Learning deep structured network for weakly supervised change detection

Salman Khan[1,2*], **Xuming He**[3†], **Fatih Porikli**[2], **Mohammed Bennamoun**[4]
Ferdous Sohel[5] **and Roberto Togneri**[4]
[1]Data61-CSIRO, Canberra, Australia
[2]Australian National University, Canberra, Australia
[3]ShanghaiTech University, Shanghai, China
[4]The University of Western Australia, Perth, Australia
[5]Murdoch University, Perth, Australia

Abstract

Conventional change detection methods require a large number of images to learn background models or depend on tedious pixel-level labeling by humans. In this paper, we present a weakly supervised approach that needs only image-level labels to simultaneously detect and localize changes in a pair of images. To this end, we employ a deep neural network with DAG topology to learn patterns of change from image-level labeled training data. On top of the initial CNN activations, we define a CRF model to incorporate the local differences and context with the dense connections between individual pixels. We apply a constrained mean-field algorithm to estimate the pixel-level labels, and use the estimated labels to update the parameters of the CNN in an iterative EM framework. This enables imposing global constraints on the observed foreground probability mass function. Our evaluations on four benchmark datasets demonstrate superior detection and localization performance.

1 Introduction

Identifying changes of interest in a given set of images is a fundamental task in computer vision with numerous applications in fault detection, disaster management, crop monitoring, visual surveillance, and scene analysis in general. When there are only two images available, existing approaches mostly resort to strong supervision, thus require large amounts of training data with accurate pixel-level annotations to perform pixel-level analysis. To comprehend the significant amount of effort needed for such a formidable task, we consider the example of CDnet-2014 [Wang *et al.*, 2014], which is the largest dataset for video based change detection. This dataset required manual annotations for $\sim$8 billion pixel locations. Although sophisticated methods have been investigated to reduce the human effort, e.g., by expert feedback in case of ambiguity [Jain and Grauman, 2013; Gueguen and Hamid, 2015], semi-automatic propagation of annotations [Badrinarayanan *et al.*, 2013], and point-wise supervision [Russakovsky *et al.*, 2015], acquisition of accurate and dense pixel-wise labels still remains a daunting task [Lin *et al.*, 2014; Song *et al.*, 2015].

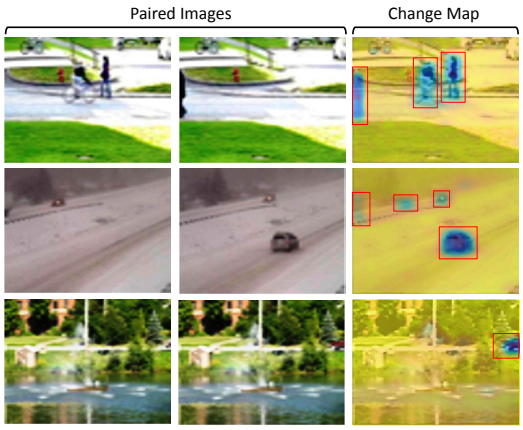

Figure 1: Change Detection in a pair of images. Our approach uses only image-level labels to detect and localize changed pixels. **Left:** pair of images. **Right-most** our change localization results (blue denotes a high change region, enclosed in a red box for clarity). Note that the paired images have rich backgrounds with different motion patterns (e.g., fountain in the *third row*) and subtle changes (e.g., small vehicles in the *second* row), which make the detection task very challenging.

Here, we address the problem of change detection within a pair of images and present a solution that uses only *image-level* labels to detect and localize changed regions (Fig. 1). Our method drastically reduces the effort required to collect annotations and provides an alternative to video change detection that requires a large number of consecutive frames to model the background scene. In many real-world applications, a continuous stream of images may not be always available due to a number of reasons such as challenging acquisition conditions, limited data storage, latency in processing, and long intervals before changes happen. For example, the analysis of aerial images for change detection, in particular for damage detection, often is formulated for a pair of images acquired at different times. Other examples where only a pair

*Corresponding author: salman.khan@anu.edu.au
†This work was done when the author was at Data61/ANU.

images might be available include structural defect identification, face rejuvenation tracking, and updating city street-view models.

Our algorithm jointly predicts the image-level change label and a segmentation map indicating the location of changes for a given pair of images. The central component of our method is a novel two-stream deep network model with structured outputs (Sec. 2). This model operates on a pair of images and does not need the images to be registered precisely. It can be trained with only weak image-level labels (Sec. 4.2). The network has a Directed Acyclic Graph (DAG) architecture where the initial layers are shared, while the latter part splits into two branches that make separate (but coupled) predictions for change detection and localization. In this manner, our deep network is different from the popular single-stream convolutional neural networks (CNN) for object classification [Khan *et al.*, 2015], detection [Girshick *et al.*, 2014; Khan *et al.*, 2016] and semantic labeling tasks [Papandreou *et al.*, 2015; Long *et al.*, 2015; Pinheiro and Collobert, 2015].

In order to jointly predict the image-level and pixel-level labels, we introduce a constrained mean-field inference algorithm (Sec. 2.3), that employs a factorizable approximate posterior distribution with global linear constraints. Using a global constraint on the foreground (changed pixels) probability mass function, we suppress the bias towards the background (no-change labeled pixels) and encourage the assignment of change labels to nonidentical regions. Such global constraints enable us to derive an efficient mean-field inference procedure, while eliminating the need of approximate biases [Papandreou *et al.*, 2015] and object based priors [Pinheiro and Collobert, 2015; Russakovsky *et al.*, 2015]. Furthermore, based on the novel inference algorithm, we apply a variational Expectation-Maximization (EM) learning algorithm that maximizes the lower bound of the log-likelihood of image-level labels. We extensively evaluate our approach on three publicly available datasets (CDnet-2014, PCD-2015 and AICD-2012) and a custom built satellite image dataset (GASI-2015) (Sec. 4.2). Our experimental results demonstrate that the proposed approach outperforms the state-of-the-art by a large margin (Sec. 4.3). The key contributions of our work include:

- To the best of our knowledge, this is the first work to address the weakly supervised change detection problem.
- Our proposed CNN model jointly detects and localizes changes in image pairs.
- We present a modified mean-field algorithm with additional constraints to efficiently localize changes.
- We introduce a new satellite image dataset (GASI-2015) for change detection. Furthermore, we perform a rigorous evaluation on three other relevant datasets.

2 Two-stream CNNs for Change Localization

We address the problem of joint change detection and localization with only image-level weak supervision. To this end, we propose a two-stream deep convolutional neural network model with structured outputs, which can be learned with weakly labeled image pairs. We describe our model next.

2.1 Model Overview

Given a pair of input data, which can be images or (short) video clips, our goal is to predict the categories of change events in the data pair and localize the change more precisely at the pixel level. For simplicity, we focus on the image pair scenario in the following and video clips can be processed in a similar manner.

Specifically, let each input consists of a pair of images, $\mathbf{x} = \{\mathbf{I}^1, \mathbf{I}^2\}$. We associate an image-level output label vector $\mathbf{y} = \{y_1, y_2\}$ to indicate the occurred change events i.e., change, no-change and $y_1, y_2 \in \{0, 1\}$. It is important to note here that the no-change category (i.e., $\mathbf{y} = \mathbf{0}$) refers to the static-background, irrelevant changes and the dynamic background change patterns while those change categories refer to changes of interest. In order to localize the change events at the pixel level, we introduce a set of binary variables $\mathbf{h}$ to denote the labels of individual pixel locations for each image pair $\mathbf{x}$. Assume the image has m pixel locations, $\mathbf{h} = \{h_1, \cdots, h_m\} \in \{0, 1\}^m$.

We formulate the change detection and localization problem as the joint prediction of its image-level and pixel-level change variables. To achieve this, we consider a deep structured model that defines a joint probabilistic model on $\mathbf{y}$ and $\mathbf{h}$ as $P(\mathbf{y}, \mathbf{h}|\mathbf{x}; \theta) = \frac{1}{Z(\mathbf{x})} \exp(-E(\mathbf{y}, \mathbf{h}|\mathbf{x}; \theta))$, where the Gibbs energy is defined as:

$$E(\mathbf{y}, \mathbf{h}|\mathbf{x}; \theta) = \Phi_l(\mathbf{y}|\mathbf{x}; \theta_l) + \Phi_u(\mathbf{h}|\mathbf{x}; \theta_u) + \Phi_p(\mathbf{h}, \mathbf{y}|\mathbf{x}, \theta_p), \quad (1)$$

where $\Phi_l(\mathbf{y}|\mathbf{x}; \theta_l)$ is the unary term for image-level label $\mathbf{y}$, modeled by a CNN with parameter θ_l, and $\Phi_u(\mathbf{h}|\mathbf{x}; \theta_u)$ is the unary term for pixel-level labels, modeled by a Fully Convolutional Network (FCN) with parameter θ_u. The pairwise energy $\Phi_p(\mathbf{h}, \mathbf{y})$ consists of two terms, ψ_p and ψ_u, which enforces the spatial smoothness of pixel-level labels and captures the coupling between image- and pixel-level labels, respectively. The joint prediction can be formulated as inferring the MAP estimation of the model distribution,

$$\mathbf{y}^*, \mathbf{h}^* = \arg\max_{\mathbf{y}, \mathbf{h}} P(\mathbf{y}, \mathbf{h}|\mathbf{x}; \theta). \quad (2)$$

A graphical illustration of the model is shown in Fig. 2.

2.2 Deep Network Architecture

We build the deep structured model by first introducing a two-stream deep CNN for the unary terms as shown in Fig. 3. The underlying architecture of the network is similar to the VGG-net (configuration-D, the winner of the classification and localization challenge, ILSVRC'14) [Simonyan and Zisserman, 2014] but with several major differences. Most importantly, the network operates on multichannel inputs (6 channels for paired color images) and divides into two branches after the fourth pooling layer ($P4$). From our initial experiments (consistent with [Zagoruyko and Komodakis, 2015]), a multi-channel network performs better than a traditionally used Siamese network for paired images. The two branches compute the probability of the image-level and pixel-level labels, and therefore will be called as the classification and the segmentation branch, respectively. The segmentation branch in our architecture is similar to FCN-VGG16-16s network

[Long *et al.*, 2015] which demonstrated state-of-the-art performance on the Pascal VOC segmentation dataset. The initial shared layers in our architecture combine the initial (essentially similar) portions of VGG and FCN networks, which results in a significant decrease in trainable parameters without any drop in performance. We now describe the details of the two branches of the network architecture.

Image-level change unary energy: The classification branch predicts the image-level label probability and has more layers to collapse the filter responses from the initial layers. Specifically, the classification branch output of the CNN architecture models $\Phi(y|\mathbf{x};\theta_l)$, predicting the image-level change energy as:

$$\Phi_l(y|\mathbf{x};\theta_l) = -F_{l-cnn}(\mathbf{x}; W_s, W_y), \tag{3}$$

where F_{l-cnn} is the deep network feature before the final softmax operator, W_s are the weight parameters shared with the segmentation branch, and W_y are the weight parameters for the classification branch only.

Pixel-level change unary energy: The segmentation branch generates a down-sampled coarse segmentation map (of size $o_{sz} \times o_{sz}$) for each change category. After shared layers, the branch has three fully connected layers, which are implemented as convolution layers as in the FCN [Long *et al.*, 2015]. Formally, the segmentation branch of the CNN model generates the pixel-level change label energy as follows,

$$\Phi_u(\mathbf{h}|\mathbf{x};\theta_u) = \sum_{j=1}^{m} \Phi(h_j|\mathbf{x}), \Phi(h_j|\mathbf{x}) = -F_{u-cnn}(h_j, \mathbf{x}; W_s, W_f),$$

where F_{u-cnn} denotes the segmentation branch scores of the CNN architecture before the soft-max operator and W_f are the weights for the fully connected layers.

We now describe the pairwise energy $\Phi(\mathbf{h}, y|\mathbf{x}, \theta_p)$ that encodes the compatibility relations between the image-level and the pixel-level variables as well as the spatial smoothness. Specifically, on the top of the fully connected layers, we add a densely connected Conditional Random Field (CRF) to impose the spatial smoothness of the pixel labeling. Unlike the previous models [Papandreou *et al.*, 2015], our dense CRF depends on the output label of the classification branch, and thus couples the image-level and pixel-level prediction.

Formally, we define the compatibility relations between the output variables y and $\mathbf{h}$ by the following energy functions,

$$\Phi_p(\mathbf{h}, y|\mathbf{x}, \theta_p) = \sum_j \psi_u(h_j, y, \mathbf{x}_j) + \sum_{j<k} \psi_p(h_j, h_k, \mathbf{f}_j, \mathbf{f}_k), \tag{4}$$

where ψ_u enforces all hidden variables to be zero if the category label predicts no-change and encourages h_j to take a change label otherwise:

$$\psi_u(h_j, y, \mathbf{x}_j) = [\![y = 0]\!][\![h_j = 0]\!] + [\![y = 1]\!]\big(1 + e^{-\gamma\delta(\mathbf{x}_j)}[\![h_j = 0]\!]\big),$$

where γ is a weight parameter and $\delta(\mathbf{x}_j)$ is the color difference between two images at pixel j. The fully-connected pairwise term ψ_p defines the smoothing term between the latent variables $\mathbf{h}$ given input features $\mathbf{f}_j, \mathbf{f}_k$. These energies have a functional form of the weighted Potts model in which the weight is defined using Gaussian kernels of [Krähenbühl and Koltun, 2011]:

$$\psi_p(h_j, h_k, \mathbf{f}_j, \mathbf{f}_k) = (\alpha_{ap} k_{ap} + \alpha_{sm} k_{sm})\mu(h_j, h_k), \tag{5}$$

where α_{ap}, α_{sm} are the kernel weights, $\mu(h_j, h_k)$ is the Potts compatibility while $k_{ap}(\mathbf{f}_j, \mathbf{f}_k)$, $k_{sm}(\mathbf{f}_j, \mathbf{f}_k)$ are the appearance and smoothness kernels [Krähenbühl and Koltun, 2011].

2.3 Model Inference for Change Localization

Given the two-stream CNN+CRF model, we predict the image and pixel-level change labels by inferring the MAP estimation of the joint probability model in Eq. (1). In order to compute the most likely configuration efficiently, we adopt a sequential prediction approach that first infers the image-level change label followed by the pixel-level change mask inference. Specifically, we compute the change label prediction approximately as follows,

$$y^* = \arg\min_y \Phi_l(y|\mathbf{x};\theta_l), \mathbf{h}^* = \arg\min_{\mathbf{h}} \Phi_u(\mathbf{h}|\mathbf{x};\theta_u) + \Phi_p(\mathbf{h}, y^*|\mathbf{x};\theta_p).$$

This prioritized inference procedure allows us to compute the (more reliable) image-level label first and to run an efficient mean-field inference for the pixel-level labels only once[1].

We now derive a constrained mean-field inference algorithm for inferring the pixel-level change labeling $\mathbf{h}$. We note that the efficient mean-field algorithm [Krähenbühl and Koltun, 2011] usually leads to an over-smoothing of the pixel-level labeling and assigns most of the pixels to the 'no-change' class. In this work, we incorporate an additional global constraint on the proportion of 'change' label values in the image. Unlike previous methods (e.g., [Papandreou *et al.*, 2015; Pinheiro and Collobert, 2015; Russakovsky *et al.*, 2015]), we enforce such constraints on the approximate probability family which allows us to derive an efficient modified mean-field procedure.

Formally, we assume the foreground label proportion to be τ, which is fixed during training by cross-validation. For each test image pair, we find K closely matching pairs from the training set using a KNN search and average their foreground label proportion to estimate τ (details in Sec. 4.3). To enforce the proportion constraint, we introduce the following factorized approximate probability family with a global constraint:

$$Q(\mathbf{h}|\mathbf{x}, y^*) = \textstyle\prod_j \mathbf{q}_j(h_j), \text{ and } \sum_j [0,1]\mathbf{q}_j = \tau, \tag{6}$$

where, $\mathbf{q}_j = [\mathbf{q}_j(0), \mathbf{q}_j(1)]^T$ and the constraint implies that the overall foreground probability mass $\sum_j \mathbf{q}_j(1)$ is τ. Following [Krähenbühl and Koltun, 2013], we minimize the approximate KL-divergence,

$$\begin{aligned} D_{Q||P} =& \textstyle\sum_j \left(\mathbf{q}_j^T \log \mathbf{q}_j + \mathbf{q}_j^T \mathbf{u}_j\right) + \frac{1}{2} \sum_{j,k} \mathbf{q}_j^T \Psi_{jk} \mathbf{q}_k + C, \\ \text{with} \quad & \mathbf{1}^T \mathbf{q}_j = 1, \textstyle\sum_j [0,1]\mathbf{q}_j = \tau, \end{aligned} \tag{7}$$

where $\mathbf{u}_j$ is the unary term vector (including $P(h_j|\mathbf{x})$ and $\psi_u(h_j, y^*, \mathbf{x}_j)$) and Ψ_{jk} is the compatibility matrix computed from ψ_p, and $C = \log Z$ is the log partition function. We use the CCCP algorithm [Yuille and Rangarajan, 2003] to minimize $D_{Q||P}$ iteratively.

[1] In general, we note that we can compute the MAP estimation jointly by enumerating y's values and running mean-field inference multiple times, which is less efficient.

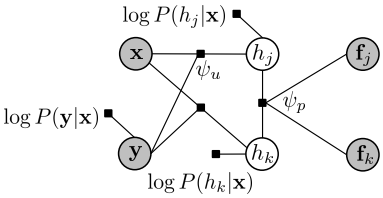

Figure 2: **Factor graph representation** of the weakly supervised change detection model. The shaded and non-shaded nodes represent the observable and hidden variables, respectively. The dense connections between hidden nodes are not shown here for clarity.

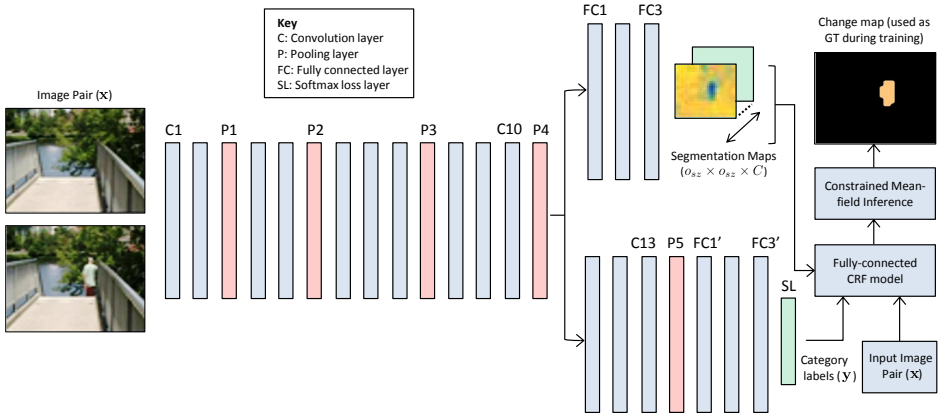

Figure 3: **CNN architecture:** The network operates on paired images and divides in two branches after the fourth pooling layer ($P4$). The classification branch (lower) is trained using the image-level labels (y). The hidden variables h_j are estimated using the constrained mean-field algorithm, which are iteratively used to update the CNN parameters.

3 EM Learning with Weak Supervision

We now consider a weakly supervised learning approach to estimate the parameters of the two-stream CNN+CRF model (Sec. 2). In particular, as the labeling of the pixel-level change pattern is tedious and impractical, we assume only image-level change annotations are available, which can be obtained with much less effort. Let us denote the dataset $\mathcal{D}$ comprising of N labeled image pairs: $\mathcal{D} = \{\mathbf{x}^n, y^n\}^{1\times N}$.

The learning objective is to maximize the log conditional likelihood and we consider a variational mean-field energy lower bound as follows,

$$\sum_n \log P(y^n|\mathbf{x}^n;\theta) \geq \sum_n \sum_{\mathbf{h}^n} Q(\mathbf{h}^n|y^n,\mathbf{x}^n) \log \frac{P(y^n,\mathbf{h}^n|\mathbf{x}^n;\theta)}{Q(\mathbf{h}^n|y^n,\mathbf{x}^n)}$$
$$= E_Q[\log P(y^n,\mathbf{h}^n|\mathbf{x}^n;\theta)] + H(Q(\mathbf{h}^n|\mathbf{x}^n,y^n)),$$

where, $E_Q[\cdot]$ and $H(\cdot)$ denote the expected value and the entropy function respectively, and $Q(\mathbf{h}^n|\mathbf{x}^n,y^n)$ is an approximate posterior probability factorizing over $\{h_j^n\}$ as defined in Eq. (6). In other words, the posterior probability can be expressed as the product of independent marginals: $Q(\mathbf{h}^n|\mathbf{x}^n,y^n) = \prod_j q_j^n(h_j^n)$. We then derive a variational expectation-maximization (EM) algorithm for learning our two-stream CNN+CRF in the following, which alternately maximizes the objective function above.

3.1 Mean-field E Step

We update the approximate Q function by maximizing the objective w.r.t the Q function given the model parameter θ from the previous iteration. Note that given the model structure, this leads to a mean-field updating equation to compute $q(h_j^n)$. The updating equation requires message passing between all the h_j and h_k, which is computationally expensive. Efficient message passing is achieved using the high dimensional Gaussian filtering by considering the permutohedral lattice structure [Adams *et al.*, 2010].

Given the approximate posterior marginals, we can compute the (approximate) most likely configuration of the latent variables $\mathbf{h}^n$,

$$h_j^{n*} \leftarrow \underset{h_j^n}{\operatorname{argmax}} \prod_{j=1}^{m} q(h_j^n|\mathbf{x}^n,y^n). \quad (8)$$

The marginal mode $\mathbf{h}^{n*}$ will be used in the M step for the CNN+CRF learning.

3.2 M Step for CNN+CRF Training

Once we have the posterior marginal distribution $q(h_j^n)$ and its mode, we update the model parameters θ with the posterior mode configuration $\{\mathbf{h}^{n*}\}$ and ground-truth $\{y^n\}$. Specifically, we treat them as the ground-truth for the pixel and image-level labels, and learn the two-stream deep CNN+CRF in a stage-wise manner. Our stage-wise learning first estimates the parameters in the unary terms, i.e., the two deep CNNs, and then validates the parameters in the pairwise term. This strategy is similar to the piece-wise learning in the CRF literature.

We first use back-propagation to train the two branches of the deep CNN separately with the corresponding training data. More precisely, the averaged gradient from two streams is back-propagated to update the shared parameters (W_s), while the individual gradients are computed using y^n and $\mathbf{h}^{n*}$ as ground-truths to update W_y and W_f for the classification and segmentation branches, respectively. Concretely, the model parameters are updated to maximize the data likelihood as follows,

$$W_s^* \leftarrow \underset{W_s}{\operatorname{argmax}} \sum_n \Big(\log P(y^n|\mathbf{x}^n;\theta_l) + \log P(\mathbf{h}^{n*}|\mathbf{x}^n;\theta_u)\Big),$$
$$W_y^* \leftarrow \underset{W_y}{\operatorname{argmax}} \sum_n \log P(y^n|\mathbf{x}^n;\theta_l),$$
$$W_f^* \leftarrow \underset{W_f}{\operatorname{argmax}} \sum_n \log P(\mathbf{h}^{n*}|\mathbf{x}^n;\theta_u).$$

After the two-stream deep network component is trained, we estimate the parameters θ_p in Eq. (4) by cross-validation.

The overall EM procedure starts with an M step with an initial value of $\mathbf{h}^n$. We assume the initial hidden variable states ($\mathbf{h}_0^n$) to be consistent with the image-level labels: $\mathbf{h}_0^n = y^n$.

The model parameters are fine-tuned by training the two-stream CNN+CRF with those initial labels. This is important because the CNN is pre-trained for object recognition on ImageNet and therefore the estimation of change regions in the initial E-step does not generate reasonable ground-truths.

4 Experiments

4.1 CNN Implementation

The network weights are initialized from a pre-trained VGG network (on ImageNet). The network splits into two portions after the fourth pooling layer. As we need a coarse segmentation map (32×32) at the output of the segmentation branch, enlarged paired images of size 512×512 are fed to the CNN. Moreover, the convolution filter size in FC1 (segmentation branch) is kept to 1×1 (in contrast to a 7×7 filter size in FC1$'$) to avoid the additional decrease in resolution of the 32×32 output map.

The unary energies of our CRF model are defined using the CNN activations, while the Gaussian edge potentials proposed by [Krähenbühl and Koltun, 2011] are used as pairwise terms. Note that changes of interest can occur in any of the two paired images, and therefore it is not desirable to remain restricted to the detection of changes in only one of the images (w.r.t the other image). For this purpose, the ground-truth with which we compare our final segmentation results include the changes in both images (see Fig. 4). During the mean-field inference step, we find the segmentation map of both images using their respective edge potentials. Subsequently, the two output maps are combined to get the final estimate of hidden variables. The resulting segmentation map is used as ground-truth during the CNN training (M step).

4.2 Datasets and Protocols

We evaluate our method on the following four datasets. All of them include pixel level change ground truth, from which we derive the image-level annotations for weakly supervised learning. The pixel-level labels are not used in the training of our deep network.

CDnet 2014 Dataset: The original video database consists of 53 videos with frame-by-frame ground-truths available for $\sim 90,000$ frames in specified regions-of-interests (ROIs). Various types of changes (e.g., shadow, object motion and motion blur) under different conditions (e.g., challenging weather, air turbulence and dynamic background) are included in this database [Wang *et al.*, 2014]. It is also important to note that the paired images are not registered and therefore background can change across paired images [Wang *et al.*, 2014]. A total of $91,595$ distinct image pairs are generated at random from the video sequences. In each pair, both images belong to the same video but they are captured at different time instances.

AICD 2012 Dataset: Aerial Image Change Detection (AICD) dataset [Bourdis *et al.*, 2011] consists of 1000 pairs of large sized images (800×600). It is a synthetic dataset in which the images are generated using a realistic rendering engine of a computer game (Battle Station 2). A total of 100 scenes are included in this dataset containing several real-world objects including buildings, trees and vehicles. The scenes are generated under varying conditions with significant changes in viewpoint, shadows and types of changes. Because the change regions are very small in satellite/aerial images, we work on the patch level and extract 48 patches of size 122×122 from each image with minimal overlap. This provides a total of $24,000$ paired images, facilitating the training of a model with a large number of parameters.

GASI 2015 Dataset: Geoscience Australia Satellite Image (GASI) dataset is a custom built dataset based on the changes occurred during $1999-2015$ in a $\sim 100 \times 100\, km^2$ area in the east of city of Melbourne in Victoria, Australia [Khan *et al.*, 2017]. For each region of interest, we have a time lapse sequence (between 1999-2015) of surface reflectance data and the corresponding pixel quality maps. Due to the severe artifacts caused due to clouds and band saturation, the modelling of the temporal trends is very challenging. In contrast, the acquisition of paired images captured at different times is much easier.

The annotations for two types of changes are provided in the GASI dataset namely: fire and harvests. We generate pairs of image patches for 67 distinct regions of interest which were identified by experts. In total, ~ 300 pairs are generated for each region of interest which makes a total of $\sim 20,000$ pairs. Since the raw data contains artifacts, we improved it's quality by filling data across different time instances.

There exists a large disparity among the sizes of change regions in the GASI dataset. For very large sized regions, we cropped the region bounded by a tightly fit bounding box. For small regions (mostly changes due to forest harvesting) with area $< 5\%$ of the total image area, we crop a bounding box with dimensions equal to three times that of a tightly bounded box. Since, there are large variations between the size of changes in the identified change regions, we converted all the regions to a uniform size of 224×224 to get a consistent segmentation map.

PCD 2015 Dataset: Panoramic change detection dataset [Sakurada and Okatani, 2015] consists of 200 pairs of panoramic images of street scenes and tsunami-hit areas. The image size is 224×1024, from which we extract 122×122 patches with a minimal overlap for training and testing. This gives us a total of $3,600$ pairs (18/panoramic image). It is important to mention that the two images are not perfectly registered. As a result, there are temporal differences in camera viewpoints, illumination and acquisition conditions.

4.3 Results

The change detection results of our approach on the four CD datasets are shown in Table 1. As a baseline, we only consider the classification branch of the CNN network initialized with the pre-trained VGG-net (configuration D, see $2-3$ columns of Table 1). Paired images are fed to this network architecture and 4096 dimensional feature vectors are extracted from the FC2$'$ layer. A linear SVM classifier is then trained for classification using the lib-linear package [Fan *et al.*, 2008]. On both datasets, the average precision (AP) and the overall

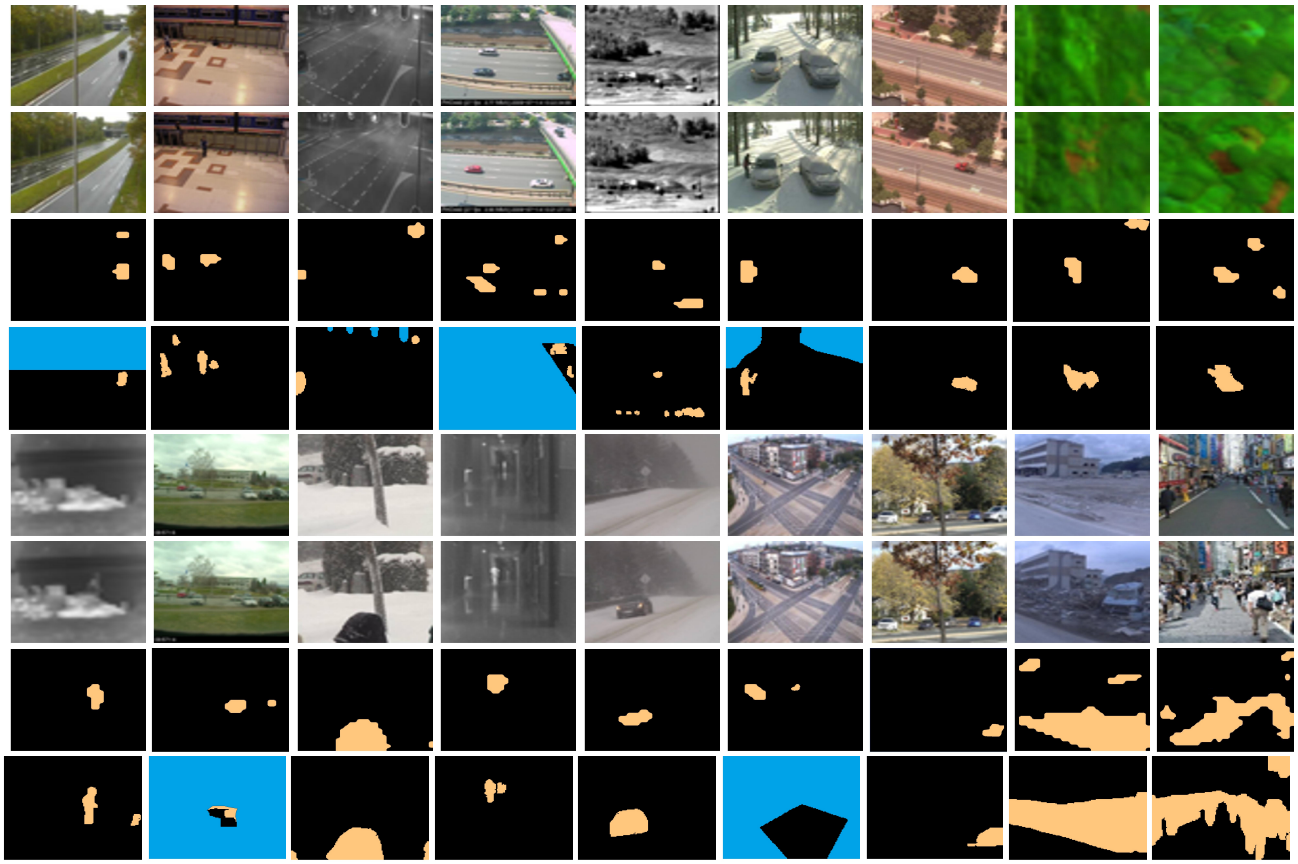

Figure 4: **Qualitative results on the CDnet-2014, GASI-2015 and PCD-2015 datasets:** Rows $1-2$ and $5-6$ show image pairs, our results are shown in rows 3 and 7 and the ground-truths are shown in rows 4 and 8. No-change pixels are shown in *black*, regions outside the ROIs are shown in *sky-blue* while the changes are shown in *coral-pink*.

Dataset	Classification Branch		Fine-tuned Classification Branch		This Paper	
	AP	Acc.	AP	Acc.	AP	Acc.
CDnet-2014	88.8	92.0	94.0	96.6	**95.6**	**98.7**
AICD-2012	92.7	95.4	95.7	98.0	**97.3**	**99.1**
GASI-2015	80.3	82.5	83.0	84.4	**85.6**	**86.5**
PCD-2015	67.1	73.5	72.9	81.1	**74.9**	**84.2**

Table 1: **Detection results** in terms of average precision (%) and overall accuracy (%) are listed above. Our approach clearly outperforms the baseline networks with only the classification branch.

accuracy of our approach was significantly higher than that of the baseline procedure (specifically 6.8% and 5.3% boost in AP for the CDnet and GASI datasets, respectively). As a stronger baseline, we also report performance of the network when only the fine-tuned classification branch was used (columns $4-5$, Table 1). We note that our full model outperformed the results from the fine-tuned classification branch.

We report the segmentation performance of our approach in Table 2 in terms of the mean intersection over union (mIOU) score. To compare our change localization results, we report four baseline procedures. Specifically, we compare against random segmentation masks (2^{nd} column-RS), thresholding applied to a difference map obtained from the pair of images (3^{rd} column-DT), thresholding applied to the output from a pre-trained network (weights initialized for segmentation branch using VGG-net, 4^{th} column-PN), thresholding applied to the output from the fine-tuned network (5^{th} column-Th.) and the graph-cuts inference [Boykov *et al.*, 2001] using CNN outputs as unaries and contrast based pairwise potentials with a Potts model (6^{th} column-GC). We note that random segmentation provides a lower baseline, while our results after training with ground-truths (shown in last column, Table 2) sets an upper bar on the performance. Another important trend is that the thresholding approach and graph-cuts performances do not differ by a large margin. However, our weakly supervised approach was able to achieve significantly higher mIOU scores due to the additional potentials and constraints (Sec. 2).

We also report segmentation results on two additional baselines which use cardinality based pattern potentials (Table 3). These baselines include the higher order potential (HOP) based dense and grid CRF models of Vineet *et al.* [Vineet *et al.*, 2014] and Kohli *et al.* [Kohli *et al.*, 2007] respectively. For both these baselines, we define HOPs on segments generated using mean-shift segmentation. Due to absence of pixel level supervision, we use the parameters from [Kohli *et al.*, 2007]. We note that the dense CRF model with P^n HOP [Vineet *et al.*, 2014] performs better than the grid CRF model [Kohli *et al.*, 2007], however our deep structured prediction

Dataset	Baseline Approaches (mIOU-%)					This Paper	Fully-supervised
	RS	DT	PN	Th.	GC[Boykov *et al.*, 2001]	(mIOU-%)	(mIOU-%)
CDnet-2014	16.4	36.8	37.4	35.9	37.1	**46.2**	59.2
AICD-2012	16.8	55.0	48.1	59.5	60.7	**64.9**	71.0
GASI-2015	18.3	40.5	40.7	41.6	42.2	**55.3**	62.4
PCD-2015	16.5	41.7	39.3	35.9	39.5	**47.7**	58.8

Table 2: **Segmentation Results** and comparisons with baseline methods. Note that all results (except the last column) are reported for the weakly-supervised setting.

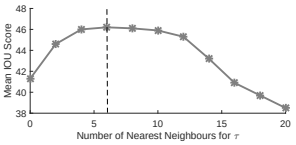

Figure 5: **Sensitivity analysis** on the number of nearest neighbours used to estimate τ (CDnet-2014).

Method	Dense CRF + P^n HOP [Vineet *et al.*, 2014]	Grid CRF + P^n HOP [Kohli *et al.*, 2007]	This Paper
mIOU%	42.0	38.3	46.2

Table 3: Comparisons for **segmentation performance** with methods using cardinality potentials on the CDnet-2014.

Method	Segmentation Results (mIOU)
with only segmentation branch	40.7
w/o CD fine-tuning	37.4
w/o difference term	41.5
w/o proportion constraint	41.3

Table 4: **Ablative Analysis** on the CDnet-2014 Dataset. Change localization performance decreases without our full model. Both the CNN architecture and the proposed CRF model contribute towards the accurate change detection.

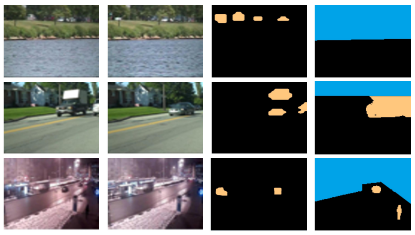

Figure 6: **Error Cases:** We present example cases, for which the ground-truths didn't exactly match with the generated results.

Normalized τ	0.1	0.2	0.3	0.4	0.5	0.6
mIOU% (CDnet-14)	30.8	28.4	36.7	41.1	32.5	25.4

Table 5: **Segmentation performance** for different τ values.

model outperforms both these strong baselines by a fair margin of $\sim 4-8\%$ in terms of mIOU score.

The qualitative results for change localization on the CDnet-2014, GASI-2015 and PCD-2015 datasets are shown in Fig. 4. The proposed approach performed well in localizing small as well as large sized changes (e.g., 1^{st} col, Fig. 4). Moreover, it showed good results for images acquired in varying conditions (e.g., night, snow, rainfall, dynamic background) and with different capturing devices (e.g., thermal camera, PTZ). For the CDnet-2014 dataset, it is interesting to note that our method localized several changes in the regions outside the ROIs (shown in *blue* color in the ground-truth). Similarly, the qualitative results indicate the good performance of our method for satellite image based change detection.

We performed an ablative study on the CDnet-2014 dataset for the change segmentation task (Table 4). The experimental results show that the localisation performance decreases without the feedback from the classification branch (whose predictions are more accurate). Moreover, since the pre-trained network is not trained to detect changes from multichannel inputs, the performance is considerably lower than that of the fine-tuned network. The difference term in the unary potential of the dense CRF and the global proportion constraint on the foreground probability mass also contributes a fair share in the final mIOU score.

During test, we use KNN to estimate an image-adaptive τ, which gives better estimate of foreground proportion and performance. We perform the KNN search using Euclidean distance on the features from the FC1 layer of the CNN model (classification branch). We use the fast approximate KNN search method based on KD-tree which has an average complexity of at most O(log(n)). Note that for the training, we validate a fixed-value τ, which is faster than using KNN. As the pixel-level labels are unavailable in training, we set τ to a value which gives coverage of at least 15% of each image on a validation set. To compare the performance of image-adaptive τ with a fixed-value τ, we include the test segmentation scores on the CDnet-14 dataset with different fixed values of τ in Table 5. Furthermore, we evaluate the segmentation performance with different numbers of nearest neighbours used to estimate τ and notice that the best performance is achieved when the K is set to 6 (for KNN) to estimate the normalized foreground probability mass (Fig. 5). Finally, we present some error cases of our approach in Fig. 6.

5 Conclusion

This paper tackles the problem of weakly supervised change detection in paired images. We developed a novel CNN based model, which predicts change events and their location. Our approach defines a dense CRF model on top of the CNN activations and uses a modified mean-field inference procedure to enforce the compatibility between image and pixel level predictions. The proposed algorithm achieved a significant boost both in the case of detection and localisation of change events compared to strong baseline procedures. Our work is the first effort in the area of weakly supervised change detection using paired images and will find possible applications in damage detection, structural monitoring and automatic 3D model updating systems. In future, we will explore the possibility of multi-class change detection in pair of images/videos.

References

[Adams *et al.*, 2010] Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*, volume 29, pages 753–762. Wiley Online Library, 2010.

[Badrinarayanan *et al.*, 2013] Vijay Badrinarayanan, Ignas Budvytis, and Roberto Cipolla. Semi-supervised video segmentation using tree structured graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2751–2764, 2013.

[Bourdis *et al.*, 2011] Nicolas Bourdis, Denis Marraud, and Hichem Sahbi. Constrained optical flow for aerial image change detection. In *IGARSS*, pages 4176–4179. IEEE, 2011.

[Boykov *et al.*, 2001] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587. IEEE, 2014.

[Gueguen and Hamid, 2015] Lionel Gueguen and Raffay Hamid. Large-scale damage detection using satellite imagery. *CVPR*, 2(2):3, 2015.

[Jain and Grauman, 2013] Suyog Dutt Jain and Kristen Grauman. Predicting sufficient annotation strength for interactive foreground segmentation. In *ICCV*, pages 1313–1320. IEEE, 2013.

[Khan *et al.*, 2015] 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.

[Khan *et al.*, 2016] 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.

[Khan *et al.*, 2017] Salman H Khan, Xuming He, Fatih Porikli, and Mohammed Bennamoun. Forest change detection in incomplete satellite images with deep neural networks. *IEEE Transactions on Geosciences and Remote Sensing*, 2017.

[Kohli *et al.*, 2007] Pushmeet Kohli, M Pawan Kumar, and Philip HS Torr. P3 & beyond: Solving energies with higher order cliques. In *CVPR*, pages 1–8. IEEE, 2007.

[Krähenbühl and Koltun, 2011] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011.

[Krähenbühl and Koltun, 2013] Philipp Krähenbühl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *ICML*, pages 513–521, 2013.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440. IEEE, 2015.

[Papandreou *et al.*, 2015] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *CVPR*, pages 1742–1750. IEEE, 2015.

[Pinheiro and Collobert, 2015] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, pages 1713–1721. IEEE, 2015.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Amy L Bearman, Vittorio Ferrari, and Fei-Fei Li. What's the point: Semantic segmentation with point supervision. *arXiv preprint arXiv:1506.02106*, 2015.

[Sakurada and Okatani, 2015] Ken Sakurada and Takayuki Okatani. Change detection from a street image pair using cnn features and superpixel segmentation. In *BMVC*, 2015.

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

[Song *et al.*, 2015] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.

[Vineet *et al.*, 2014] Vibhav Vineet, Jonathan Warrell, and Philip HS Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *International Journal of Computer Vision*, 110(3):290–307, 2014.

[Wang *et al.*, 2014] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. Cdnet 2014: An expanded change detection benchmark dataset. In *CVPR Workshops*, pages 393–400. IEEE, 2014.

[Yuille and Rangarajan, 2003] Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.

[Zagoruyko and Komodakis, 2015] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. *arXiv preprint arXiv:1504.03641*, 2015.