Weakly Supervised Change Detection in a Pair of Images

Salman H Khan · Xuming He · Mohammed Bennamoun · Fatih Porikli · Ferdous Sohel · Roberto Togneri

Received: date / Accepted: date

Abstract Conventional change detection methods require a large number of images to learn background models. The few recent approaches that attempt change detection between two images either use handcrafted features or depend strongly 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 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 large benchmark datasets demonstrate superior detection and localization performance.

Keywords change detection · weakly supervised segmentation · deep learning · structured learning

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 [38], which is the largest dataset for video based change detection.
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.

This dataset required manual annotations for ~8 billion pixel locations. Although sophisticated methods have been investigated to reduce the human effort, e.g., by expert feedback in case of ambiguity [11,13], semi-automatic propagation of annotations [2], and point-wise supervision [27], acquisition of accurate and dense pixel-wise labels still remains a daunting task [18,32].

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 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 3). 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 5.2). 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 detection [10,12] and semantic labeling tasks [19,22,25].

In order to couple the image-level and pixel-level predictions, we introduce a constrained mean-field inference algorithm (Sec 3.3), 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 [22] and object based pri-
ors [25,27]. 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 (CDnet-2014, PCD-2014, and AICD-2012) and a custom built satellite image dataset (GASI-2015) (Sec 5.3). Our experimental results demonstrate that the proposed approach outperforms the state-of-the-art by a large margin (Sec 5.3). The key contributions of our work include:

- To the best of our knowledge, this is the first work to address the problem of weakly supervised change detection.
- 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 Related Work

Change Detection (CD) in before-and-after images has mostly been investigated for the case of satellite images [6,11]. In these scenarios, the acquisition of two images of the same location at different time instances is more feasible than acquisition of a video sequence. Traditional approaches normally use multiple images acquired from multiple views and at different times to create appearance or geometry based models (e.g., using SFM [21]) that capture the background state [6, 7, 26, 35]. Query images are then compared against the learnt model to detect changes. Recently, Gueguen and Hamid [11] automatically analyzed paired images for damage detection caused by calamities such as an earth-quake or a flood. These methods, however, mostly use hand-crafted features, multiple images from different viewpoints [7] (or videos [38]) and do not explore weakly supervised learning to localize the changes.

Anomaly Detection problem is closely related to change detection [20]. The training data consists of a large number of positive samples that show the usual trend under normal conditions. The goal is to detect new or unusual trends in the test data which were absent in the training data [24]. This task is different from our problem because it involves learning a one-class model, and consequently there is no supervision for anomalies. Some representative approaches include distance based methods [9], probabilistic approaches which model the underlying data distribution [5], data transformation based methods [14, 31], and classification boundary based approaches [29].

Deep Networks for CD: In the last few years, CNNs have demonstrated their superior learning capability for a number of computer vision tasks. Following the trend, few recent methods have explored the possibility of change detection using deep neural networks [28, 33, 41]. However, [28] uses CNN only as a feature extraction module. Similarly, [33, 41] do not localize the dissimilarities between patches and therefore do not explore the weakly supervised segmentation.

Weakly Supervised Segmentation: Several recent works have explored the weakly supervised semantic labeling problem using CNNs [22, 23, 25, 39]. Xu et al [39] proposed a unified model to make use of different forms of weak labels, e.g., bounding boxes. However, their approach employs an R-CNN as a feature extractor and does not explore end-to-end training. Russakovsky et al [27] used an objectness prior to improve the performance of weakly supervised semantic segmentation. The pixel-level activations were aggregated in
to train an end-to-end network using image-level annotations. Similar to [27], they used a segmentation prior to locate the salient object. Our work is closer to that of [22, 23], which deal with a single image semantic labeling under weak supervision. In contrast, our approach performs joint prediction of image and pixel level change labels using a pair of images.

Cardinality Constraints in CRF: The Potts model based CRF approaches [15, 37] add cardinality constraints over segments generated using clustering algorithms. The existing CRF inference methods with cardinality potentials are designed for certain classes of graphical models. The approach presented in [36] is applicable to CRFs with sparse connectivities, while [34] is applied to RBMs. In contrast, we develop a principled way to add the cardinality constraint in the mean-field inference for fully-connected CRF.

3 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 first introduce the model design and develop a constrained mean-field inference algorithm for the change label prediction and localization in this section. The model parameter learning with weak-supervision will be detailed in Sec 4.

3.1 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, \( x = \{I^1, I^2\} \). We associate an image-level output label vector \( y = \{y_1, \ldots, y_C\} \) to indicate the occurred change events, where \( C \) is the number of change categories and \( y_c \in \{0, 1\} \) denotes whether the change class \( c \) occurs. It is important to note here that the no-change category (i.e., \( y = 0 \)) 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 \( h \) to denote the labels of individual pixel locations for each image pair \( x \). Assume the image has \( m \) pixel locations, \( h = \{h_1, \ldots, h_m\} \in \{0, 1, \ldots, C\}^m \).

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 \( y \) and \( h \) as follows,

\[
P(y, h|x; \theta) \propto P(y|x; \theta_l)P(h|x; \theta_u)\Phi(h, y|x, \theta_p)
\]  

(1)

where \( P(y|x; \theta_l) \) is the unary term for image-level label \( y \), modeled by a CNN with parameter \( \theta_l \), and \( P(h|x; \theta_u) \) is the unary term for pixel-level labels, modeled by a Fully Convolutional Network (FCN) with parameter \( \theta_u \). The pairwise potential \( \Phi \) consists of two terms, \( \psi_p \) and \( \psi_u \), 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,

$$y^*, h^* = \arg \max_{y, h} P(y \mid x; \theta_l) P(h \mid x; \theta_u) \Phi(h, y \mid x, \theta_p)$$  \hspace{1cm} (2)

A graphical illustration of the model is shown in Fig 2. In the current work, we only consider two classes (change or no-change); therefore, $C = 1$ and we denote the image-level label as $y$ in the following for notation clarity. Note that our approach can be applied to detect multiple change classes ($C > 1$), where only a single dominant change category appear in a given pair of images.

### 3.2 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) [30] 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 the results of Zagoruyko et al. [41]), 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 [19] 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 potential:** 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 $P(y \mid x; \theta_l)$, predicting the conditional probability of image-level change as:

$$P(y \mid x; \theta_l) \propto \exp(F_l - \text{cnn}(x; W_s, W_y))$$ \hspace{1cm} (3)

where $F_l - \text{cnn}$ 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 potential: 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 [19]. Formally, the segmentation branch of the CNN model generates the pixel-level change label probability as follows,

$$P(h_x; \theta_u) \propto \prod_{j=1}^{o_{sz}} P(h_j | x), \quad P(h_j | x) \propto \exp(F_{u-cnn}(h_j, x; W_s, W_f))$$

(4)

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 potential $\Phi(h, y | x, \theta_p)$ 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 CRF to impose the spatial smoothness of the pixel labeling. Unlike the previous models [22], 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 $h$ by the following potential functions,

$$\Phi(h, y | x, \theta_p) = \prod_j \psi_u(h_j, y, x_j) \prod_{j<k} \psi_p(h_j, h_k, f_j, f_k)$$

(5)

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

$$\psi_u(h_j, y, x_j) = [y = 0][h_j = 0] + [y = 1](1 + e^{-\gamma \delta(x_j) [h_j = 0]})$$

(6)

where $\gamma$ is a weight parameter and $\delta(x_j)$ is the color difference between two images at pixel $j$. The fully-connected pairwise term $\psi_p$ defines the smoothing term between the latent variables $h$ given input features $f_j, f_k$. These potentials have a functional form of the weighted

![Fig. 3 CNN architecture](image-url)
Weakly Supervised Change Detection in a Pair of Images

Potts model in which the weight is defined using Gaussian kernels of Krahenbuhl et al [16]:

\[ \psi_p(h_j, h_k, f_j, f_k) = (\alpha_{ap} k_{ap} + \alpha_{sm} k_{sm}) \mu(h_j, h_k) \]  

where \( \alpha_{ap} \) and \( \alpha_{sm} \) are the kernel weights, \( \mu(h_j, h_k) \) is the Potts compatibility while \( k_{ap}(f_j, f_k) \) and \( k_{sm}(f_j, f_k) \) are the appearance and smoothness kernels as in [16].

3.3 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,

\[ y^* = \arg \max_y P(y|x; \theta_l), \quad h^* = \arg \max_h P(h|x; \theta_u) \Phi(h, y^* | x; \theta_p) \]  

This prioritized inference procedure allows us to compute the image-level label first, which is usually more reliable, and to run an efficient mean-field inference for the pixel-level labels only once.

We now derive a constrained mean-field inference algorithm for inferring the pixel-level change labeling \( h \). We note that the efficient mean-field algorithm [16] 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., [22, 25, 27]), 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 \( \tau \), 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 \( \tau \) (see more details in Sec 5.3). To enforce the proportion constraint, we introduce the following factorized approximate probability family with a global constraint:

\[ Q(h|x, y^*) = \prod_j q_j(h_j), \quad \text{and} \quad \sum_j [0, 1] q_j = \tau \]  

where, \( q_j = [q_j(0), q_j(1)]^T \) and the constraint implies that the overall foreground probability mass \( \sum_j q_j(1) \) is \( \tau \). Following [17], we minimize the approximate KL-divergence,

\[ D_{Q||P} = \sum_j (q_j^T \log q_j + q_j^T u_j) + \frac{1}{2} \sum_{j,k} q_j^T \Psi_{jk} q_k + C \]  

where \( u_j \) is the unary term vector (including \( P(h_j|x) \) and \( \psi_u(h_j, y^*, x_j) \)) and \( \Psi_{jk} \) is the compatibility matrix computed from \( \psi_p \), and \( C = \log Z \) is the log partition function. We use the CCCP algorithm [40] to minimize \( D_{Q||P} \) iteratively. At each iteration \( t \), the CCCP

---

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.
algorithm first linearizes the concave term \( \sum_j q_j^T u_j + \frac{1}{2} \sum_{j,k} q_j^T \Psi_{jk} q_k \) by taking its gradient,

\[
e_j^{(t)} = u_j + \sum_k \Psi_{jk} q_k^{(t-1)}
\]  

which can be computed efficiently by fast filtering as in [16]. Then the objective becomes a convex function with linear equality constraints, which can be minimized by solving the KKT condition:

\[
\log q_j^{(t)} + \lambda [0, 1]^T + e_j^{(t)} = 0, \quad 1^T q_j^{(t)} = 1, \quad \sum_j [0, 1] q_j = \tau 
\]

where, \( \lambda \) and \( \eta \) are the Lagrange multipliers which ensure that the \( q_j \) is a valid probability distribution and the global proportion constraint is satisfied. Eq (12) leads to the following mean field update equation:

\[
q_j^{(t)} = \frac{1}{Z_j} \exp \left( -e_j^{(t)} + [0, \eta]^T \right)
\]

At each iteration, we carry out a line search on \( \eta \) to make sure the second constraint in Eq (12) is satisfied.

4 EM Learning with Weak Supervision

We now consider a weakly supervised learning approach to estimate the parameters of the two-stream DCNN model (Sec 3). 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 \( D \) comprising of \( N \) labeled image pairs: \( D = \{x^n, y^n\}_{1 \times N} \).

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

\[
\sum_n \log P(y^n | x^n; \theta) \geq \sum_n \sum_{h^n} Q(h^n | y^n, x^n) \log \frac{P(y^n, h^n | x^n; \theta)}{Q(h^n | y^n, x^n)} \quad (14)
\]

\[
= E_Q[\log P(y^n, h^n | x^n; \theta)] + H(Q(h^n | x^n, y^n)) \quad (15)
\]

where, \( E_Q[\cdot] \) and \( H(\cdot) \) denote the expected value and the entropy function respectively, and \( Q(h^n | x^n, y^n) \) is an approximate posterior probability factorizing over \( \{h^n\} \) as defined in Eq (9). In other words, the posterior probability can be expressed as the product of independent marginals: \( Q(h^n | x^n, y^n) = \prod_i q_i^n(h^n_i) \). We then derive a variational expectation-maximization (EM) algorithm for learning our two-stream deep CNN in the following, which alternately maximizes the objective function above.
4.1 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 \( q(h^n) \) as in Eq (11) and (13). 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 [1].

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

\[
  h_j^{n^*} \leftarrow \arg\max_{h_j} \prod_{j=1}^{m} q(h_j^n | x^n, y^n) \tag{16}
\]

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

4.2 M step for CNN+CRF training

Once we have the posterior marginal distribution \( q(h^n_j) \) and its mode, we update the model parameters \( \theta \) with the posterior mode configuration \( \{h^{n^*}\} \) 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 \( (W) \), while the individual gradients are computed using \( y^n \) and \( h^{n^*} \) 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,

\[
  W_s^* \leftarrow \arg\max_{W_s} \sum_n \left( \log P(y^n | x^n; \theta_l) + \log P(h^{n^*} | x^n; \theta_u) \right) \\
  W_y^* \leftarrow \arg\max_{W_y} \sum_n \log P(y^n | x^n; \theta_l), \quad W_f^* \leftarrow \arg\max_{W_f} \sum_n \log P(h^{n^*} | x^n; \theta_u) \tag{17}
\]

After the two-stream deep network component is trained, we estimate the remaining parameters \( \theta_p \) in Eq (5) by cross-validation.

The overall EM procedure starts with an M step with an initial value of \( h^n \). We assume the initial hidden variable states \( (h_0^n) \) to be consistent with the image-level labels: \( h_0^n = y^n \). The model parameters are fine-tuned by training the two-stream CNN 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.

---

2 We note that it is possible to implement an end-to-end learning procedure to learn all the parameters based on back-propagation and unfolded mean-field inference but we found the stage-wise learning is more efficient and works reasonably well in this case.
5 Experiments

5.1 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 ($32 \times 32$) at the output of the segmentation branch, enlarged paired images of size $512 \times 512$ are fed to the CNN. Moreover, the convolution filter size in FC1 (segmentation branch) is kept to $1 \times 1$ (in contrast to a $7 \times 7$ filter size in FC1') to avoid the additional decrease in resolution of the $32 \times 32$ output map.

The unary potentials of our CRF model are defined using the CNN activations, while the Gaussian edge potentials proposed by Krahenbuhl et al. [16] 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 for examples). 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 the hidden variables. The resulting segmentation map is used as ground-truth during the CNN training (M step).

5.2 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 [38]. 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 [3] consists of 1000 pairs of large sized images ($800 \times 600$). Because the change regions are very small in satellite/aerial images, we work on the patch level and extract 48 patches of size $122 \times 122$ 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. 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 regions of interest which were identified by experts. In total, nearly $\sim 300$ pairs are generated for each region of interest which makes a total of $\sim 20,000$
Fig. 4 Qualitative results on the CDnet-2014 dataset: Rows 1 – 2 show image pairs, our results are shown in row 3 and the ground-truths are shown in row 4. No-change pixels are shown in black, regions outside the ROIs are shown in sky-blue while the changes are shown in coral-pink.

pairs. Since the raw data contains artifacts, we improved it’s quality by filling data across different time instances.

PCD 2015 Dataset: Panoramic change detection dataset 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.

5.3 Results

The change detection results of our approach on the four major CD datasets are shown in Table 2. 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 2). 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. On all four datasets, the average precision (AP) and the overall accuracy of our approach was significantly higher than that of the baseline procedure (specifically 6.8%, 4.6%, 5.3%, 7.8% boost in AP for the CDnet, AICD, GAS1 and PCD 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 2). 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 1 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 (2nd column-RS), thresholding applied to a difference map obtained from the pair of images (3rd column-DT), thresholding applied to the output from a pre-trained network (weights initialized for segmentation branch using VGG-net, 4th column-PN), thresholding applied

More details about the dataset are in the supplementary material.
Table 1 Segmentation Results and comparisons with baseline methods. Note that all results (except the last column) are reported for the weakly-supervised setting.

| Dataset  | Baseline Approaches (mIOU-%) | This Paper Fully-supervised |
|----------|------------------------------|-----------------------------|
|          | RS  | DT  | PN  | Th. | GC  | (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 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.

Table 3 Ablative Analysis on the CDnet-2014 Dataset. Change localization performance decreases when the full model is not used. Specifically, both the CNN architecture and the proposed CRF model contribute towards the accurate change detection.

| 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                        |

We also report segmentation results on two additional baselines which use cardinality based pattern potentials (Table 4). These baselines include the higher order potential (HOP) based dense and grid CRF models of Vineet et al. [37] and Kohli et al. [15] 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 [15]. We note that the dense CRF model with $P^n$ HOP [37] performs better than the grid CRF model [15], however our deep structured prediction 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 dataset are shown in Fig 4. The proposed approach performed well in localizing small as well as large sized

---

*More example results can be found in the supplementary material.*
Weakly Supervised Change Detection in a Pair of Images

| Method               | Dense CRF + HOP [37] | Grid CRF + HOP [15] | This Paper |
|----------------------|----------------------|---------------------|------------|
| mIOU% (CDnet-14)     | 42.0                 | 38.3                | 46.2       |

Table 4 Comparisons for segmentation performance with methods using cardinality potentials.

Fig. 5 Qualitative results on the GASI-2015 and PCD-2015 datasets: Paired images (rows 1 – 2), our results (row 3) and the ground-truths (row 4) are shown. No-change pixels are shown in black while the changes are shown in coral-pink.

changes (e.g., 1st 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 on GASI-2015 and PCD-2015 datasets (shown in Fig 5) indicate the good performance of our method for satellite image based change detection and damage detection. It is worth noticing that the ground-truths available particularly for the case of GASI-2015 are not very accurate and the generated results are correct in several cases in contrast to the available ground-truths.

Some examples of ambiguous cases are shown in Fig 6. One major problem with our results is that full object coverage was not provided in the predicted masks, although the specific changes were precisely determined in most cases (e.g., 2nd column, Fig 6). We attribute this behaviour to the fact that change localisation is performed using a 32 \times 32 resolution output from the network. Moreover, we were not able to use priors to encourage a full object coverage (e.g., as in [27]) because of the diversity of datasets (and their constituent objects) used for evaluation. In some cases, our method also missed some subtle changes, e.g., a person in 3rd column, Fig 6.

We performed an ablative study on the CDnet-2014 dataset for the change segmentation task (Table 3). 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 \( \tau \), which gives better estimate of foreground proportion and performance. We perform the KNN search using Euclidean
Fig. 6 Ambiguous Cases: We present some example cases, for which the ground-truths didn’t exactly match with the generated results. No-change pixels are shown in black, regions outside the ROIs are shown in sky-blue, while the changes are shown in coral-pink.

Table 5 Segmentation performance for different fixed values of $\tau$.

| Normalized $\tau$ | 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|

Fig. 7 Sensitivity analysis on the number of nearest neighbours used to estimate foreground probability mass parameter ($\tau$) for the CDnet-2014 dataset.

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 precisely locates them. 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

1. Adams, A., Baek, J., Davis, M.A.: Fast high-dimensional filtering using the permutohedral lattice. In: Computer Graphics Forum, vol. 29, pp. 753–762. Wiley Online Library (2010)
2. Badrinarayanan, V., Budvytis, I., Cipolla, R.: Semi-supervised video segmentation using tree structured graphical models. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(11), 2751–2764 (2013)
3. Bourdis, N., Marraud, D., Sahbi, H.: Constrained optical flow for aerial image change detection. In: IGARSS, pp. 4176–4179. IEEE (2011)
4. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(11), 1222–1239 (2001)
5. Clifton, D.A., Hugueny, S., Tarassenko, L.: Novelty detection with multivariate extreme value statistics. Journal of Signal Processing Systems 65(3), 371–389 (2011)
6. Crispell, D., Mundy, J., Taubin, G.: A variable-resolution probabilistic three-dimensional model for change detection. IEEE Transactions on Geoscience and Remote Sensing 50(2), 489–500 (2012)
7. Eden, I., Cooper, D.B.: Using 3d line segments for robust and efficient change detection from multiple noisy images. In: ECCV, pp. 172–185. Springer (2008)
8. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874 (2008)
9. Ghoting, A., Parthasarathy, S., Otey, M.E.: Fast mining of distance-based outliers in high-dimensional datasets. Data Mining and Knowledge Discovery 16(3), 349–364 (2008)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp. 580–587. IEEE (2014)
11. Guéguen, L., Hamid, R.: Large-scale damage detection using satellite imagery. CVPR 2(2), 3 (2015)
12. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: ECCV, pp. 345–360. Springer (2014)
13. Jain, S.D., Grauman, K.: Predicting sufficient annotation strength for interactive foreground segmentation. In: ICCV, pp. 1313–1320. IEEE (2013)
14. Kassab, R., Alexandre, F.: Incremental data-driven learning of a novelty detection model for one-class classification with application to high-dimensional noisy data. Machine learning 74(2), 191–234 (2009)
15. Kohli, P., Kumar, M.P., Torr, P.H.: P3 & beyond: Solving energies with higher order cliques. In: Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on, pp. 1–8. IEEE (2007)
16. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS, pp. 109–117 (2011)
17. Krähenbühl, P., Koltun, V.: Parameter learning and convergent inference for dense random fields. In: ICML, pp. 513–521 (2013)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV, pp. 740–755. Springer (2014)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440. IEEE (2015)
20. Markou, M., Singh, S.: Novelty detection: a review part 1: statistical approaches. Signal processing 83(12), 2481–2497 (2003)
21. Matzen, K., Snively, N.: Scene chronology. In: ECCV, pp. 615–630. Springer (2014)
22. Papandreou, G., Chen, L.C., Murphy, K., Yuille, A.L.: Weakly and semi-supervised learning of a dcmn for semantic image segmentation (2015)
23. Pathak, D., Krahenbuhl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation (2015)
24. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: CVPR, pp. 1713–1721. IEEE (2015)
25. Pollard, T., Mundy, J.L.: Change detection in a 3-d world. In: CVPR, pp. 1–6. IEEE (2007)
26. Russakovsky, O., Bearman, A.L., Ferrari, V., Li, F.F.: What’s the point: Semantic segmentation with point supervision. arXiv preprint arXiv:1506.02106 (2015)
27. Sakurada, K., Okatani, T.: Change detection from a street image pair using cnn features and superpixel segmentation. In: BMVC (2015)
28. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C.: Support vector method for novelty detection. In: NIPS, vol. 12, pp. 582–588 (1999)
29. Simonov, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
31. Singh, S., Markou, M.: An approach to novelty detection applied to the classification of image regions. IEEE Transactions on Knowledge and Data Engineering 16(4), 396–407 (2004)
32. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: CVPR, pp. 567–576 (2015)
33. Stent, S., Gherardi, R., Stenger, B., Cipolla, R.: Detecting change for multi-view, long-term surface inspection (2015)
34. Swersky, K., Sutskever, I., Tarlow, D., Zemel, R.S., Salakhutdinov, R., Adams, R.P.: Cardinality restricted boltzmann machines. In: Advances in Neural Information Processing Systems, pp. 3302–3310 (2012)
35. Taneja, A., Ballan, L., Pollefeys, M.: Image based detection of geometric changes in urban environments. In: ICCV, pp. 2336–2343. IEEE (2011)
36. Tarlow, D., Swersky, K., Zemel, R.S., Adams, R.P.: Fast exact inference for recursive cardinality models. In: Proceedings of the Twenty-Eighth Conference Conference on Uncertainty in Artificial Intelligence. AUAI Press (2012)
37. Vineet, V., Warrell, J., Torr, P.H.: 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)
38. Wang, Y., Jodoin, P.M., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P.: Cnet 2014: An expanded change detection benchmark dataset. In: CVPR, pp. 393–400. IEEE (2014)
39. Xu, J., Schwing, A.G., Urtasun, R.: Learning to segment under various forms of weak supervision. In: CVPR, pp. 3781–3790. IEEE (2015)
40. Yuille, A.L., Rangarajan, A.: The concave-convex procedure. Neural computation 15(4), 915–936 (2003)
41. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. arXiv preprint arXiv:1504.03641 (2015)