model for 3,700 steps (∼81 epochs).

B. Siamese Network

We implemented the Siamese network approaches using Keras. Details of how to load or fine-tune a pre-trained model in Keras can be found on the Keras website[6] (we used Xception for feature extraction). We used a batch size of 100 and learning rate of 0.001. We used the Adam optimizer with hyperparameters β_1 0.9, β_1 0.999, ϵ 1e-7, and decay 0.001/4551 (learning rate divided by number of training images) [50]. Validation loss was minimized after training for 19 epochs for the grayscale representations and 200 epochs for the bottleneck representations.

C. Convolutional Autoencoder

We implemented the convolutional autoencoder using Keras. The size (and thus number of feature maps) is given in Table IV. We used 3×3-pixel kernels for convolution and 2×2-pixel kernels for max pooling, both using a stride size of 1 pixel. We used the Adam optimizer with β_1 0.9, β_1 0.999, ϵ 1e-7, and decay 0.0. We used a batch size of 100 and learning rate of 0.001. We used the binary cross-entropy loss function. We trained the autoencoder for 50 epochs.

APPENDIX C
ADDITIONAL CHANGE MAPS

In Section VII, we showed an example change map for a region of Garni crater computed by convolving the Inception-V3 model fine-tuned with absolute difference image representations over the entire image with a stride size of 2 pixels (Figure 13). In Figures 14-18, we show the change maps computed by convolving additional approaches tested over the same image pair as Figure 13.

REFERENCES

[1] A. S. McEwen, E. M. Eliason, J. W. Bergstrom, N. T. Bridges, C. J. Hansen, W. A. Delamere, J. A. Grant, V. C. Gulick, K. E. Herkenhoff, L. Keszthelyi, R. L. Kirk, M. T. Mellon, S. W. Squyres, N. Thomas, and C. M. Weitz, "Mars Reconnaissance Orbiter's High Resolution Imaging Science Experiment (HiRISE)," *Journal of Geophysical Research*, vol. 112, no. E5, p. E05S02, May 2007.

[2] M. C. Malin, J. F. Bell, B. A. Cantor, M. A. Caplinger, W. M. Calvin, R. T. Clancy, K. S. Edgett, L. Edwards, R. M. Haberle, P. B. James, S. W. Lee, M. A. Ravine, P. C. Thomas, and M. J. Wolff, "Context Camera Investigation On Board the Mars Reconnaissance Orbiter," *Journal of Geophysical Research*, vol. 112, no. E5, p. E05S04, May 2007.

[3] D. E. Stillman, T. I. Michaels, R. E. Grimm, and K. P. Harrison, "New Observations of Martian Southern Mid-Latitude Recurring Slope Lineae (RSL) Imply Formation by Freshwater Subsurface Flows," *Icarus*, vol. 233, pp. 328–341, May 2014.

[5] https://www.tensorflow.org/hub/tutorials/image_retraining

[6] https://keras.io/applications

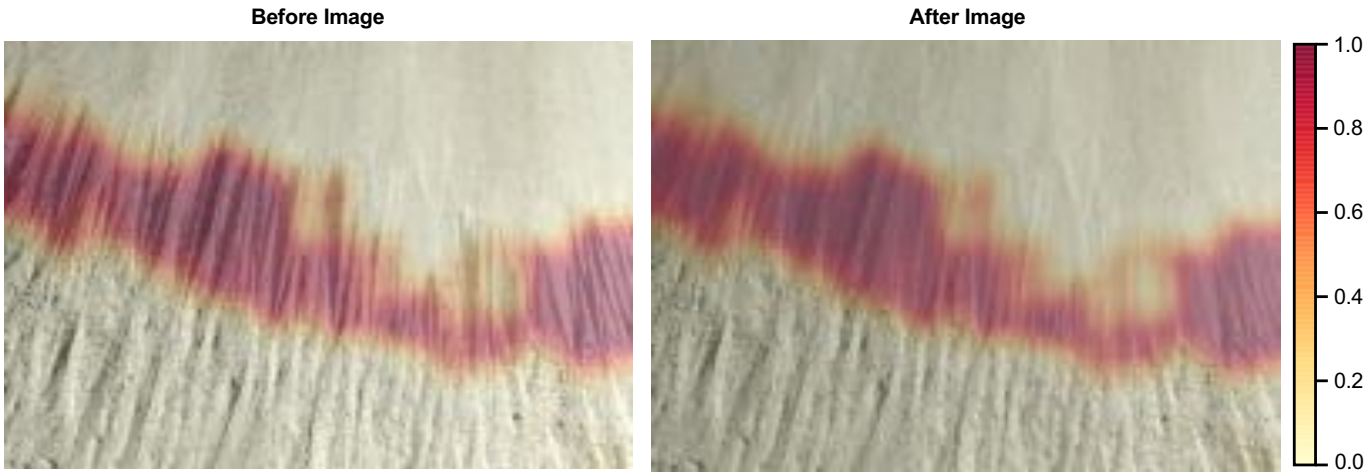

Fig. 14. Change map visualization for example HiRISE RSL image pair showing fading RSL, computed by convolving the **Inception-V3** with **signed difference** representations change classifier over the image pair. Scale bar represents the average likelihood of change estimated in each pixel.

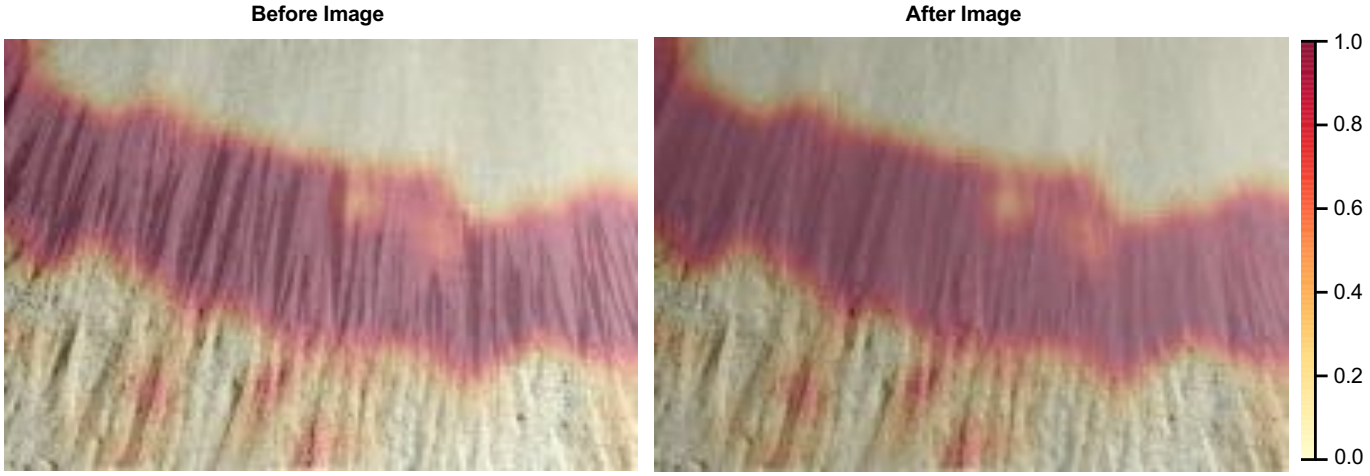

Fig. 15. Change map visualization for example HiRISE RSL image pair showing fading RSL, computed by convolving the **Inception-V3** with **composite grayscale** representations change classifier over the image pair. Scale bar represents the average likelihood of change estimated in each pixel.

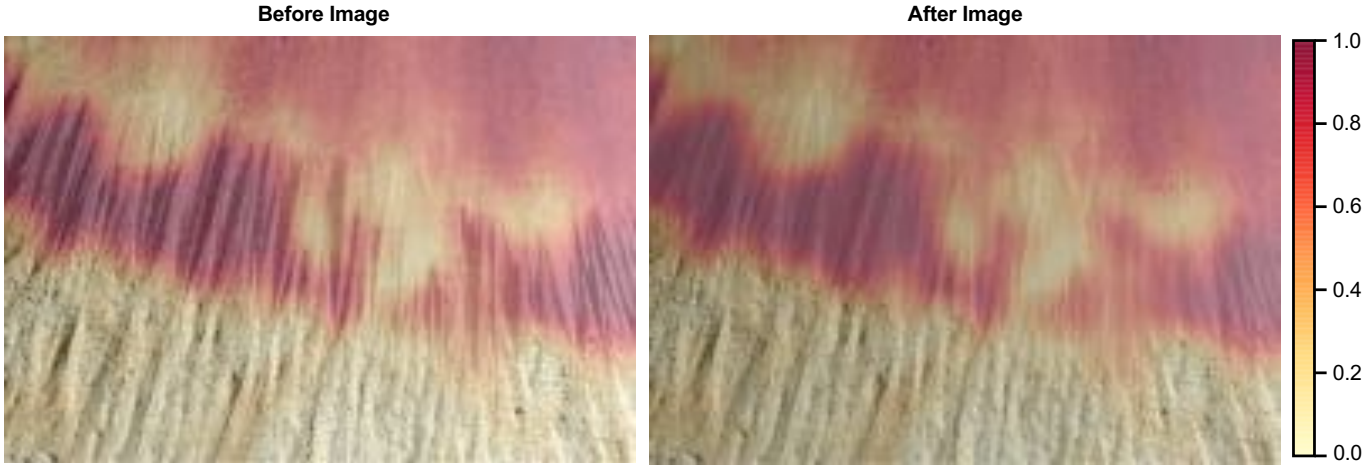

Fig. 16. Change map visualization for example HiRISE RSL image pair showing fading RSL, computed by convolving the **Inception-V3** with **autoencoder bottleneck** representations change classifier over the image pair. Scale bar represents the average likelihood of change estimated in each pixel.

[4] I. Daubar, A. McEwen, S. Byrne, M. Kennedy, and B. Ivanov, "The Current Martian Cratering Rate," *Icarus*, vol. 225, no. 1, pp. 506–516, Jul. 2013.

[5] E. J. Speyerer, R. Z. Povilaitis, M. S. Robinson, P. C. Thomas, and

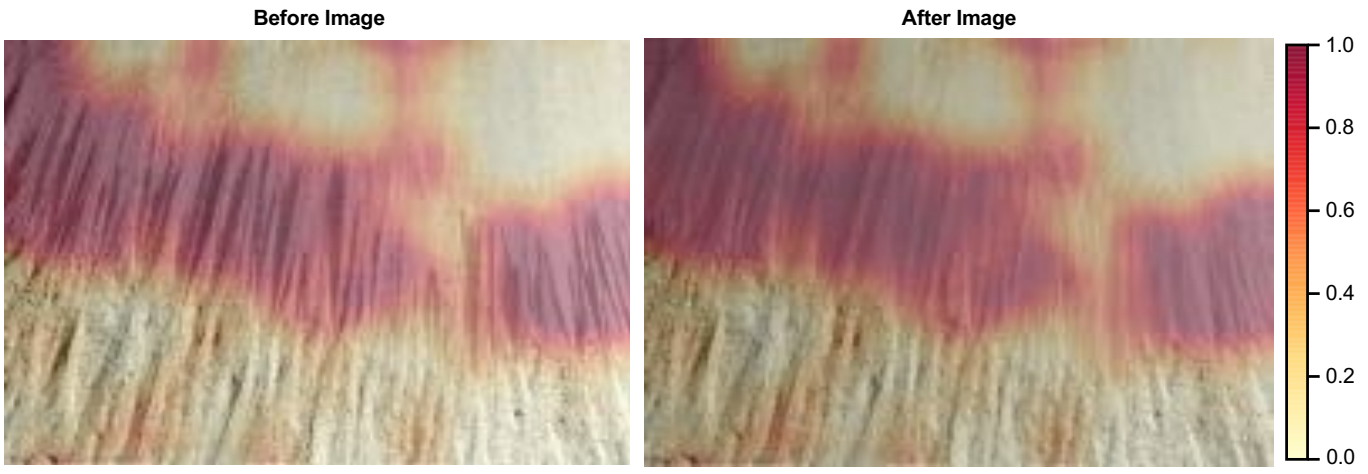

Fig. 17. Change map visualization for example HiRISE RSL image pair showing fading RSL, computed by convolving the **Siamese network** with **grayscale** representations change classifier over the image pair. Scale bar represents the average likelihood of change estimated in each pixel.

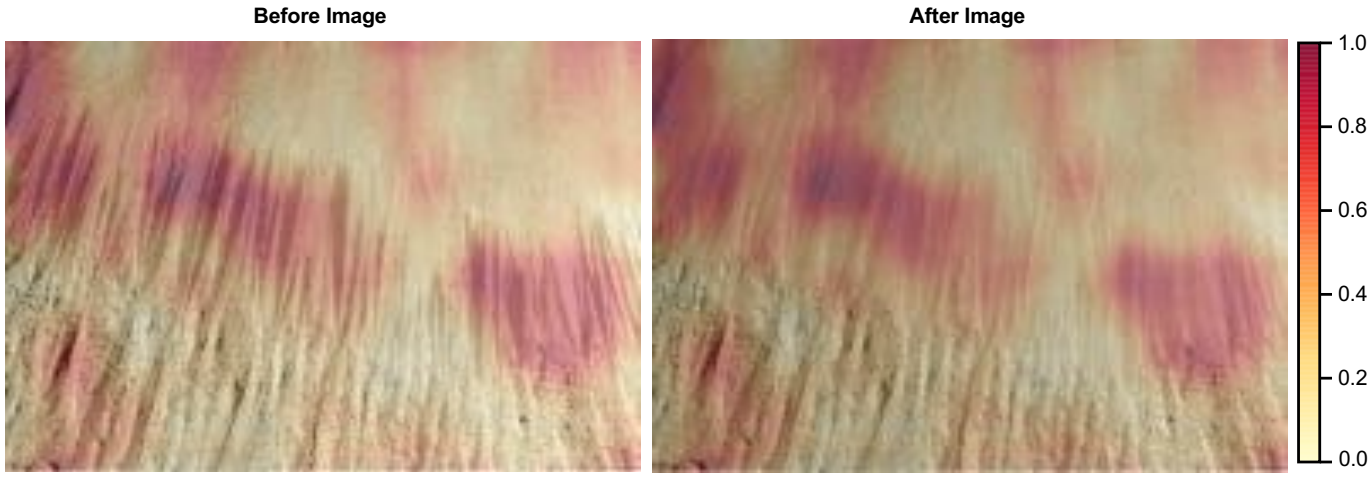

Fig. 18. Change map visualization for example HiRISE RSL image pair showing fading RSL, computed by convolving the **Siamese network** with **autoencoder bottleneck** representations change classifier over the image pair. Scale bar represents the average likelihood of change estimated in each pixel.

R. V. Wagner, "Quantifying Crater Production and Regolith Overturn on the Moon with Temporal Imaging," *Nature*, vol. 538, no. 7624, pp. 215–218, Oct. 2016.

[6] C. Munyati, "Wetland Change Detection on the Kafue Flats, Zambia, by Classification of a Multitemporal Remote Sensing Image Dataset," *International Journal of Remote Sensing*, vol. 21, no. 9, pp. 1787–1806, Jan. 2000.

[7] J. Nichol and M. S. Wong, "Satellite Remote Sensing for Detailed Landslide Inventories using Change Detection and Image Fusion," *International Journal of Remote Sensing*, vol. 26, no. 9, pp. 1913–1926, May 2005.

[8] A. Shalaby and R. Tateishi, "Remote Sensing and GIS for Mapping and Monitoring Land Cover and Land-use Changes in the Northwestern Coastal Zone of Egypt," *Applied Geography*, vol. 27, no. 1, pp. 28–41, Jan. 2007.

[9] F. Yuan, K. E. Sawaya, B. C. Loeffelholz, and M. E. Bauer, "Land Cover Classification and Change Analysis of the Twin Cities (Minnesota) Metropolitan Area by Multitemporal Landsat Remote Sensing," *Remote Sensing of Environment*, vol. 98, no. 2-3, pp. 317–328, Oct. 2005.

[10] X. Dai and S. Khorram, "The Effects of Image Misregistration on the Accuracy of Remotely Sensed Change Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 36, no. 5, pp. 1566–1577, 1998.

[11] A. P. Tewkesbury, A. J. Comber, N. J. Tate, A. Lamb, and P. F. Fisher, "A Critical Synthesis of Remotely Sensed Optical Image Change Detection Techniques," *Remote Sensing of Environment*, vol. 160, pp. 1–14, Apr. 2015.

[12] T. Hame, I. Heiler, and J. San Miguel-Ayanz, "An Unsupervised Change Detection and Recognition System for Forestry," *International Journal of Remote Sensing*, vol. 19, no. 6, pp. 1079–1099, Jan. 1998.

[13] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, "PCA-Based Land-Use Change Detection and Analysis using Multitemporal and Multisensor Satellite Data," *International Journal of Remote Sensing*, vol. 29, no. 16, pp. 4823–4838, Aug. 2008.

[14] M. A. Torres-Vera, R. M. Prol-Ledesma, and D. Garcia-Lopez, "Three Decades of Land Use Variations in Mexico City," *International Journal of Remote Sensing*, vol. 30, no. 1, pp. 117–138, Jan. 2009.

[15] O. Abd El-Kawy, J. Rød, H. Ismail, and A. Suliman, "Land Use and Land Cover Change Detection in the Western Nile Delta of Egypt using Remote Sensing Data," *Applied Geography*, vol. 31, no. 2, pp. 483–494, Apr. 2011.

[16] R. Peiman, "Pre-Classification and Post-Classification Change-Detection Techniques to Monitor Land-Cover and Land-Use Change using Multi-Temporal Landsat Imagery: A Case Study on Pisa Province in Italy," *International Journal of Remote Sensing*, vol. 32, no. 15, pp. 4365–4381, Aug. 2011.

[17] A. A. Alesheikh, A. Ghorbanali, and N. Nouri, "Coastline Change Detection using Remote Sensing," *International Journal of Environmental Science & Technology*, vol. 4, no. 1, pp. 61–66, Dec. 2007.

[18] A. Singh, "Review Article: Digital Change Detection Techniques using Remotely-Sensed Data," *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 989–1003, Jun. 1989.

[19] G. Castilla and G. J. Hay, "Image Objects and Geographic Objects," in *Object-Based Image Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 91–110.

[20] L. Bruzzone and D. Prieto, "An Adaptive Semiparametric and Context-Based Approach to Unsupervised Change Detection in Multitemporal Remote-Sensing Images," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 452–466, Apr. 2002.

[21] V. Walter, "Object-Based Classification of Remote Sensing Data for Change Detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no. 3-4, pp. 225–238, Jan. 2004.

[22] J. Im and J. R. Jensen, "A Change Detection Model Based on Neighborhood Correlation Image Analysis and Decision Tree Classification," *Remote Sensing of Environment*, vol. 99, no. 3, pp. 326–340, Nov. 2005.

[23] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Rojo-Alvarez, and M. Martinez-Ramon, "Kernel-Based Framework for Multitemporal and Multisource Remote Sensing Data Classification and Change Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 6, pp. 1822–1835, Jun. 2008.

[24] M. N. Klaric, B. C. Claywell, G. J. Scott, N. J. Hudson, O. Sjahputera, Yonghong Li, S. T. Barratt, J. M. Keller, and C. H. Davis, "GeoCDX: An Automated Change Detection and Exploitation System for High-Resolution Satellite Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 2067–2086, Apr. 2013.

[25] M. Volpi, D. Tuia, F. Bovolo, M. Kanevski, and L. Bruzzone, "Supervised Change Detection in VHR Images using Contextual Information and Support Vector Machines," *International Journal of Applied Earth Observation and Geoinformation*, vol. 20, pp. 77–85, Feb. 2013.

[26] P. Gray, J. Ridge, S. Poulin, A. Seymour, A. Schwantes, J. Swenson, D. Johnston, P. C. Gray, J. T. Ridge, S. K. Poulin, A. C. Seymour, A. M. Schwantes, J. J. Swenson, and D. W. Johnston, "Integrating Drone Imagery into High Resolution Satellite Remote Sensing Assessments of Estuarine Environments," *Remote Sensing*, vol. 10, no. 8, p. 1257, Aug. 2018.

[27] D. E. Stillman, "Unraveling the Mysteries of Recurring Slope Lineae," in *Dynamic Mars: Recent and Current Landscape Evolution of the Red Planet*, R. J. Soare, S. J. Conway, and S. M. Clifford, Eds. Elsevier, 2018, ch. 2, pp. 51–85.

[28] G. Doran, D. R. Thompson, and T. Estlin, "Precision Instrument Targeting via Image Registration for the Mars 2020 Rover," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016, pp. 3352–3358.

[29] M. S. Robinson, S. M. Brylow, M. Tschimmel, D. Humm, S. J. Lawrence, P. C. Thomas, B. W. Denevi, E. Bowman-Cisneros, J. Zerr, M. A. Ravine, M. A. Caplinger, F. T. Ghaemi, J. A. Schaffner, M. C. Malin, P. Mahanti, A. Bartels, J. Anderson, T. N. Tran, E. M. Eliason, A. S. McEwen, E. Turtle, B. L. Jolliff, and H. Hiesinger, "Lunar Reconnaissance Orbiter Camera (LROC) Instrument Overview," *Space Science Reviews*, vol. 150, no. 1-4, pp. 81–124, Jan. 2010. [Online]. Available: http://link.springer.com/10.1007/s11214-010-9634-2

[30] W.-H. Ip, J. Yan, C.-L. Li, and Z.-Y. Ouyang, "Preface: The Chang'e-3 lander and rover mission to the Moon," *Research in Astronomy and Astrophysics*, vol. 14, no. 12, pp. 1511–1513, Dec. 2014. [Online]. Available: http://stacks.iop.org/1674-4527/14/i=12/a=001?key=crossref.9bfd149bf7dda6d9668c908849b8f2f1

[31] Planet Team, "Planet Application Program Interface: In Space for Life on Earth," 2018. [Online]. Available: https://api.planet.com/

[32] R. Houborg and M. F. McCabe, "High-Resolution NDVI from Planet's Constellation of Earth Observing Nano-Satellites: A New Data Source for Precision Agriculture," *Remote Sensing*, vol. 8, no. 9, p. 768, 2016.

[33] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, Jun. 2014, pp. 512–519.

[34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How Transferable are Features in Deep Neural Networks?" in *Advances in Neural Information Processing Systems*. MIT Press, 2014, pp. 3320–3328.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2016, pp. 2818–2826.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, Jun. 2009, pp. 248–255.

[37] J. Bromley, I. Guyon, Y. Lecun, E. Sickinger, and R. Shah, "Signature Verification using a "Siamese" Time Delay Neural Network," in *Advances in Neural Information Processing Systems (NIPS)*, J. Cowan and G. Tesauro, Eds., 1993, pp. 737–744.

[38] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[39] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[40] G. Koch, "Siamese Neural Networks for One-Shot Image Recognition," Ph.D. dissertation, University of Toronto, 2015.

[41] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *2017 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017, pp. 1251–1258.

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017, pp. 4278–4284.

[43] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *eprint arXiv:1409.1556*, Sep. 2014.

[44] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction," in *International Conference on Artificial Neural Networks (ICANN): Artificial Networks and Machine Learning*. Springer, Berlin, Heidelberg, 2011, pp. 52–59.

[45] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science (New York, N.Y.)*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[46] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. JMLR.org, 2015, pp. 448–456.

[47] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. Morgan Kaufmann, 1994, pp. 474–485.

[48] H. Zhang, "The Optimality of Naive Bayes," in *Florida Artificial Intelligence Research Society (FLAIRS) Conference*, 2004.

[49] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A System for Large-Scale Machine Learning," 2016.

[50] D. P. Kingma and J. L. Ba, "Adam: a Method for Stochastic Optimization," *International Conference on Learning Representations 2015*, pp. 1–15, Dec. 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

Hannah R. Kerner received the B.S. in computer science from the University of North Carolina at Chapel Hill in 2014. She is currently pursuing the Ph.D. degree from Arizona State University.

Her doctoral research focuses on machine learning solutions to challenges in remote sensing, especially for Mars exploration missions. She has worked at NASA's Jet Propulsion Laboratory, Goddard Space Flight Center, and Langley Research Center as well as commercial remote sensing company Planet, Inc.

Kiri L. Wagstaff received the B.S. degree in computer science from the University of Utah in 1997, the M.S. and Ph.D. degrees in computer science from Cornell University in 2000 and 2002 respectively, the M.S. degree in geological sciences from the University of Southern California in 2008, and the MLIS degree in library and information science from San Jose State University in 2017.

She is a Principal Research Technologist in artificial intelligence and machine learning at the Jet Propulsion Laboratory. Her research focuses on developing new machine learning and data analysis methods for spacecraft.

Brian Bue

Patrick C. Gray received the B.A. in computer science from the University of North Carolina at Chapel Hill in 2014. He is currently pursuing the Ph.D. degree from Duke University.

His doctoral research in the Duke Marine Robotics and Remote Sensing Lab focuses on machine learning approaches to remote sensing analysis, coordinating satellite, drone, and in-situ environmental monitoring, and understanding how autonomous systems will benefit field scientists—all with a focus on coastal and polar environments.

James F. Bell

Heni Ben Amor Heni Ben Amor received the Dipl.-Inf. degree in computer science from the University of Koblenz-Landau in 2005 and the Ph.D. in robotics and computer science from the Technical University Bergakademie Freiberg in 2010.

He is an Assistant Professor for robotics at Arizona State University where he directs the ASU Interactive Robotics Laboratory. Prior to joining ASU, he was a research scientist at Georgia Tech, a postdoctoral researcher at the Technical University Darmstadt (Germany), and a visiting research scientist in the Intelligent Robotics Lab at the University of Osaka (Japan). His primary research interests lie in the fields of artificial intelligence, machine learning, robotics, and human-robot interaction. He received the NSF CAREER Award in 2018, the Fulton Outstanding Assistant Professor Award in 2018, and the Daimler-and-Benz Fellowship in 2012.