Forest Change Detection in Incomplete Satellite Images with Deep Neural Networks

Salman H Khan, Xuming He, Fatih Porikli, and Mohammed Bennamoun

***Abstract*—Land cover change monitoring is an important task from the perspective of regional resource monitoring, disaster management, land development and environmental planning. In this study, we analyze imagery data from remote sensing satellites to detect forest cover changes over a period of 29 years ($1987-2015$). Since the original data is severely incomplete and contaminated with artifacts, we first devise a spatiotemporal inpainting mechanism to recover the missing surface reflectance information. The spatial filling process makes use of the available data of the nearby temporal instances followed by a sparse encoding based reconstruction. We formulate the change detection task as a region classification problem. We build a multi-resolution profile of the target area and generate a candidate set of bounding box proposals that enclose potential change regions. In contrast to existing methods that use handcrafted features, we automatically learn region representations using a deep neural network in a data-driven fashion. Based on these highly discriminative representations, we determine forest changes and predict their onset and offset timings by labeling the candidate set of proposals. Our approach achieves state-of-the-art average patch classification rate of 91.6% (an improvement of $\sim 16\%$) and mean onset/offset prediction error of 4.9 months (an error reduction of 5.0 months) compared to a strong baseline. We also qualitatively analyze the detected changes in the unlabeled image regions, which demonstrate that the proposed forest change detection approach is scalable to new regions.**

***Index Terms*—Change detection, Multi-temporal spectral data, Remote sensing, Deep learning, Image inpainting.**

I. Introduction

Ecosystem management and socioeconomic studies at regional, national and international scale require the detection and monitoring of land cover changes. In particular, forest change detection is crucial for continuous environmental monitoring to closely investigate pressing environmental issues such as natural resource depletion, biodiversity loss and deforestation. Change detection can also provide essential information to help in disaster management, policy making, area planning and efficient land management. In Australia alone, forests occupy 125 million hectares, which corresponds to 16% of the total continent's land and nearly 3% of the total forest area in the world. Forests are regularly disturbed by significant changes, e.g., during 2006-07 to 2010-11, an area of approximately 39 million hectares was destroyed by fires and 9 thousand hectares were yearly harvested in Australia [1]. These disturbances need to be frequently monitored and analyzed to develop competent response procedures for forest ecosystems.

Current studies using medium spatial resolution satellite imagery usually perform a synoptic analysis over a temporal scale of one or more years [2–5]. The traditionally used Landsat imagery based systems work at a longer time scale due to their low coverage across the globe and low repeat frequency in contrast to the coarse spatial resolution satellite imagery sources e.g., Moderate-Resolution Imaging Spectrometer (MODIS), National Oceanic and Atmospheric Administration (NOAA), and Advanced Very High Resolution Radiometer (AVHRR). Moreover, climate and weather conditions (e.g., continuous cloud cover) significantly restrict the acquisition of quality land cover data.

In this work, we introduce an automatic solution for forest monitoring at a sub-annual level for applications that require a more frequent analysis, including grazing land management, crops safety examinations, and natural hazard analysis. Our solution is also applicable to regions undergoing a rapid forest regeneration (thus requiring a more frequent analysis) to avoid excessive omission error [6]. We utilize the publicly accessible Landsat data and monitor changes at a much finer timescale of 2 months as opposed to several years. The primary challenge, however, is the severe missing data problem in the Landsat imagery due to limited camera aperture, cloud occlusion and sensor artifacts. To address this issue, we propose a two-stage strategy for the fine-grain change detection task (see Fig. 1). In the **first** stage, we take a data-driven approach to fill in the missing spatial data and achieve higher temporal resolution from the available Landsat spectral data sequences (Sec. IV). Our technique is based on image inpainting using sparse encoding. The key idea here is to exploit the temporal and spatial continuity of the underlying events and use statistics of the observed image patches to fill in small gaps (Sec. IV-A, IV-B). The resulting high temporal resolution image sequences enable us to analyze data at a much finer temporal scale.

After obtaining the inpainted time-lapse satellite imagery, we tackle the change detection problem in the **second** stage. We focus on two sub-problems under the scope of change detection. The *first* is the detection of multiple classes and instances of change events in a specified region. The *second* is the estimation of the start and end time of the detected change event. For this purpose, we consider the unconstrained change discovery in a large geographic area by selecting class-independent change event candidate regions, and predicting the likelihood of certain change event types along-with their start

S. H. Khan is with the Data61-CSIRO and the Australian National University (ANU), Canberra ACT 0200, Australia. Email:salman.khan@anu.edu.au

X. He is with the ShanghaiTech University, Pudong Xinqu, Shanghai Shi, China. This work was performed when he was with Data61/ANU. Email:xuming.he@anu.edu.au

F. Porikli is with the Australian National University, Canberra ACT 0200, Australia. Email:fatih.porikli@anu.edu.au

M. Bennamoun is with the University of Western Australia, Crawley WA 6009, Australia. Email:mohammed.bennamoun@uwa.edu.au

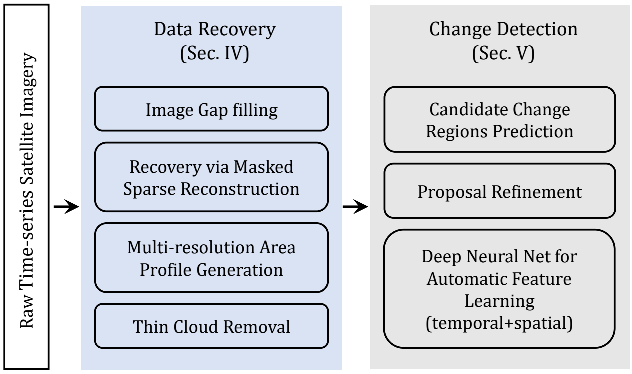

Raw Time-series Satellite Imagery
Data Recovery (Sec. IV)
Image Gap filling
Recovery via Masked Sparse Reconstruction
Multi-resolution Area Profile Generation
Thin Cloud Removal
Change Detection (Sec. V)
Candidate Change Regions Prediction
Proposal Refinement
Deep Neural Net for Automatic Feature Learning (temporal+spatial)

Fig. 1: An overview of our approach for change detection in incomplete satellite images.

and end times. Our main contributions are threefold:

- In contrast to the existing approaches that rely strongly on expert's domain knowledge to extract features, we employ a deep learning approach to automatically capture the most appropriate features from the inpainted image data at the finer temporal scale (Sec. V). Such deep neural network based approaches have shown superior performance in most computer vision tasks such as classification, detection, and segmentation [7–10], and are particularly suitable for representing signals and their spatiotemporal context.
- Unlike the traditional pixel-based local change detection techniques [11–14], our method incorporates contextual information in the form of spatial, spectral and temporal relationships in a novel deep convolutional neural network (CNN) model (Sec. V-C). Thus, our method can be categorized among the object-based change detection methods that are more robust than their pixel-based counterparts [15].
- Conventional object-based change detection methods heavily rely on image segmentation, which often leads to over (excessively large regions) and under (incorrectly too small) partitioning of change areas [4, 16]. To alleviate this problem, we generate change box proposals and select a candidate set with the help of multi-resolution area profiles (Sec.V-A,V-B).

As a case-study, we analyze time-series satellite imagery of the north-east region of Melbourne, Victoria, Australia (Sec. III). Our region-of-interest is a rectangular section with an area of 20,016.1 km^2 (7,728.2 mi^2) lying between a latitude and longitude of $36^0 00' 00.0$"S $146^0 00' 00.0$"E and $38^0 00' 00.0$"S $147^0 00' 00.0$"E. We detect potential change regions in this area and predict their onset and offset timings. Since annotations are available for only a few selected change regions, we perform both a quantitative and qualitative analysis to assess the performance of our approach on both labeled and unlabeled patches, respectively. Through extensive experiments, we show that our approach outperforms all baseline techniques by a significant margin. Our method attains a mean-IOU score of 84.9% and an average recall rate of 77.7% for the temporal change detection and patch-wise classification tasks. In terms of the start and end time predictions for detected change events, our method predicts the onset and offset times with an average error margin of $\sim$3 months and $\sim$6 months, respectively. This performance is remarkably better than the current state-of-the-art approaches, which yield error margins in years scale (Sec. VI-D).

The rest of the paper is organized as follows. We discuss related literature in the next section (Sec. II). Data description is provided in Sec. III and data recovery approach is detailed in Sec. IV. Next, we explain our change detection approach in Sec. V. The experimental results are reported in Sec. VI and the paper finally concludes in Sec. VII.

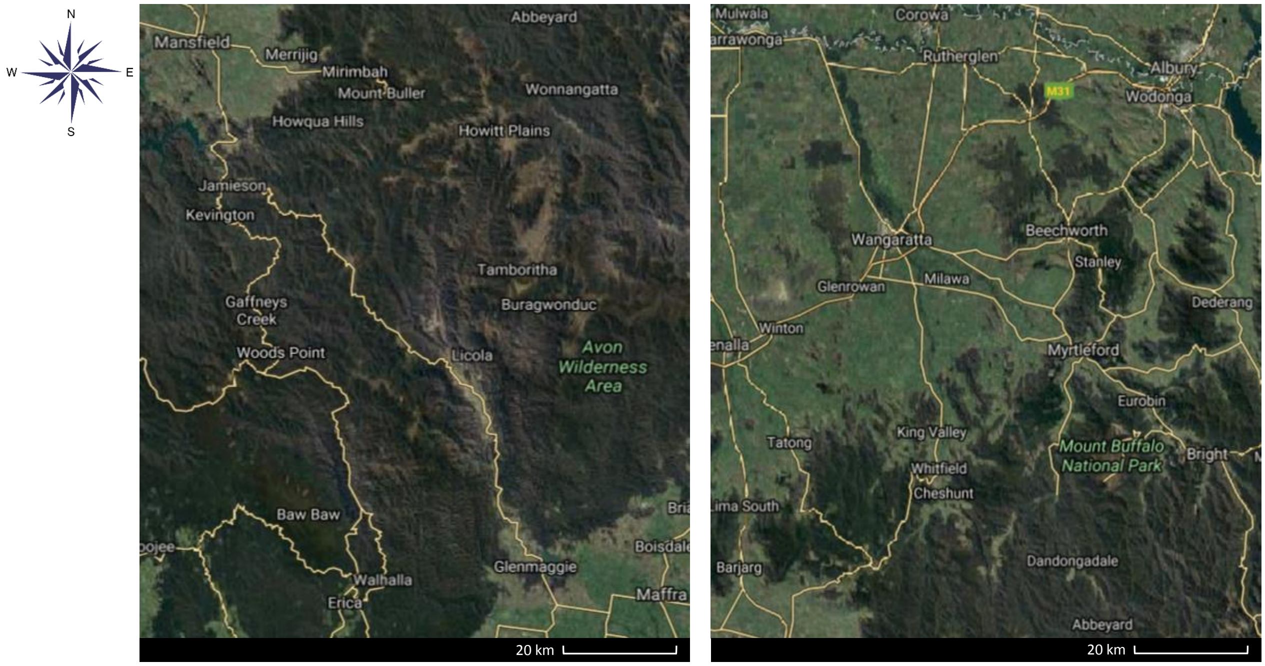
II. Related Work

The prevalent approaches for change detection in remotely sensed data can be categorized into two major classes; low-level local approaches and object-based approaches [4]. The low-level approaches use statistical indices derived from the pixel values of spectral images [17]. They are limited to pixel-level analysis, thus they remain agnostic to the valuable contextual information. A conventional approach to pixel-level change detection directly compares the contrast of bi-temporal (pair of) images acquired at selected dates when high-quality data was available [18]. Similarly, [11] extracts spectral indices to compare and detect changes in a pair of images. To study seasonal trends in multiple images, the temporal trajectories of coarse to moderate spatial-resolution spectral data have also been analyzed [12]. [19] proposed a pixel-level forest trend index and studied its performance on the Australian continent Landsat imagery. Compared to our approach, they perform analysis at a much coarse temporal scale (only 10 images during 1989-2006) and work on clean data acquired during dry seasons.

Other pixel-level change detection techniques use a vegetation index [20, 21], change vectors [22], spectral mixture analysis [23] and local texture [13]. Machine learning based classifiers such as Multi-layer Perceptron [24], Decision Trees [25] and Support Vector Machine (SVM) [14] have also been used for pixel-level change detection. However, these methods mainly use handcrafted features based on domain expertise.

The object-based approaches consider the contextual information by working on the homogeneous pixels, which are usually grouped together based on their appearance (spectral information), location and/or temporal properties [15]. One of the earliest work in object based change detection also uses the geometrical information of urban structures for object based analysis [26]. In most cases, standard unsupervised segmentation and grouping procedures are used to generate such pixel clusters [27]. Since these approaches work on region or object level, they are less prone to spectral variability, geo-referencing effects and errors in detecting land cover changes compared to pixel-level approaches [4]. Some object based approaches [13, 28] directly compare objects from different images to account for changes. In contrast, the approaches from [29] and [30] compare the extracted objects for change detection only after they are categorized into one of the desired classes.

One problem with object-based methods is that they heavily depend on the segmentation methods used for the generation



Db-37: [37^0S,146^0E]-[38^0S,147^0E]

Db-36: [36^0S,146^0E]-[37^0S,147^0E]

Fig. 2: The two study regions (*left* and *right*) for change detection are forests in Victoria, Australia (courtesy of Google Maps).

of objects [15, 31]. Not all objects generated in this manner are of the same size, and therefore over and under segmentation errors lead to less accurate change detection results [4]. To avoid such errors, we propose to generate bounding box candidates at multiple scales to detect interesting changes of varying sizes. Moreover, existing works use hand-crafted features or spectral indices derived from the objects for change monitoring [18, 31]. In contrast, this work automatically learns useful feature representations and predicts change likelihoods using a deep neural network.

Spectral remote sensing data suffers from several artifacts and various approaches have been proposed in the literature for preprocessing and data recovery [32, 33]. The preprocessing techniques deals with problems such as image registration for mosaic generation and the radiometric, atmospheric and topographic corrections needed to improve raw spectral data [34, 35]. From the perspective of frequent change analysis, a more crucial issue is the recovery of data missed due to sensor errors, seasonal and weather conditions. Data recovery approaches normally use image inpainting, multi-spectral and multi-temporal information [36].

Image inpainting approaches (e.g., [37, 38]) give visually pleasing results, however they fail to recover very large regions of missing data and the recovered information is not reliable for change analysis. Multi-spectral approaches (e.g., [39, 40]) use spectral information from other bands or sensors (e.g., MODIS in [41]) to estimate missing information in the Landsat ETM+ (Enhanced Thematic Mapper Plus) images. However, the spectral bands from other sensors suffer from differences in spatial resolution and bandwidths. The method in [42], called Automated Cloud Cover Assessment (ACCA), uses the reflective and thermal properties of the captured image for cloud cover estimation. This technique fails for the case of thin cirrus clouds (present at higher altitudes) because of their weak thermal signature. Compared to ACCA, Function of Mask (Fmask) [31] method for cloud and their shadow detection performs slightly better but still misses very thin cirrus clouds. Two types of auxiliary images are used by [33] to combine the high-frequency and the low-frequency information for data recovery.

Our approach for data recovery lies under the category of multi-temporal imagery based methods. These methods rely on both the temporal and spatial contextual information and work best for the recovery of large missing regions. One such approach from [32] assumes that land cover changes are insignificant over a short time-duration and use cloud-free patches to recover contaminated data. Similarly, other approaches (e.g., [43–45]) present sophisticated methods to perform data recovery across temporal domain by either re-adjusting the patch statistics or directly predicting the intensities. In contrast to these approaches, our method performs data recovery using reliable temporal information and can also recover regions contaminated by transparent clouds. Furthermore, the proposed approach is fairly straightforward and uses multi-resolution profiles which keep the recovered data consistent and reliable for valid change analysis.

The combination of complementary information obtained from multiple remote sensors has also been studied in the literature to remove mutual inconsistensies [46]. This proves to be useful because different data modalities have varying measurement resolutions, failure rates and sensitivity to atmospheric conditions (e.g., cloud cover). Shen *et. al*[47] fused high frequency and high spatial-resolution data streams to leverage the benefits of both for surface urban heat island anal-

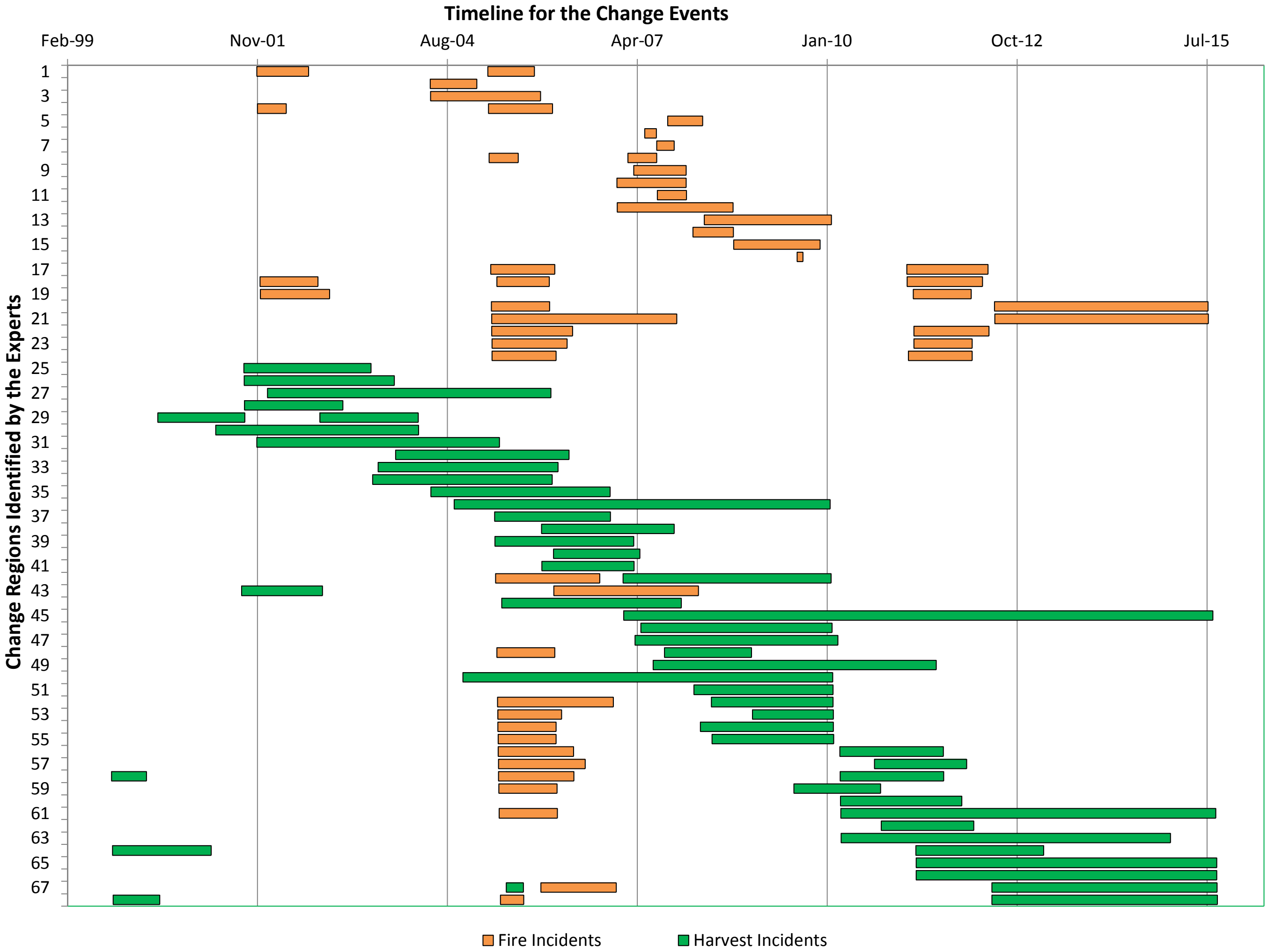


Fig. 3: Gantt chart of the fire and harvest incidents in the regions of interest identified during the period 1999-2015. Fire regions are usually recovered in a shorter period compared to the harvest regions.

ysis. Multi-sensory information was jointly used to produce better estimates of urban growth maps in densely populated regions [48]. Fablet and Rousseau [49] suggested an inpainting approach to benefit from the mutual strengths of microwave and infrared measurements for sea surface temperature. Apart from applications in geophysical analysis, interpolation of missing data using multiple data sources has also been used in biophysical monitoring e.g., vegetation mapping [50]. Different to these approaches, we only consider output from a single remote sensor to interpolate missing information to enable more frequent forest cover analysis.

More recently, convolutional neural networks (CNN) have been used for object detection and segmentation in remotely sensed multi-spectral images [16]. Penatti *et. al* [51] found that deep features that are extracted from a network pretarined on regular color images, generalize very well to satellite images. Transfer learning paradigm has also been investigated to learn better representations from remotely sensed data. Gueguen and Hamid [52] used a CNN model, fine-tuned on a large number of satellite images, for damage detection. Multi-scale convolutional architectures were also learned to obtain pixel-level segmentation in satellite images [53]. Among other applications, CNN models have been used for high-resolution remotely-sensed scene classification [54, 55], road network segmentation [56] and vehicle detection [57]. In contrast to these techniques, our approach deals with change detection in forest cover and provides a mechanism to extract and combine localized feature representations using a CNN.

III. Study Area

We analyzed a $222.4 \times 90.0\ km^2$ rectangular area in the north-east of Melbourne city in Victoria, Australia (Figure 2). The remote sensing satellite data is provided by the Australian Reflectance Grid (ARG) from the Geoscience Australia (GA). ARG is a medium resolution ($0.00025^0 \cong 25$m) grid of surface reflectance data based on United States Geological Survey's (USGS) Landsat TM/ETM+ imagery. To make the data comparable, robust physical models of [59, 60] are used to remove the differences caused due to sensor geometry, surface geometry, sun and atmospheric characteristics. With each of the surface reflectance image, a corresponding map of pixel quality flags is provided. For each grid cell, this map indicates

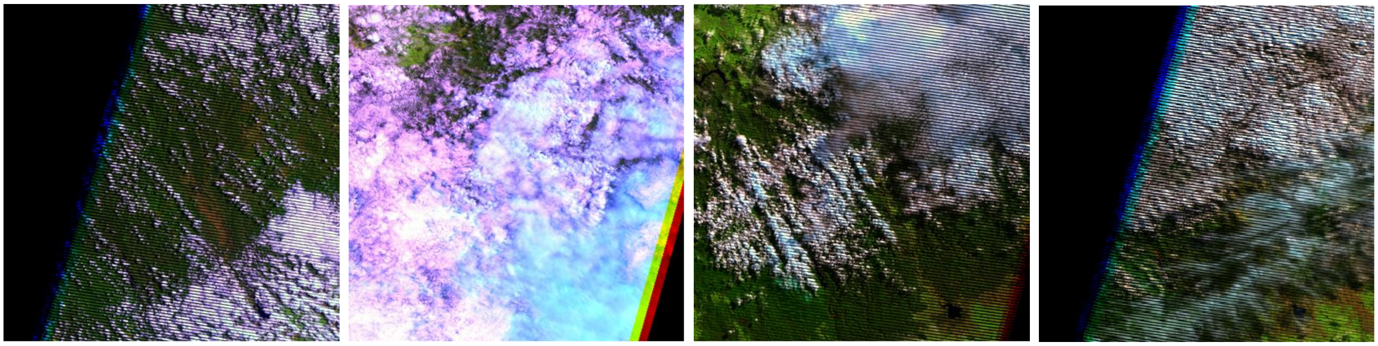


Fig. 4: Examples of artifacts in the data. There exist large regions of missing data along with strong clouds and their shadows. SLC-off artifacts are shown in the two *right-most* images which appear as slanted wedge-shaped regions of missing data (*figure best seen when enlarged*). Artifacts constitute almost **75.9%** of the data.

the presence or absence of null values, band saturation, clouds and cloud shadows. The processed reflectance data is referred to as the Landsat Nadir Bidirectional Reflectance Distribution Function (BRDF)-Adjusted (NBAR) images.

The flags included in the pixel quality map are shown in Table I. Notice that, the two cloud flags are included in the map based on two different methods. The first cloud detection method used is the ACCA algorithm of [42, 58]. The second cloud detection method used is the Fmask algorithm proposed by [31]. Fmask utilizes Top of Atmosphere Reflectance (TOAR) for cloud detection and performs better than the ACCA algorithm. Therefore, in this work, we use clouds detected with the Fmask algorithm during the preprocessing phase. It is important to note that very thin clouds are still missed by both methods and therefore we describe our approach to remove such clouds in Sec. IV-C.

The study area is divided into two regions of equal dimensions. Since the available data of both regions belongs to different time-ranges, we refer to the region between coordinates $37^0 00'00.0"$S $146^0 00'00.0"$E and $38^0 00'00.0"$S $147^0 00'00.0"$E as Db-37 and the region between coordinates $36^0 00'00.0"$S $146^0 00'00.0"$E and $37^0 00'00.0"$S $147^0 00'00.0"$E as Db-36. For Db-37, we have a time lapse sequence between 1999-2015 (17 years) of surface reflectance data and the corresponding pixel quality maps. For Db-36, we have surface reflectance data and pixel quality maps for years 1987-2014 (28 years).

TABLE I: The flags included in the pixel quality map available with the Landsat NBAR images. The remaining bit locations are not currently used.

Bit Position	Flag Purpose
0-4	Respective band 1-5 is saturated
5	Band 6-1 is saturated
6	Band 6-2 is saturated
7	Band 7 is saturated
8	Contiguity (No Null Values)
9	Land or Sea
10	Clouds (ACCA [42, 58])
11	Clouds (Fmask [31])
12	Cloud shadows (ACCA)
13	Cloud shadows (Fmask)
14	Topographic Shadow

The remote sensing data is labeled with two types of forest changes, namely harvests and fire incidents. During the period of 17 years in the region Db-37, a total of 99 incidents were manually identified by experts, out of which 50 were fire incidents while the remaining 49 were harvest incidents. These 99 change incidents happened at 68 distinct sites. Similarly, a total of 49 incidents were recorded in Db-36 during the 28 years period, out of which 14 were fire incidents while 35 were harvest incidents. These change events took place at 29 different sites. The Gantt chart representation for both types of annotations in Db-37 is shown in Figure 3. Note that the fire incidents usually last for a much shorter period (and also recover quickly) compared to the harvest incidents.

IV. Data Recovery

The data under investigation contains several artifacts due to which land cover is not always visible in the ARG (see examples in Figure 4). These artifacts include missing surface reflectance data, heavy clouds and saturated channels in remotely sensed data. Moreover, black stripes (wedge shaped gaps) appear in the Landsat-7 ETM+ imagery due to the failure of the scan line corrector (SLC) in 2003. There is no temporal relationship between the missing data locations, i.e., these locations do not remain consistent at different instances of time. To illustrate by an example, approximately 40.7% of the total reflectance data in the Db-37 is missing while nearly 35.2% of the data is cloudy. For land cover change analysis and detection, it is necessary to remove these artifacts, which make a staggering $\sim$**75.9%** of the reflectance data in Db-37. In the current work, we do not aim to remove light cloud shadows or topographical shadows, which also create visual artifacts but are not as severe as the artifacts described above.

To fill-in the missing data and the residual cloudy regions, we design a three-stage image completion process that exploits the redundancy in the raw image data. The first stage deals with large gaps by assessing the reliability of data along the temporal domain (Sec. IV-A). The second stage performs a spatial refinement to remove noisy data and ensure spatially consistency (Sec. IV-B). The last stage performs further refinement by removing very thin and transparent clouds

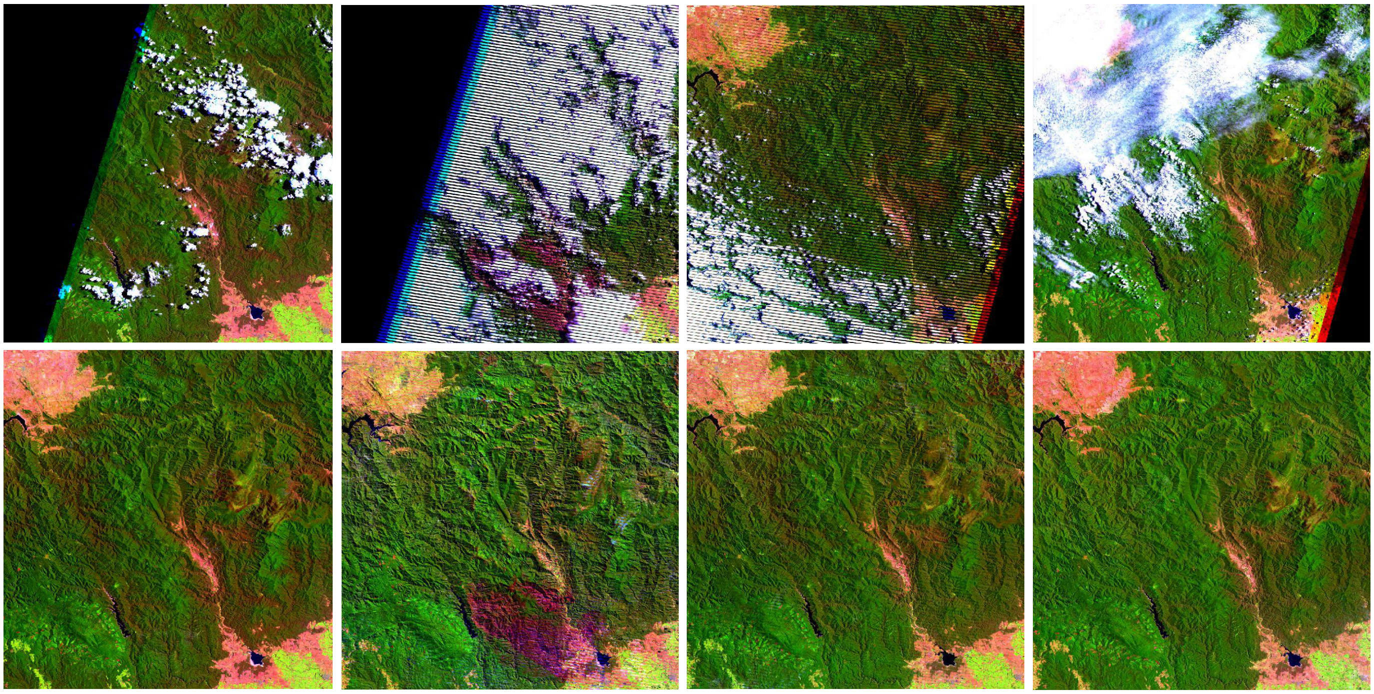


Fig. 5: Data recovery results on single frames: The *top row* shows raw spectral data, which suffers from several artifacts including weather conditions. The *bottom row* illustrates our recovered images, which are visually more pleasing and more suitable for further analysis.

(Sec. IV-C). These three stages are elaborated in the following sections.

A. Gap Filling

In the first stage, we fuse the reliable data along the temporal dimension to generate one representative image for a period of approximately two months using the corresponding flags in the available pixel quality map. Then, we construct a mean image from the representative images to obtain a yearly background profile, which we employ consecutively to fill the remaining missing pixels in the original images. In our experiments, this temporal strategy yields better performance than pixel-wise interpolation that only uses spatial information leading into additional artifacts. Since the original satellite images were acquired at an average frequency of 12 days, this stage can fill in a large percentage of missing pixels without affecting forest change events, which are usually slower processes.

B. Masked Sparse Reconstruction

In the second stage, we further enhance the image frames using masked sparse reconstruction to enforce the spatial consistency and remove possible artifacts generated from the first stage. We elaborate our approach below.

Given a set of input images $\{\mathcal{I}\}_{1\times N}$, we first extract same size overlapping patches with dimensions $s \times s$ and a uniform step of p. These patches form a set $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{M}$, where normally M is a considerably large number. To make the dictionary learning step computationally feasible, we randomly choose a relatively smaller set of patches denoted by $\hat{\mathcal{P}} = \{\mathbf{p}_i\}_{i=1}^{m}$. Typically, the learned dictionary is composed of r basis vectors, where $r << m$[1]. The objective minimized during the dictionary learning process is defined as follows:

$$\min_{\mathbf{D}\in\mathcal{C}} \frac{1}{m}\sum_{i=1}^{m} \min_{\alpha_i\in\mathbb{R}^r} \left(\frac{1}{2} \parallel \mathbf{p}_i - \mathbf{D}\alpha_i \parallel_2^2\right) + \lambda \parallel \alpha_i \parallel_1 + \gamma \parallel \alpha_i \parallel_2^2 \quad (1)$$

where, λ and γ are the regularization parameters which enforce a sparse solution for α_i. The set $\mathcal{C}$ is the constraint set of matrices defined as follows:

$$\mathcal{C} = \{\mathbf{D} \in \mathbb{R}^{q\times r} \quad s.t., \parallel \mathbf{d}_j \parallel_2^2 \leq 1, j \in [1, r]\}. \quad (2)$$

The above constraints on the basis vectors (columns of dictionary $\mathbf{D}$) avoid arbitrary large values in the learned dictionary. Notice that, we form an over-complete dictionary by setting a small patch size (s), therefore $r > s^2$. The sparse coding problem posed in Eq. 1 is solved using the online dictionary learning algorithm of [61].

Once a dictionary has been learned, each image patch $\mathbf{p}_i$ can be reconstructed using a sparse combination of basis vectors in $\mathbf{D}$. However, as discussed in Sec. IV, the spectral data has severe artifacts and it is quite possible that some of the regions are still not fully recovered during the first-stage inpainting procedure. If we perform a normal reconstruction step using all the pixels in a given patch, it will lead to errors because some of the patch information may not be valid (appearing usually as black regions). Therefore, during the sparse reconstruction step, we only reconstruct the valid regions (original valid data and the recovered regions in the inpainting step in Sec. IV-A)

[1] In our experiments, the following parameter settings were used: $s = 8, p = 2, m = 5 \times 10^5$ and $r = 512$. The total number of patches (M) were $\sim 2.0 \times 10^9$ and $\sim 3.8 \times 10^9$ for Db-37 and Db-36 respectively.

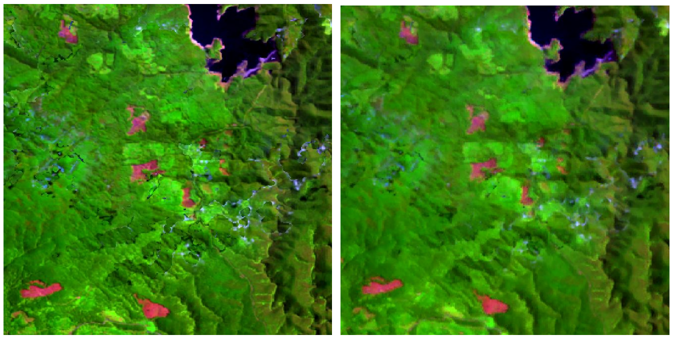


Fig. 6: *Left:* Gap filling output, *Right:* The masked sparse reconstruction step reduces noise and removes boundary effects caused by the gap filling.

and do not include the missing pixels in the approximation process (Eq. 3). This step fills in small regions of missing data and removes abrupt changes in pixel contrast since the dictionary $\mathbf{D}$ is constructed from only clean patches. The objective function for this recovery step can be formulated as follows:

$$\min_{\alpha_i \in \mathbb{R}^r} \frac{1}{2} \parallel \mathbf{M}_i(\mathbf{x}_i - \mathbf{D}\alpha_i) \parallel_2^2 + \lambda' \parallel \alpha_i \parallel_0, \qquad \forall i \in [1, M], \tag{3}$$

where, λ' is a regularization parameter to enforce sparsity, $\mathbf{M}_i \in \mathbb{R}$ is a mask defined as a diagonal matrix: $\mathbf{M}_i = \text{diag}(\beta_i)$ and $\beta_i \in \{0,1\}^{s^2 \times 1}$. The mask $\mathbf{M}_i$ encodes the validity of each pixel. More precisely, the pixels that are not recovered during the first stage of recovery process are marked as invalid pixels.

The optimization problem in Eq. 3 is solved using the orthogonal matching pursuit (OMP) algorithm [62]. A final complete image is obtained by combining all the small patches $\mathbf{p}_i$ and performing an averaging operation over overlapped regions. Finally, note that the sparse reconstruction step is performed individually for each channel of the reflectance data by learning a separate dictionary. This preserves the distinct information in each spectral band and ensures a consistent recovery of the missing information. The improvement is illustrated via an example in Fig. 6.

C. Thin Cloud Removal

The third stage of our data recovery addresses the residual thin clouds in the recovered images. At this stage, all the missing data regions are filled-in, however, some partially-missing regions can still occur due to the thin clouds. Note that, state-of-the-art cloud detection methods (ACCA and Fmask) fail to find thin layers of clouds. Besides, the pixel quality map (Sec. III) does not indicate their location. These translucent regions cause problems during the later stages of change detection (e.g., region proposal generation). Therefore, we devise an efficient approach based on color heuristics to remove the thin clouds (Figure 7).

In the forest region under consideration, thin clouds appear in the Band 1 of the surface reflectance data (blue pixels in Figure 4). More importantly, these thin clouds appear and disappear abruptly and do not occupy one spatial location for a long time period. To detect thin clouds, we build a multi-resolution profile (MRP) of a spatial region. Note that the multiple resolutions are considered along the temporal domain to build MRP.. The MRP has two distinct levels, the higher level consists of the back-ground profile of an area generated by averaging all the valid pixels in the entire time-range i.e., 1999-2015. The lower level comprises of the yearly profile of an area generated by averaging all the valid pixels within one year. In order to detect thin shadows, we first compare the yearly profiles with the background profile and compute a thresholded-difference image using Band-1. This band captures the wave-length range (0.45-0.52 μm) where thin clouds are clearly visible. The detected regions in the yearly profiles are replaced with the background profile data to remove thin clouds.

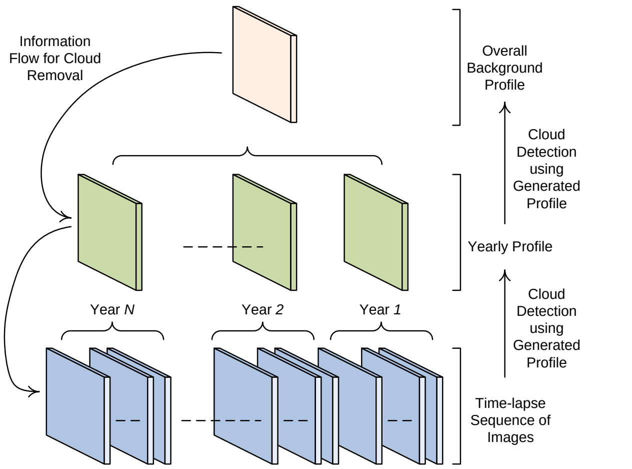


Fig. 7: Our approach to detect and remove thin translucent clouds which are missed by the current stat-of-the-art techniques.

At the next level, we repeat a similar procedure with the images and the yearly profiles. A thresholded-difference image is computed by comparing each image with its yearly profile image. The detected regions in the images are replaced by the values in the corresponding locations in the yearly profile image. This hierarchical procedure (along the temporal domain) has the advantage of being simple and efficient, while only affecting Band-1 and therefore landscape change regions remain unaltered. The use of multi-resolution average profile during the hierarchical restoration process ensures that the irreverent noisy information (e.g., clouds and topographical shadows) is filtered out and the filled in values are reliable for change detection. After the data recovery process (see Figure 5 for examples), we now have complete image frames for the time-lapse sequence (approximately one frame for every two months), which are used for the forest change detection described in the next section.

V. Change Detection

We formulate the change detection task as a region classification problem where we first identify change area proposals

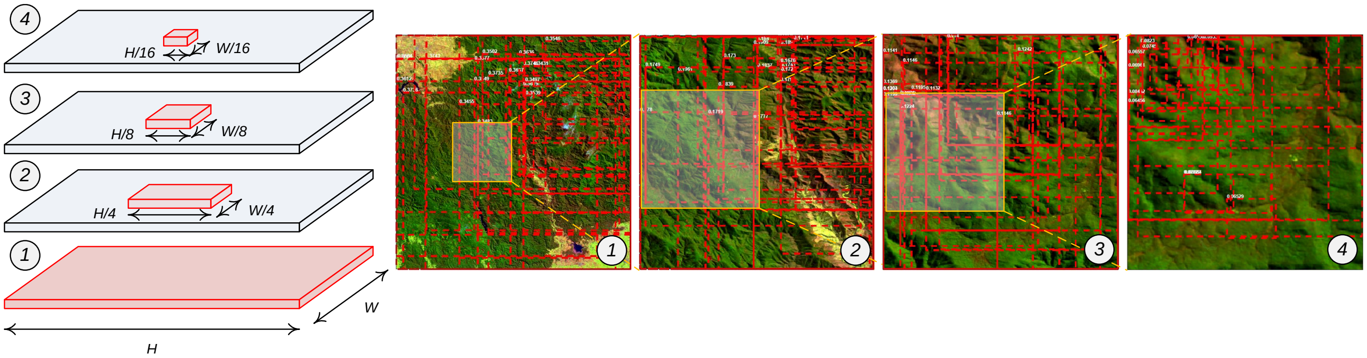


Fig. 8: Box proposals are generated at multiple scales to capture all sizes of change events. The constants 'H' and 'W' denote the height and the width of the original image respectively.

and then apply a deep CNN to detect the change or no-change. More specifically, we consider the healthy forest cover under normal conditions as a no-change region. A forest region which undergoes a change event is labeled as a change region. If a change event is detected, we predict the type of change. With the proposed approach we are able to detect multiple change events (of same or different type) happening at a particular location.

A. *Multiscale Region Proposal Generation*

To detect forest changes, our approach initiates with the generation of a set of candidate change regions. We use a selected set of spectral bands (more details in Sec. VI-B) which provide an appropriate visualization for fire and harvest changes. Since these changes have different appearance, texture and shape characteristics, we can apply standard computer vision methods (usually tailored for RGB images) to this problem.

The region candidates are generated using the MRP described in Sec. IV-C. Since the changes of interest mostly span a time frame comparable to one year, we generate the initial set of potential change region candidates using each of the yearly profiles in the MRP. The bounding boxes enclosing the regions of interest are generated using the edge based object proposal method (EdgeBox) [63]. Note that the notion of edge and contours is similar in both the reflectance data and the color images. We use the structured edge detector model pre-trained on the Berkeley Segmentation Data Set 500 (BSDS-500 dataset) for edge detection [64]. Since the model is pre-trained on regular color images, we obtain RGB images from the reflectance data by selecting the relevant channels which provide a natural-looking visualization of vegetation and fires (see Sec.VI-B for details).

From the expert annotated change regions, we observe that there exists a large disparity among the relative sizes of bounding boxes for different change events. For example, the fire regions are usually large (up to $\sim$80% of the total area enclosed by the entire image) and the harvest regions are usually tiny (up to $< 0.005\%$ of the image area). To tackle this problem, we propose a scheme that uses the original image as well as the patches extracted at multiple scales to generate bounding-box proposals.

Our approach to use multiscale patches is illustrated in Figure 8. We use four scales during the proposal generation process, each with different sized patches. More precisely, the sizes relative to the original image dimensions are : 1×1, $\frac{1}{4} \times \frac{1}{4}$, $\frac{1}{8} \times \frac{1}{8}$ and $\frac{1}{16} \times \frac{1}{16}$. To avoid missing any change regions which appear close to the patch boundaries, we extract overlapping patches with a step size equal to 80% of the shortest patch dimension. For every patch, we allow a fixed maximum number of boxes (M_{box}) with scores higher than S_{box} to be detected. By varying M_{box} and S_{box}, we can obtain varying number of boxes. We note that generating a large number of proposals gives a better overlap ratio with the manually identified change regions by the experts. However, it also results in redundant proposals and a high computational load during the subsequent processing steps.

Next, we describe our approach to refine the initial candidate set by removing the redundant and unwanted box proposals.

B. *Candidate Suppression*

The initial set of box candidates is further refined to reduce the computational load without affecting the detection accuracy. First, we generate a change map by comparing the yearly profile to the overall background profile in the MRP. Since a change map consists of pixel-wise intensity differences, it captures any visible changes happened on the image plane. The change map is refined by morphological operations (erosion followed by dilation). Afterwards, we retain the candidate boxes whose at least 20% area was changed. We also ignore the box proposals that enclose a very high percentage ($> 90\%$) of the total changed area. This results in the suppression of several small and unnecessary box proposals which do not fully cover a particular change event. Finally, we perform a non-maximal suppression of bounding boxes to remove redundant proposals. This suppression step ignores the lower scored bounding box for each pair of overlapping boxes (overlap ratio defined by IOU). The recall rates of a varying number of proposals generated in this manner (for different values of M_{box} and S_{box}) are shown in Figure 10. We note that the generated proposals provide a reasonable coverage ($> 94\%$) for $2000 - 4000$ box proposals in an area of $\sim 10^4\ km^2$.

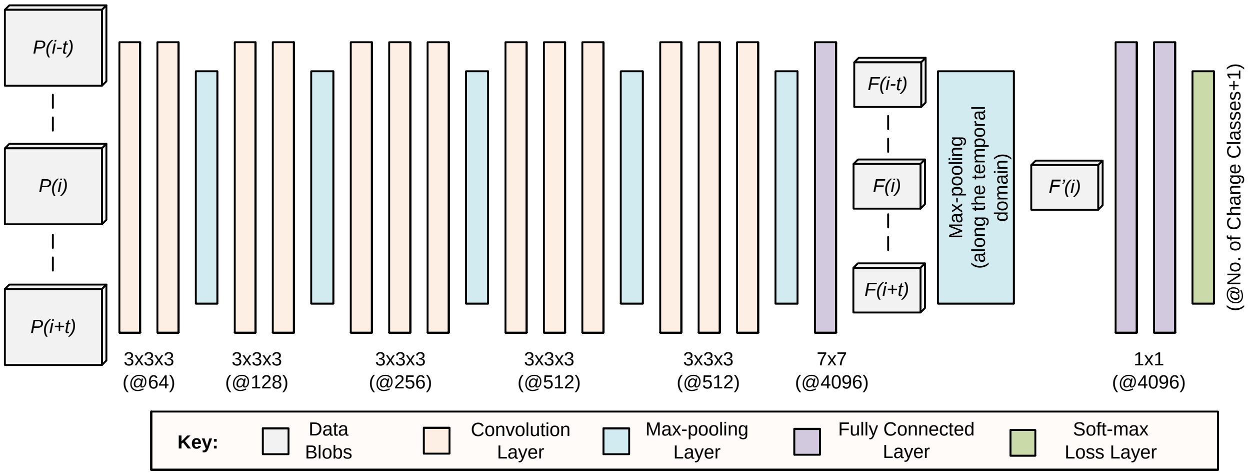


Fig. 9: The CNN architecture used for forest change detection. The network takes a series of patches $(P(i-t)\dots P(i+t))$ centered at a given time instance for each change area proposal. The feature representations are fused together after the first FC layer using a max-pooling operation to produce temporally consistent and smooth features.

C. Deep Convolutional Neural Network

We use a deep CNN model to map the raw patch data to a discriminative feature space which is then used to detect relevant changes. Given a candidate set of change regions, the deep neural network predicts whether each patch belongs to no-change category or a change category (either fire or harvest change). The network architecture comprises of 17 weight layers, whose filter sizes and number of filters are shown in Fig 9. Upto the 14^{th} weight layer (first fully connected (FC) layer), the network architecture is identical to the state-of-the-art VGG-16 net (configuration-D) [8]. Afterwards, we perform feature pooling along the temporal domain to generate temporally smooth predictions. Notice that, for every time instance i, the network is fed with $2t$ before and after frames whose features $(F(i-t)\dots F(i+t))$ are pooled with the current frame (i^{th}) features to generate a refined representation $(F'(i))$. The max-pooling operation is used for feature fusion and performs better than the average pooling in our experiments. The temporally pooled feature representation $F'(i)$ is then used by the subsequent FC layers and finally the output layer to predict the change class. We set the temporal window size $t = 3$ by cross-validation, which gives a moderate boost over the non-pooled features (see Sec. VI-D).

The input patches fed to the network are of varying sizes, including some very large as well as very small bounding boxes. To remove this disparity, we ensure that the smaller dimension of an image patch is within the range $[224, 480]$ by proper upsampling or down-sizing. From each input image patch, we extract 224×224 windows with step size of 64 to feed equal-sized inputs to the network. The 4096 dimensional feature vectors of all these windows (obtained after the first FC layer) are then max-pooled to obtain a single representation of each distinct patch. The mean image is also subtracted from each input patch which enhances the discriminative ability of features.

The network parameters are comparatively huge (around 139 million) compared to the available patch labels for harvest and fire changes. Therefore, we initialize the first 14 layers from the network pretrained on the ImageNet dataset [8] and perform fine-tuning using the available surface reflectance data. The last two FC layers are initialized with random weights and learned from scratch for change detection. We also note that since the fire events last for a relatively short time, their representation is comparatively lower in the training set which results in a lower test performance. To avoid this class imbalance problem, we upsample the less frequent change event data to make sure that both types of change events have nearly an equal representation in the training set. The upsampling is achieved by adding identical, flipped, rotated and cropped copies of the less frequent class samples.

During the test phase, we input multiple patches to the network (similar to the training phase) and perform temporal feature pooling after the first FC layer. The predictions made by the network are temporally smooth and directly compared with the ground-truths for evaluation (see example predictions in Figure 13). More experimental and evaluation details are described in the next section.

VI. Experimental Analysis

A. Evaluation Tasks

We test our algorithm on four standard tasks. The first two tasks pertain to an ablative analysis to study the localisation and patch-level classification performance of our approach. The next two tasks relate to the time-series change detection and start/end time prediction for change events.

1) Tasks for Ablation Study:

a) Localisation of Change Events: In this task, we assess the quality of the bounding-box proposals generated by our approach. Since only a limited number of change locations have been identified in the available annotations, we quantify the quality of proposals by finding the proportion of labeled change boxes that are matched by the generated proposals.

b) Patch-level Change Classification: For this task, we treat the change detection problem as a classification task. Therefore, for a given time-lapse sequence, we treat each frame as an independent instance and predict whether or not a change happened in a given frame. For evaluation, we use the overall accuracy and the recall measure averaged over the classes.

2) Tasks for Change Detection:

a) Time-series Change Detection: For this task, we make use of the temporal information while making change predictions. To enforce a temporal consistency in the predictions, we perform feature fusion in a small window defined over features computed for the same region at adjacent time instances. We also, smooth the output predictions from the baseline approaches to have a uniform detection pattern. The evaluation metric used for this case is the average intersection over union (IOU) score obtained over all the labeled change regions.

b) Change On/Offset Prediction: In this task, the onset and offset of a change event is predicted for a given region. Information across multiple time instances is used to predict a smooth change sequence and to avoid multiple noisy spikes in the prediction. For evaluation, we use the mean taxi-cab distance for both the onset and offset points of change event predictions.

B. Experimental Settings

In all our results, we report performances on the complete dataset including both the original and the recovered regions. It is important to note that the recovered regions make a significant portion of the dataset under investigation and therefore make the change detection highly challenging. We perform 10 fold cross-validation by keeping the train vs. test split to 90% vs. 10%. Mutually exclusive sets of change locations are used for training and testing procedures, and care has been taken to ensure that an event does not split between the train and the test set.

We use a combination of bands 5, 4 and 1 from the Landsat 7 imagery and bands 6, 5, and 2 from the Landsat 8 imagery for training and testing. These band combinations for Landsat 7 and 8 are suitable for natural-looking visualization of vegetation and fires (see Figures 5, 14 and 15). The healthy, dry and sparse vegetation appears in bright green, orange and brown colors respectively. Grasslands appear in light green color while water is usually blue. The fire regions appear in dark red color. Since these band combinations provide a natural looking visualization of forest cover, we can apply standard computer vision algorithms and pre-trained models on the spectral data.

To enhance the contrast of the image, we perform a uniform rescaling of the red, green and blue channels within the ranges of 0.0055-0.0463, 0.0132-0.0600 and 0.0029-0.0175, respectively. This helps in the feature extraction process and the uniform mapping ensures that multiple frames remain comparable to each other for multi-temporal analysis.

C. Baseline Approaches

We compare our approach with strong baselines which use popular handcrafted features and strong machine learning classifiers. These baselines are described next.

1) Handcrafted Features for Classification: We use dense Scale Invariant Feature Transform (SIFT) descriptors as a baseline for change detection. Based on these features, we experiment with three classifiers: i) linear support vector machine (SVM) for max-margin classification, ii) kernel SVM for nonlinear classification, and iii) random forest (RF) for ensemble learning based classification. For the kernel SVM, we use the efficient homogeneous kernel mapping [65] to approximate the χ^2 kernel. The SIFT descriptors are computed on a dense grid and the classifier is directly trained on these local features. Note that this was feasible because the pixel labelling of the change regions is known within each patch. During the testing phase, we classify a given image patch as a change region if at least 15% of the SIFT descriptors are classified as the fire or harvest change. This percentage was set using cross-validation experiments, which provided approximately equal true-positive and true-negative rates.

2) Bag-of-Visual-Words (BoW) for Classification: For the BoW baseline, we use dense SIFT as local features and efficiently compute a dictionary using the k-means clustering. The number of bins is set to 600 by cross validation. All the features are then represented in terms of associations with the dictionary atoms. A conventional BoW model does not preserve the spatial information. However, this information can be useful to categorize change patterns with distinctive shapes. Therefore, to incorporate the spatial information, we use disjoint spatial bins to compute histograms which are then stacked together to obtain a final representation. Similar to the previous baseline, we use linear SVM, χ^2-kernel SVM, and RF classifier for prediction.

D. Results

1) Ablative Analysis: We first evaluate the performance of our bounding-box proposal generation scheme on the Db-37 region. The varying number of bounding-box proposals affect the amount of coverage for the labeled change regions. The trend is illustrated in Figure 10. We consider a successful match between the ground-truth bounding-box and the generated proposal if their IOU > 0.1. We generate different number of bounding-box proposals by changing values of the constants M_{box} and S_{box}. The higher number of box proposals provide more coverage but also require more computational resources for further processing. To make a balanced choice, we set $M_{box} = 30$ and $S_{box} = 0.05$ in our experiments to generate ∼1900 box proposals, which cover 94% of the labeled change regions.

We have also experimented with other box proposal generation methods and analyzed their performance compared to EdgeBox. These box proposal methods include selective search [66], constrained parametric min-cuts (CPMC [67]) and objectness measure [68]. The parameters of these models were set to generate nearly the same number of boxes as generated by the EdgeBox. In the cases where the number of generated boxes was very large (e.g., objectness measure), we only considered the box proposals with the highest score for evaluation. For each of these methods, we recorded the

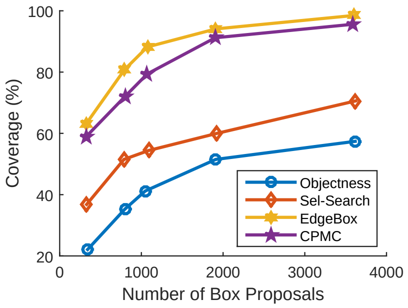


Fig. 10: Labeled change region coverage by the different number of bounding-box change proposals in Db-37.

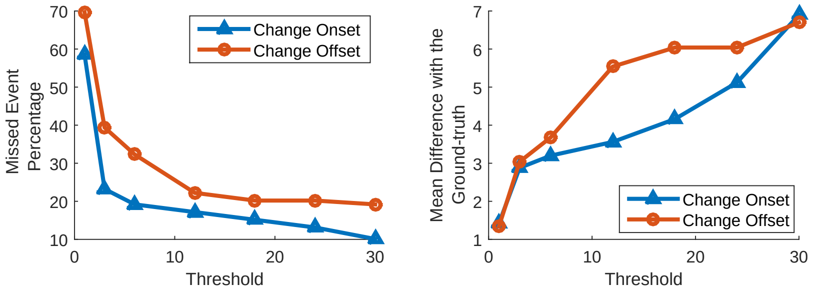
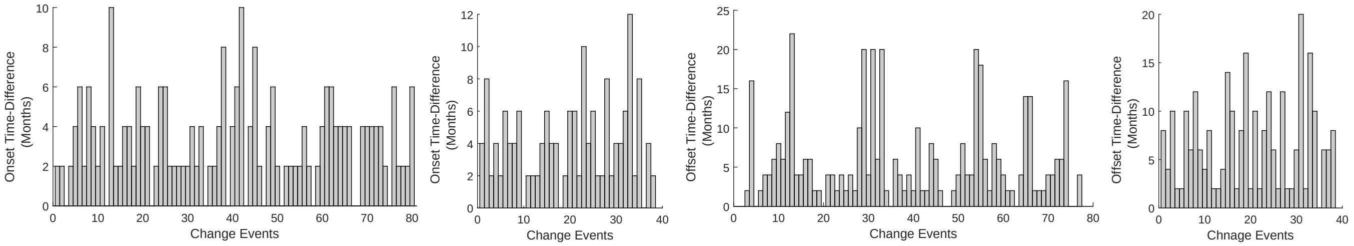


Fig. 11: The trend of missed events and mean onset/offset difference when the temporal threshold for a valid detection in Db-37 is changed from low to high.



(a) Onset time differences between actual and predicted outputs for Db-37 (*left*) and Db-36 (*right*) respectively.

(b) Offset time differences between actual and predicted outputs for Db-37 (*left*) and Db-36 (*right*) respectively.

Fig. 12: Onset and offset detection results for individual fire and harvest events. The bar plot shows differences in months, the events which were not detected have been excluded. It is important to notice that the change detection is performed on multi-temporal images with nearly two months gap.

TABLE II: Db-37 Region: Patch-wise classification and detection results for the temporal sequence are summarized below. All performance numbers are in percentages (%).

Method	Accuracy	Avg. Recall	Mean IOU
SIFT+l-SVM	68.1	57.3	50.2
SIFT+k-SVM	71.3	61.4	54.8
SIFT+RF	69.7	58.8	50.6
BoW+l-SVM	72.6	63.1	54.8
BoW+k-SVM	74.1	64.9	57.1
BoW + RF	71.7	64.0	54.7
This paper	**92.0**	**84.6**	**84.7**

TABLE III: Db-36 Region: Patch-wise classification and detection results for the temporal sequence are reported below. All performance numbers are in percentages (%).

Method	Accuracy	Avg. Recall	Mean IOU
SIFT+l-SVM	69.0	57.0	54.2
SIFT+k-SVM	71.8	59.3	57.8
SIFT+RF	71.3	58.6	55.1
BoW+l-SVM	74.2	63.0	61.5
BoW+k-SVM	76.5	65.6	64.9
BoW + RF	73.9	61.5	60.0
This paper	**91.3**	**70.8**	**85.4**

percentage coverage of the ground-truth change regions when the number of box proposals is increased in Fig. 10. Furthermore, we have also evaluated the performance of the random box generation around the change regions. Specifically, for random box generation, we obtain a thresholded change mask for each yearly profile image and generate randomly sized boxes with in the range of ground-truth change box sizes (by varying the box diagonal). Afterwards, the desired number of boxes are randomly selected as the candidate set for further processing. With the same number of boxes as the ones used for EdgeBox ($\sim$ 1900), the coverage rate turned out to be very low (17.6%). However, we noticed a consistent increase in the coverage rate and for a very huge number of boxes (50,000), we obtained a good coverage rate of 86.7%.

We performed patch level change/no-change prediction by treating the problem as a classification task. In Table II), we report the overall accuracy, the average recall rate and the mean IOU (averaged over all classes). We noticed a comparatively higher performance when features were fused in a small window along the temporal dimension to get an improved feature representation at each time instance. We tested different sized windows and observed that a medium sized window (size = 7) performed best. The baseline approaches perform fairly low (accuracy and recall difference as much as $\sim$24% and $\sim$27%, respectively) compared to our proposed approach. Among the two baseline techniques, the bag-of-words based procedure performs better than the low-level SIFT features. In terms of classifiers, the homogeneous approximation of χ^2-kernel outperforms consistently the linear SVM and RF alternatives.

TABLE IV: Our results for onset/offset detection and comparisons with several baseline techniques are reported for Db-37 region. The error units are months (Mn) and it is defined as the mean taxicab distance.

Method	Onset Error (Mn)	Offset Error (Mn)
SIFT+l-SVM	8.7 ± 4.1	15.1 ± 7.5
SIFT+k-SVM	8.3 ± 4.1	14.9 ± 7.2
SIFT+RF	8.9 ± 4.3	15.9 ± 7.7
BoW+l-SVM	7.4 ± 3.6	13.5 ± 6.9
BoW+k-SVM	7.1 ± 3.4	12.6 ± 6.8
BoW + RF	7.4 ± 3.7	13.8 ± 7.1
This paper	**3.2 ± 2.3**	**5.5 ± 5.5**

TABLE V: Onset/offset detection results and comparisons with several baseline techniques are reported for Db-36 region.

Method	Onset Error (Mn)	Offset Error (Mn)
SIFT+l-SVM	9.2±3.7	17.4±7.7
SIFT+k-SVM	8.4±3.8	15.5±7.0
SIFT+RF	9.0±3.5	17.5±7.7
BoW+l-SVM	7.8±3.5	14.2±6.0
BoW+k-SVM	6.8±3.1	12.9±5.7
BoW + RF	7.5±3.3	14.4±6.1
This paper	**4.1 ± 2.7**	**6.9 ± 4.8**

2) Change Detection Results: For the temporal change detection task, since it is highly unlikely to have forest changes taking place abruptly at close-by time instances, we further smooth the output predictions made by the baseline procedures. For this purpose we used a uni-dimensional median filter with a comparatively higher window size of 5 (equivalent to ∼10 months data). We note that the outputs from our CNN based approach with feature fusion are already smooth and do not need further processing. Therefore, our final classification and detection results which are reported in Tables II and III use feature level fusion but do not use any output prediction smoothing. Sample results of ground-truth and predicted sequences for the case of fire and harvest events are shown in Figure 13. Our approach provides temporally-smooth labelings and it was able to detect multiple changes of similar and different types occurring at a particular change site.

We analyze the accuracy of onset/offset prediction for each change event in Tables IV and V. We show the differences between the onset/offset points in the output predictions and ground-truth annotations for each distinct event in Figure 12. On average, the start point of each predicted change event in Db-37 and Db-36 differs from the ground-truth change sequence by 3.2 ± 2.3 and 4.1 ± 2.7 months respectively. The average end point difference between the predicted and ground-truth changes in Db-37 and Db-36 is 5.5 ± 5.5 and 6.9 ± 4.8 months respectively. For the change onset, we consider a valid detection to be the one which lies within one year of the ground-truth change event start point. For the case of change offset, the maximum permissible gap between ground-truth and predicted end-times is set to two years because the changes recover slowly and there is no definite change end time. An event is considered as missed if the predicted onset and offset time is higher than the permissible limits. With the above mentioned limits, 19.2% and 22.4% change onsets are missed while 22.2% and 22.4% change offsets are missed for Db-37 and Db-36 respectively. It is important to note here that the change patterns are not very clear in most cases and the ground-truth annotations (especially for offsets) are based on a subjective judgement.

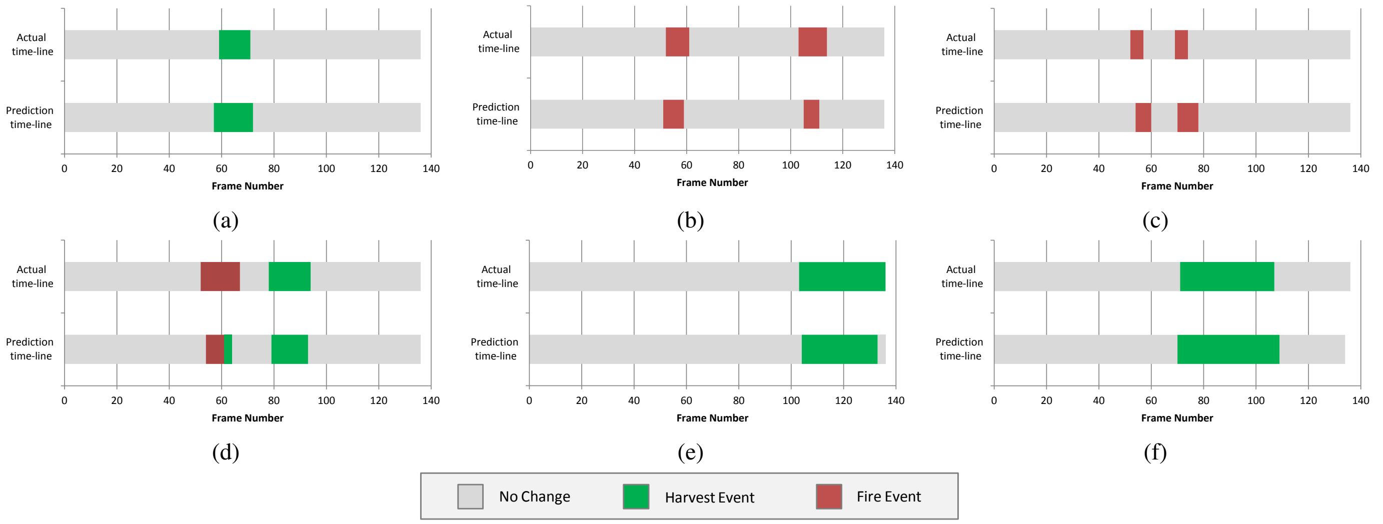
To study the effect of permissible limits (error threshold) on the change on/offset performance and the percentage of missed events, we experiment with different thresholds. The trends of missed events and performance can be seen in Figure 11. We note that as the onset/offset error threshold (in months) is increased, the percentage of missed events decreases steadily. However, the mean error in terms of taxicab distance increases with the increase in error threshold. Another important observation is that the error for change onset is comparatively lower than the change offset error. This can be explained by the fact that although the change events usually start at one particular point in time, the recovery process is slow and does not finish at a single time instance.

We also qualitatively analyze our detection results on unlabeled regions with in the study area. Figure 14 shows our detection results on the full image plane (example frame taken from year 2003). In addition to the labeled change sites, our approach was able to identify new change sites (see *bottom* two rows) and also predict their change on/offset points in the time-series data. Only in the shown example, our method discovered more than 10 new change sites. We also noticed a few false detections e.g., one in the middle-right of the top image in Figure 14. Our approach can drastically reduce the human effort required for full change annotations by introducing a human in the loop to eliminate any false detections on the unlabeled patches. The proposed system can then be trained on the enhanced training set (including the newly generated training data) which will further improve the detection ability.

The qualitative results of our approach on selected parts of the time-series data are shown in Figure15. In particular, we show three challenging image sequences where our approach was not fully accurate. The illustrated examples include both the fire and harvest changes. The first challenge was that the changes recover slowly, thus a mismatch between the predicted offset timings was evident in some cases (e.g., the top and bottom sequences in Figure 15). Secondly, change events of different types occurring in nearby regions affect the overall performance. For example, a fire event was predicted before the harvest change due to a near-by fire event in the middle sequence in Figure 15. Finally, minor errors in prediction can occur when very small change regions are involved as the one shown at the bottom of Figure 15.

VII. Conclusion

Existing approaches to detect changes in forest land-cover work at a larger temporal scale and use hand-crafted features designed from the landscape attributes. Our proposed approach is capable of performing change analysis at a much finer temporal resolution and automatically learns strong features from



(a) (b) (c)

(d) (e) (f)

Fig. 13: (a-f) Sample results of the ground-truth change patterns and the change sequences predicted by our approach. In each plot, the top bar shows ground truth, and the bottom bar shows prediction from our approach.

the raw surface reflectance data. To achieve a finer temporal resolution, we perform data inpainting using the reliable data values and sparse coding. For change detection, our approach works on the object-level by identifying a candidate set of change regions using multi-resolution area profiles. We use both spatial and temporal contextual information in the deep CNN model which helps in making better predictions. Our method can precisely localize the change regions and predict their on/offset timings accurately within an error margin of 3 to 6 months. In future, the possibility of creating a large-scale annotated dataset will be investigated. This will enable the training of large-scale data driven models from scratch. Since interesting changes are scarce in practical settings, we will also investigate class-imbalanced learning of deep networks for change detection [69].

ACKNOWLEDGEMENTS

The authors would like to thank Geoscience Australia for providing the data and expert annotations. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. This work is supported by Data61/CSIRO, UWA through IPRS, and under the Australian Research Council's Discovery Projects funding scheme (project DP150104645).

REFERENCES

[1] ABARES. Australias state of the forests five-yearly report 2013. *Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES), Department of Agriculture, Australian Government*, 2013.

[2] Robert E Kennedy, Zhiqiang Yang, and Warren B Cohen. Detecting trends in forest disturbance and recovery using yearly landsat time series: 1. landtrendrtemporal segmentation algorithms. *Remote Sensing of Environment*, 114(12):2897–2910, 2010.

[3] Matthew C Hansen and Thomas R Loveland. A review of large area monitoring of land cover change using landsat data. *Remote sensing of Environment*, 122:66–74, 2012.

[4] Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, 80:91–106, 2013.

[5] Txomin Hermosilla, Michael A Wulder, Joanne C White, Nicholas C Coops, and Geordie W Hobart. Regional detection, characterization, and attribution of annual forest change from 1984 to 2012 using landsat-derived time-series metrics. *Remote Sensing of Environment*, 170:121–132, 2015.

[6] Ross S Lunetta, David M Johnson, John G Lyon, and Jill Crotwell. Impacts of imagery temporal frequency on land-cover change detection monitoring. *Remote Sensing of Environment*, 89(4):444–454, 2004.

[7] Salman H Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic shadow detection and removal from a single image. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):431–446, 2016.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[10] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Roberto Togneri, and Ferdous A Sohel. A discriminative representation of convolutional features for indoor scene recognition. *IEEE Transactions on Image Processing*, 25(7):3372–3383, 2016.

[11] George Xian, Collin Homer, and Joyce Fry. Updating the 2001 national land cover database land cover classification to 2006 by using landsat imagery change detection methods. *Remote Sensing of Environment*, 113(6):1133–1147, 2009.

[12] A Kawabata, K Ichii, and Y Yamaguchi. Global monitoring of interannual changes in vegetation activities using ndvi and its relationships to temperature and precipitation. *International Journal of Remote Sensing*, 22(7):1377–1382, 2001.

[13] Daniel Tomowski, Manfred Ehlers, and Sascha Klonus. Colour and texture based change detection for urban disaster analysis. In *Urban Remote Sensing Event (JURSE), 2011 Joint*, pages 329–332. IEEE, 2011.

[14] Chengquan Huang, Kuan Song, Sunghee Kim, John RG Townshend, Paul Davis, Jeffrey G Masek, and Samuel N Goward. Use of a dark object concept and support vector machines to automate forest cover change analysis. *Remote Sensing of Environment*, 112(3):970–985, 2008.

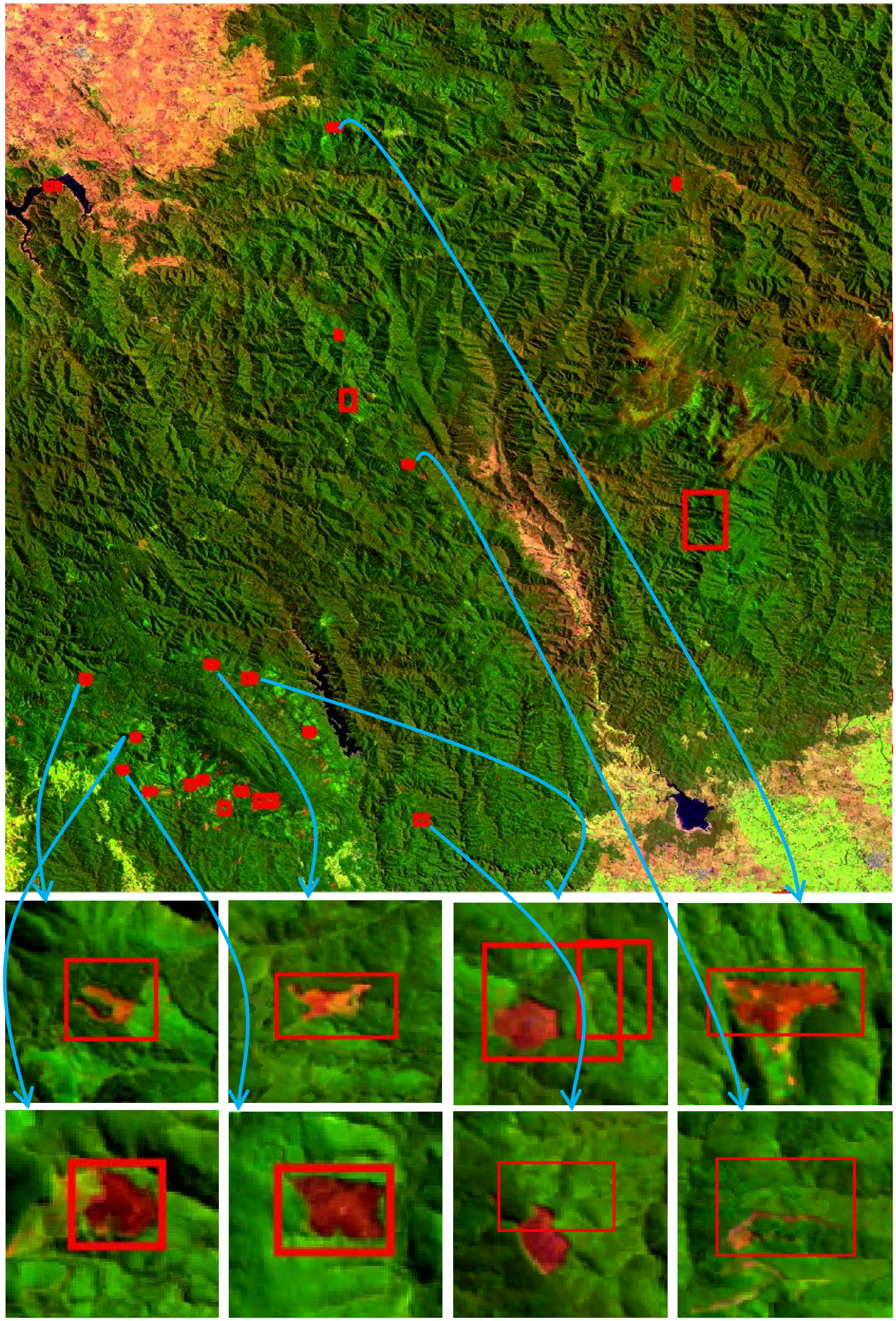
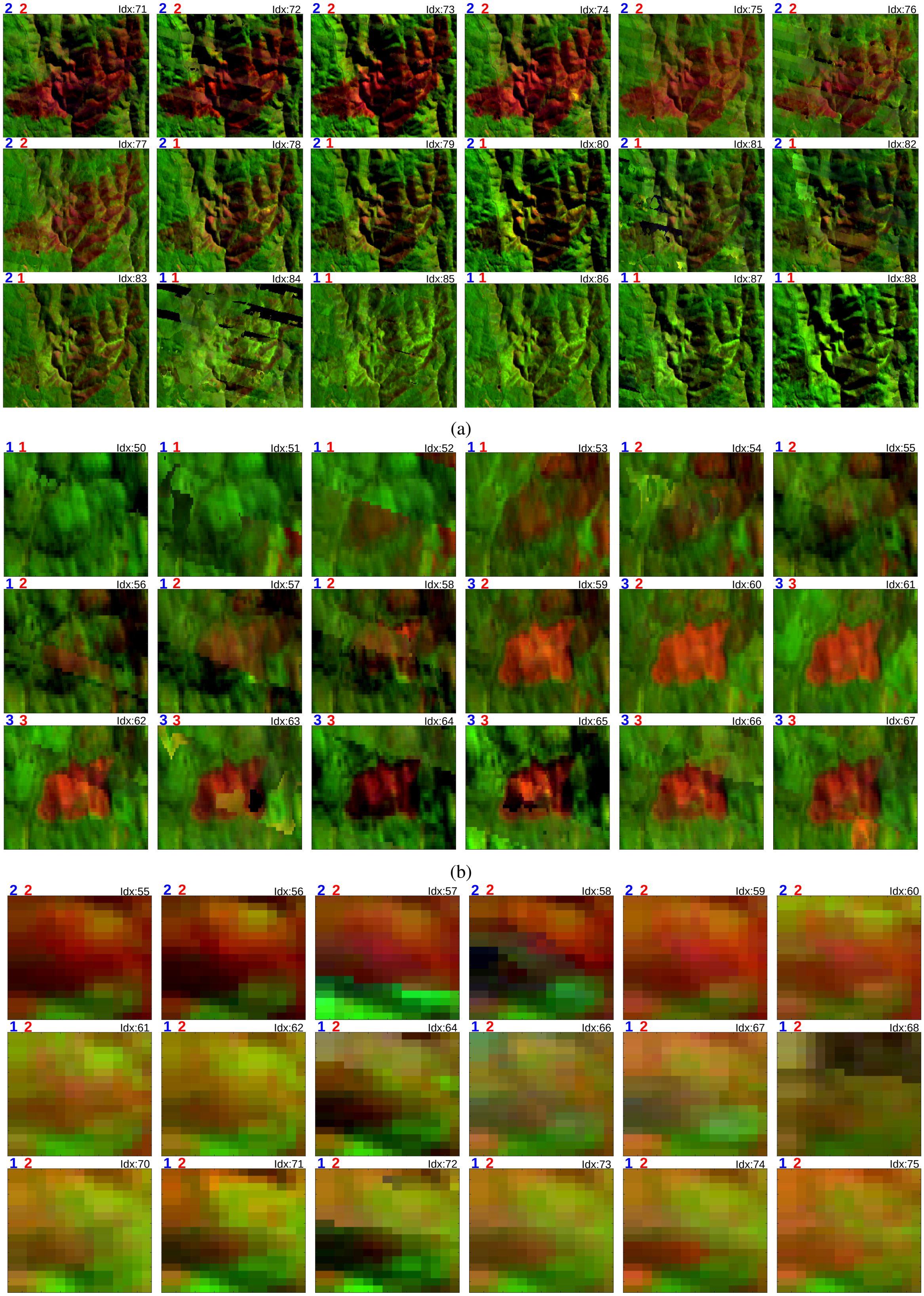


Fig. 14: This figure shows detection results on the complete image plane encompassing the forest area under investigation. We show some examples of change regions (bottom two rows) which were not labeled by the experts, yet our algorithm was successfully able to detect them.



(a)

(b)

(c)

Fig. 15: Three small portions of patch sequences are shown in the above figure. The ground-truth and the predicted change/no-change labels are shown on the top left corner in blue color and red color respectively. The digits 1, 2 and 3 on top-left represent no change, fire and harvest respectively.

[15] Gang Chen, Geoffrey J Hay, Luis MT Carvalho, and Michael A Wulder. Object-based change detection. *International Journal of Remote Sensing*, 33(14):4434–4457, 2012.

[16] Salman H Khan, Xuming He, Mohammed Bennamoun, Fatih Porikli, Ferdous Sohel, and Roberto Togneri. Learning deep structured network for weakly supervised change detection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2017.

[17] J-F Mas. Monitoring land-cover changes: a comparison of change detection techniques. *International journal of remote sensing*, 20(1):139–152, 1999.

[18] Pol Coppin, Inge Jonckheere, Kristiaan Nackaerts, Bart Muys, and Eric Lambin. Digital change detection methods in ecosystem monitoring: a review. *International journal of remote sensing*, 25(9):1565–1596, 2004.

[19] Eric A Lehmann, Jeremy F Wallace, Peter A Caccetta, Suzanne L Furby, and Katherine Zdunic. Forest cover trends from time series landsat data for the australian continent. *International Journal of Applied Earth Observation and Geoinformation*, 21:453–462, 2013.

[20] ML Nordberg and J Evertson. Vegetation index differencing and linear regression for change detection in a swedish mountain range using landsat tm® and etm+® imagery. *Land Degradation & Development*, 16(2):139–149, 2005.

[21] Chengquan Huang, Samuel N Goward, Jeffrey G Masek, Nancy Thomas, Zhiliang Zhu, and James E Vogelmann. An automated approach for reconstructing recent forest disturbance history using dense landsat time series stacks. *Remote Sensing of Environment*, 114(1):183–198, 2010.

[22] Kris Nackaerts, Krist Vaesen, Bart Muys, and Pol Coppin. Comparative performance of a modified change vector analysis in forest change detection. *International Journal of Remote Sensing*, 26(5):839–852, 2005.

[23] Anna Versluis and John Rogan. Mapping land-cover change in a haitian watershed using a combined spectral mixture analysis and classification tree procedure. *Geocarto International*, 25 (2):85–103, 2010.

[24] Bryan C Pijanowski, Snehal Pithadia, Bradley A Shellito, and Konstantinos Alexandridis. Calibrating a neural network-based urban change model for two metropolitan areas of the upper midwest of the united states. *International Journal of Geographical Information Science*, 19(2):197–215, 2005.

[25] Jungho Im and John R Jensen. A change detection model based on neighborhood correlation image analysis and decision tree classification. *Remote Sensing of Environment*, 99(3):326–340, 2005.

[26] Paul A Longley. Geographical information systems: will developments in urban remote sensing and gis lead to betterurban geography? *Progress in Human Geography*, 26(2):231–239, 2002.

[27] YH Araya and C Hergarten. A comparison of pixel and object-based land cover classification: a case study of the asmara region, eritrea. *WIT Transactions on the Built Environment, Geo-Environment and Landscape Evolution III*, 100, 2008.

[28] Antoine Lefebvre, Thomas Corpetti, and Laurence Hubert-Moy. Object-oriented approach and texture analysis for change detection in very high resolution images. In *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, volume 4, pages IV–663. IEEE, 2008.

[29] Tim De Chant and Maggi Kelly. Individual object change detection for monitoring the impact of a forest pathogen on a hardwood forest. *Photogrammetric Engineering & Remote Sensing*, 75(8):1005–1013, 2009.

[30] George Xian and Collin Homer. Updating the 2001 national land cover database impervious surface products to 2006 using landsat imagery change detection methods. *Remote Sensing of Environment*, 114(8):1676–1686, 2010.

[31] Zhe Zhu and Curtis E Woodcock. Object-based cloud and cloud shadow detection in landsat imagery. *Remote Sensing of Environment*, 118:83–94, 2012.

[32] Chao-Hung Lin, Po-Hung Tsai, Kang-Hua Lai, and Jyun-Yuan Chen. Cloud removal from multitemporal satellite images using information cloning. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(1):232–241, 2013.

[33] Bo Huang, Ying Li, Xiaoyu Han, Yuanzheng Cui, Wenbo Li, and Rongrong Li. Cloud removal from optical satellite imagery with sar imagery using sparse representation. *Geoscience and Remote Sensing Letters, IEEE*, 12(5):1046–1050, 2015.

[34] Abdolrassoul S Mahiny and Brian J Turner. A comparison of four common atmospheric correction methods. *Photogrammetric Engineering & Remote Sensing*, 73(4):361–368, 2007.

[35] Gong Jianya, Sui Haigang, Ma Guorui, and Zhou Qiming. A review of multi-temporal remote sensing data change detection algorithms. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37(B7):757–762, 2008.

[36] Chao-Hung Lin, Kang-Hua Lai, Zhi-Bin Chen, and Jyun-Yuan Chen. Patch-based information reconstruction of cloud-contaminated multitemporal images. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(1):163–174, 2014.

[37] Luca Lorenzi, Farid Melgani, and Grégoire Mercier. Inpainting strategies for reconstruction of missing data in vhr images. *Geoscience and Remote Sensing Letters, IEEE*, 8(5):914–918, 2011.

[38] Aldo Maalouf, Philippe Carré, Bertrand Augereau, and Christine Fernandez-Maloigne. A bandelet-based inpainting technique for clouds removal from remotely sensed images. *Geoscience and Remote Sensing, IEEE Transactions on*, 47(7): 2363–2371, 2009.

[39] Chuanrong Zhang, Weidong Li, and David J Travis. Restoration of clouded pixels in multispectral remotely sensed imagery with cokriging. *International Journal of Remote Sensing*, 30(9): 2173–2195, 2009.

[40] MJ Pringle, M Schmidt, and JS Muir. Geostatistical interpolation of slc-off landsat etm+ images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(6):654–664, 2009.

[41] David P Roy, Junchang Ju, Philip Lewis, Crystal Schaaf, Feng Gao, Matt Hansen, and Erik Lindquist. Multi-temporal modis–landsat data fusion for relative radiometric normalization, gap filling, and prediction of landsat data. *Remote Sensing of Environment*, 112(6):3112–3130, 2008.

[42] Richard R Irish. Landsat 7 automatic cloud cover assessment. In *AeroSense 2000*, pages 348–355. International Society for Optics and Photonics, 2000.

[43] Din-Chang Tseng, Hsiao-Ting Tseng, and Chun-Liang Chien. Automatic cloud removal from multi-temporal spot images. *Applied Mathematics and Computation*, 205(2):584–600, 2008.

[44] Quanjun Jiao, Wenfei Luo, Xue Liu, and Bing Zhang. Information reconstruction in the cloud removing area based on multi-temporal chris images. In *International Symposium on Multispectral Image Processing and Pattern Recognition*, pages 679029–679029. International Society for Optics and Photonics, 2007.

[45] EH Helmer and B Ruefenacht. Cloud-free satellite image mosaics with regression trees and histogram matching. *Photogrammetric Engineering & Remote Sensing*, 71(9):1079–1089, 2005.

[46] Benjamin W Martin and Ranga R Vatsavai. Evaluating fusion techniques for multisensor satellite image data. In *SPIE Defense, Security, and Sensing*, pages 87470J–87470J. International Society for Optics and Photonics, 2013.

[47] Huanfeng Shen, Liwen Huang, Liangpei Zhang, Penghai Wu, and Chao Zeng. Long-term and fine-scale satellite monitoring of the urban heat island effect by the fusion of multi-temporal and multi-sensor remote sensed data: A 26-year case study of the city of wuhan in china. *Remote Sensing of Environment*, 172:109–125, 2016.

[48] Patrick Griffiths, Patrick Hostert, Oliver Gruebner, and Sebastian van der Linden. Mapping megacity growth with multi-

sensor data. *Remote Sensing of Environment*, 114(2):426–439, 2010.

[49] Ronan Fablet and François Rousseau. Non-local super-resolution of missing data in multi-sensor observations of sea surface geophysical fields. In *IGARSS 2015: IEEE International Geoscience and Remote Sensing Symposium*, pages TUP–PJ, 2015.

[50] Kevin C Guay, Pieter SA Beck, Logan T Berner, Scott J Goetz, Alessandro Baccini, and Wolfgang Buermann. Vegetation productivity patterns at high northern latitudes: a multi-sensor satellite data assessment. *Global change biology*, 20(10):3147–3158, 2014.

[51] Otávio AB Penatti, Keiller Nogueira, and Jefersson A dos Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 44–51, 2015.

[52] Lionel Gueguen and Raffay Hamid. Large-scale damage detection using satellite imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1321–1328, 2015.

[53] Martin Längkvist, Andrey Kiselev, Marjan Alirezaie, and Amy Loutfi. Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sensing*, 8(4):329, 2016.

[54] Fan Hu, Gui-Song Xia, Jingwen Hu, and Liangpei Zhang. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, 7(11):14680–14707, 2015.

[55] Munawar Hayat, Salman H Khan, Mohammed Bennamoun, and Senjian An. A spatial layout and scale invariant feature representation for indoor scene classification. *IEEE Transactions on Image Processing*, 25(10):4829–4841, 2016.

[56] Jun Wang, Jingwei Song, Mingquan Chen, and Zhi Yang. Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine. *International Journal of Remote Sensing*, 36(12):3144–3169, 2015.

[57] Xueyun Chen, Shiming Xiang, Cheng-Lin Liu, and Chun-Hong Pan. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and remote sensing letters*, 11(10):1797–1801, 2014.

[58] Richard R Irish, John L Barker, Samuel N Goward, and Terry Arvidson. Characterization of the landsat-7 etm+ automated cloud-cover assessment (acca) algorithm. *Photogrammetric Engineering & Remote Sensing*, 72(10):1179–1188, 2006.

[59] Fuqin Li, David LB Jupp, Shanti Reddy, Leo Lymburner, Norman Mueller, Peter Tan, and Anisul Islam. An evaluation of the use of atmospheric and brdf correction to standardize landsat data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 3(3):257–270, 2010.

[60] Fuqin Li, David LB Jupp, Medhavy Thankappan, Leo Lymburner, Norman Mueller, Adam Lewis, and Alex Held. A physics-based atmospheric and brdf correction for landsat data over mountainous terrain. *Remote Sensing of Environment*, 124: 756–770, 2012.

[61] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[62] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[63] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision 2014*, pages 391–405. Springer, 2014.

[64] Piotr Dollár and C. Lawrence Zitnick. Structured forests for fast edge detection. In *Proceedings of the International Conference on Computer Vision*, 2013.

[65] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

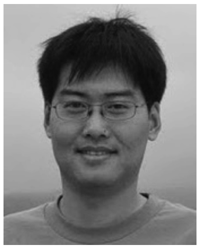
[66] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[67] Joao Carreira and Cristian Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (7):1312–1328, 2012.

[68] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.

[69] Salman H Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Cost sensitive learning of deep feature representations from imbalanced data. *arXiv preprint arXiv:1508.03422*, 2015.

Salman Khan (M'15) received the B.E. degree in electrical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2012, and the Ph.D. degree from The University of Western Australia (UWA), in 2016. He was a Visiting Researcher with National ICT Australia, CRL, during the year 2015. He is currently a Research Scientist with Data61 (CSIRO) and an Adjunct Lecturer with Australian National University (ANU) since 2016. His research interests include computer vision, pattern recognition and machine learning.



Xuming He received the Ph.D. degree in computer science from the University of Toronto, Canada, in 2008. He was an Adjunct Research Fellow with Australian National University from 2010 to 2016. He held a post-doctoral position with the University of California at Los Angeles, USA, from 2008 to 2010. He joined National ICT Australia and was a Senior Researcher from 2013 to 2016. He is currently an Associate Professor with the School of Information Science and Technology, ShanghaiTech University. His research interests include semantic image and video segmentation, 3-D scene understanding, visual motion analysis, and efficient inference and learning in structured models.

Fatih Porikli (M'99-F'13) received the Ph.D. degree from New York University in 2002. He was the Distinguished Research Scientist with Mitsubishi Electric Research Laboratories. He is currently a Professor with the Research School of Engineering, Australian National University and a Chief Scientist at the Global Media Technologies Lab at Huawei, Santa Clara. He has authored over 150 publications, coedited two books, and invented 66 patents. His research interests include computer vision, pattern recognition, manifold learning, image enhancement, robust and sparse optimization and online learning with commercial applications in video surveillance, car navigation, robotics, satellite, and medical systems. He was a recipient of the Research and Development 100 Scientist of the Year Award in 2006. He received five Best Paper Awards at premier IEEE conferences and five other professional prizes. He is serving as the Associate Editor of several journals for the past 12 years.

Mohammed Bennamoun received his M.Sc. degree in control theory from Queen's University, Kingston, Canada, and the Ph.D. degree in computer vision from Queen's University/Queensland University of Technology (QUT), Brisbane, Australia. He lectured Robotics at Queen's University and then joined QUT in 1993 as an Associate Lecturer. He is currently a Winthrop Professor and has been the Head of the School of Computer Science and Software Engineering, The University of Western Australia (UWA), Perth, Australia for five years (2007−2012). He has published over 300 journal and conference publications and secured highly competitive national grants from the Australian Research Council (ARC). His areas of interest include control theory, robotics, object recognition, artificial neural networks, signal/image processing, and computer vision.