Higher Order Conditional Random Fields in Deep Neural Networks

Anurag Arnab  Sadeep Jayasumana  Shuai Zheng  Philip Torr
University of Oxford, United Kingdom
{firstname.lastname}@eng.ox.ac.uk

Abstract

We address the problem of semantic segmentation using deep learning. Most segmentation systems include a Conditional Random Field (CRF) to produce a structured output that is consistent with the image’s visual features. Recent deep learning approaches have incorporated CRFs into Convolutional Neural Networks (CNNs), with some even training the CRF end-to-end with the rest of the network. However, these approaches have not employed higher order potentials, which have previously been shown to significantly improve segmentation performance. In this paper, we demonstrate that two types of higher order potential, based on object detections and superpixels, can be included in a CRF embedded within a deep network. We design these higher order potentials to allow inference with the differentiable mean field algorithm. As a result, all the parameters of our richer CRF model can be learned end-to-end with our pixelwise CNN classifier. We achieve state-of-the-art segmentation performance on the PASCAL VOC benchmark with these trainable higher order potentials.

1. Introduction

Semantic segmentation involves assigning a visual object class label to every pixel in an image, resulting in a segmentation with a semantic meaning for each segment. While a strong pixel-level classifier is critical for obtaining high accuracy in this task, it is also important to enforce the consistency of the semantic segmentation output with visual features of the image. For example, segmentation boundaries should usually coincide with strong edges in the image, and regions in the image with similar appearance should have the same label.

Recent advances in deep learning have enabled researchers to create stronger classifiers, with automatically learned features, within a Convolutional Neural Network (CNN) [26, 35, 45]. This has resulted in large improvements in semantic segmentation accuracy on widely used benchmarks such as PASCAL VOC [10]. CNN classifiers are now considered the standard choice for pixel-level classifiers used in semantic segmentation.

On the other hand, probabilistic graphical models have long been popular for structured prediction of labels, with constraints enforcing label consistency. Conditional Random Fields (CRFs) have been the most common framework, and various rich and expressive models [27, 29, 48], based on higher order clique potentials, have been developed to improve segmentation performance. Whilst some deep learning methods showed impressive performance in semantic segmentation without incorporating graphical models [17, 35], current state-of-the-art methods [4, 30, 34, 38, 53] have all incorporated graphical models into the deep learning framework in some form. However, we observe that the CRFs that have been incorporated into deep learning techniques are still rather rudimentary as they consist of only unary and pairwise potentials [53]. In this paper, we show that CRFs with carefully designed higher order potentials (potentials defined over cliques consisting of more than two nodes) can also be modelled as CNN layers when using mean field inference [23]. The advantage of performing CRF inference within a CNN is that it enables joint optimisation of CNN classifier weights and CRF parameters during the end-to-end training of the complete system. Intuitively, the classifier and the graphical model learn to optimally co-operate with each other during the joint training.

We introduce two types of higher order potential into the CRF embedded in our deep network: object-detection based potentials and superpixel-based potentials. The primary idea of using object-detection potentials is to use the outputs of an off-the-shelf object detector as additional semantic cues for finding the segmentation of an image. Intuitively, an object detector with a high recall can help the semantic segmentation algorithm by finding objects appearing in an image. As shown in Fig. 1, our method is able to recover from poor segmentation unaries when we have a confident detector response. However, our method is robust to false positives identified by the object detector since CRF inference identifies and rejects false detections that do not agree with other types of energies present in the CRF.

Superpixel-based higher order potentials encourage label consistency over superpixels obtained by oversegmentation.
This is motivated by the fact that regions defined by superpixels are likely to contain pixels from the same visual object. Once again, our formulation is robust to the violations of this assumption and errors in the initial superpixel generation step. In practice, we noted that this potential is effective for getting rid of small regions of spurious labels that are inconsistent with the correct labels of their neighbours.

We evaluate our higher order potentials on the PASCAL VOC 2012 semantic segmentation benchmark as well as the PASCAL Context dataset, to show significant improvements over our baseline and achieve state-of-the-art results.

2. Related Work

Before deep learning became prominent, semantic segmentation was performed with dense hand-crafted features which were fed into a per-pixel or region classifier [44]. The individual predictions made by these classifiers were often noisy as they lacked global context, and were thus post-processed with a CRF, making use of prior knowledge such as the fact that nearby pixels, as well as pixels of similar appearance, are likely to share the same class label [18, 44].

The CRF model of [44] initially contained only unary and pairwise terms in an 8-neighbourhood, which [21] showed can result in shrinkage bias. Numerous improvements to this model were subsequently proposed including: densely connected pairwise potentials facilitating interactions between all pairs of image pixels [24], formulating higher order potentials defined over cliques larger than two nodes [21, 27] in order to capture more context, modelling co-occurrence of object classes [14, 28, 39], and utilising the results of object detectors [29, 32, 49, 51].

Recent advances in deep learning have allowed us to replace hand-crafted features with features learned specifically for semantic segmentation. The strength of these representations was illustrated by [35] who achieved significant improvements over previous hand-crafted methods without using any CRF post-processing. Chen et al. [4] showed further improvements by post-processing the results of a CNN with a CRF. Subsequent works [30, 43, 53] have taken this idea further by incorporating a CRF as layers within a deep network and then learning parameters of both the CRF and CNN together via backpropagation.

In terms of enhancements to conventional CRF models, Ladicky et al. [29] proposed using an off-the-shelf object detector to provide additional cues for semantic segmentation. Unlike other approaches that refine a bounding-box detection to produce a segmentation [17, 50], this method used detector outputs as a soft constraint and can thus recover from object detection errors. Their formulation, however, used graph-cut inference, which was only tractable due to the absence of dense pairwise potentials. Object detectors have also been used by [46, 51], who also modelled variables that describe the degree to which an object hypothesis is accepted.

We formulate the detection potential in a different manner to [29, 46, 51] so that it is amenable to mean field inference. Mean field permits inference with dense pairwise connections, which result in substantial accuracy improvements [4, 24, 53]. Furthermore, mean field updates related to our potentials are differentiable and its parameters can thus be learned in our end-to-end trainable architecture.

On a separate track, [6] have used object detection ground truth to weakly supervise the training of CNNs for semantic segmentation and thus exploit the fact that there is more training data for object detection than segmentation. Gould et al. [15] used semantic segmentation to propose regions to detect. And by enforcing consistency between segmented regions, object detections and object instances in a graphical model, the two tasks of detection and segmentation were performed jointly in a unified model. However, a greedy move-making algorithm was used for inference. Such non-differentiable algorithms have not yet been incorporated into deep learning models.

We also note that while the semantic segmentation problem has mostly been formulated in terms of pixels [35, 44, 53], some have expressed it in terms of superpixels [2, 3, 7, 11]. Superpixels can capture more context than a single pixel and
computational costs can also be reduced if one considers pairwise interactions between superpixels rather than individual pixels [51]. However, such superpixel representations assume that the segments share boundaries with objects in an image, which is not always true. As a result, several authors [27, 48] have employed higher order potentials defined over superpixels that encourage label consistency over regions, but do not strictly enforce it. This approach also allows multiple, non-hierarchical layers of superpixels to be integrated. Our formulation uses this kind of higher order potential, but in an end-to-end trainable CNN.

Graphical models have been used with CNNs in other areas besides semantic segmentation, such as in pose-estimation [47] and group activity recognition [8]. However, the nature of models used in these works is substantially different to ours. Some early works that advocated gradient backpropagation through graphical model inference for parameter optimisation include [9, 19, 25] and [41].

Our work differentiates from the above works since, to our knowledge, we are the first to propose and conduct a thorough experimental investigation of higher order potentials that are based on detection outputs and superpixel segmentation in a CRF which is learned end-to-end in a deep network. Note that although [48] formulated mean field inference with higher order potentials, they did not consider object detection potentials at all, nor were the parameters learned.

3. Conditional Random Fields

We now review conditional random fields used in semantic segmentation and introduce the notation used in the paper. Take an image I with N pixels, indexed 1, 2, . . . , N. In semantic segmentation, we attempt to assign every pixel a label from a predefined set of labels L = {l1, l2, . . . , lL}. Define a set of random variables X1, X2, . . . , XN, one for each pixel, where each Xi ∈ L. Let X = [X1 X2 . . . XN]T. Any particular assignment x to X is thus a solution to the semantic segmentation problem.

We use notations {V}, and V(i) to represent the set of elements of a vector V, and the ith element of V, respectively. Given a graph G where the vertices are from {X} and the edges define connections among these variables, the pair (I, X) is modelled as a CRF characterised by Pr(X = x|I) = 1/Z(I) exp(−E(x|I)), where E(x|I) is the energy of the assignment x and Z(I) is the normalisation factor known as the partition function. We drop the conditioning on I hereafter to keep the notation uncluttered. The energy E(x) of an assignment is defined using the set of cliques C in the graph G. More specifically,

\[ E(x) = \sum_{c \in C} \psi_c(x_c), \]

where x_c is a vector formed by selecting elements of x that correspond to random variables belonging to the clique c, and \( \psi_c(.) \) is the cost function for the clique c. The function, \( \psi_c(.) \), usually uses prior knowledge about a good segmentation, as well as information from the image, the observation the CRF is conditioned on.

Minimising the energy yields the maximum a posteriori (MAP) labelling of the image. When dense pairwise potentials are used in the CRF to obtain higher accuracy, exact inference is impracticable, and one has to resort to an approximate inference method such as mean field inference [24]. Mean field inference is particularly appealing in a deep learning setting since it is possible to formulate it as a Recurrent Neural Network [53].

4. CRF with Higher Order Potentials

Many CRF models that have been incorporated into deep learning frameworks [4, 53] have so far used only unary and pairwise potentials. However, potentials defined on higher order cliques have been shown to be useful in previous works such as [21, 48]. The key contribution of this paper is to show that a number of explicit higher order potentials can be added to CRFs to improve image segmentation, while staying compatible with deep learning. We formulate these higher order potentials in a manner that mean field inference can still be used to solve the CRF. Advantages of mean field inference are twofold: First, it enables efficient inference when using densely-connected pairwise potentials. Multiple works, [25, 53] have shown that dense pairwise connections result in substantial accuracy improvements, particularly at image boundaries [4, 24]. Secondly, we keep all our mean field updates differentiable with respect to their inputs as well as the CRF parameters introduced. This design enables us to use backpropagation to automatically learn all the parameters in the introduced potentials.

We use two types of higher order potential, one based on object detections and the other based on superpixels. These are detailed in Sections 4.1 and 4.2 respectively. Our complete CRF model is represented by

\[ E(x) = \sum_i \psi_i^U(x_i) + \sum_{i<j} \psi_{ij}(x_i, x_j) + \sum_d \psi_d^\text{Det}(x_d) + \sum_s \psi_s^\text{SP}(x_s), \]

where the first two terms \( \psi_i^U(.) \) and \( \psi_{ij}(., .) \) are the usual unary and densely-connected pairwise energies [24] and the last two terms are the newly introduced higher order energies. Energies from the object detection take the form \( \psi_d^\text{Det}(x_d) \), where vector x_d is formed by elements of x that correspond to the foreground pixels of the dth object detection. Superpixel label consistency based energies take the form \( \psi_s^\text{SP}(x_s) \), where x_s is formed by elements of x that correspond to the pixels belonging to the s th superpixel.

4.1. Object Detection Based Potentials

Semantic segmentation errors can be classified into two broad categories [6]: recognition and boundary errors.
Boundary errors occur when semantic labels are incorrect at the edges of objects, and it has been shown that densely connected CRFs with appearance-consistency terms are effective at combating this problem [24]. On the other hand, recognition errors occur when object categories are recognised incorrectly or not at all. A CRF with only unary and pairwise potentials cannot effectively correct these errors since they are caused by poor unary classification. However, we propose that a state-of-the-art object detector [13, 40] capable of recognising and localising objects, can provide important information in this situation and help reduce the recognition error, as shown in Fig. 2.

A key challenge in feeding-in object-detection potentials to semantic segmentation are false detections. A naïve approach of adding an object detector’s output to a CRF formulated to solve the problem of semantic segmentation would confuse the CRF due to the presence of the false positives in the detector’s output. Therefore, a robust formulation, which can automatically reject object detection false positives when they do not agree with other types of potentials in the CRF, is desired. Furthermore, since we are aiming for an end-to-end trainable CRF which can be incorporated into a deep neural network, the energy formulation should permit a fully differentiable inference procedure. We now propose a formulation which has both of these desired properties.

Assume that we have $D$ object detections for a given image, and that the $d^{th}$ detection is of the form $(l_d, s_d, F_d)$, where $l_d \in L$ is the class label of the detected object, $s_d$ is the confidence score of the detection, and $F_d \subseteq \{1, 2, \ldots, N\}$, is the set of indices of the pixels belonging to the foreground of the detection. The foreground within a detection bounding box could be obtained using a foreground/background segmentation method (i.e. GrabCut [42]), and represents a crude segmentation of the detected object. Using our detection potentials, we aim to encourage the set of pixels represented by $F_d$, to take the label $l_d$. However, this should not be a hard constraint since the foreground segmentation could be inaccurate and the detection itself could be a false detection. We therefore seek a soft constraint that assigns a penalty if a pixel in $F_d$ takes a label other than $l_d$. Moreover, if other energies used in the CRF strongly suggest that many pixels in $F_d$ do not belong to the class $l_d$, the detection $d$ should be identified as invalid.

An approach to accomplish this is described in [29] and [51]. However, in both cases, dense pairwise connections were absent and different inference methods were used. In contrast, we would like to use the mean field approximation to enable efficient inference with dense pairwise connections [24], and also because its inference procedure is fully differentiable. We therefore use a detection potential formulation quite different to the ones used in [29] and [51].

In our formulation, as done in [29] and [51], we first introduce latent binary random variables $Y_1, Y_2, \ldots Y_D$, one for each detection. The interpretation for the random variable $Y_d$ that corresponds to the $d^{th}$ detection is as follows: If the $d^{th}$ detection has been found to be valid after inference, $Y_d$ will be set to 1, it will be 0 otherwise. Mean field inference probabilistically decides the final value of $Y_d$. Note that, through this formulation, we can account for the fact that the initial detection could have been a false positive: some of the detections obtained from the object detector may be identified to be false following CRF inference.

All $Y_d$ variables are added to the CRF which previously contained only $X_i$ variables. Let each $(X_d, Y_d)$, where $\{X_d\} = \{X_i \in \{X_i | i \in F_d\}$, form a clique $c_d$ in the CRF. We define the detection-based higher order energy associated with a particular assignment $(x_d, y_d)$ to the clique.

---

**Figure 2: Utility of object detections as another cue for semantic segmentation** For every pair, segmentation on the left was produced with only unary and pairwise potentials. Detection based potentials were added to produce the result on the right. Note how we are able to improve our segmentations for the bus, table and bird over their respective baselines. Furthermore, our system is able to reject erroneous detections such as the person in (b) and the bottle and chair in (d). Images were taken from the PASCAL VOC 2012 reduced validation set. Baseline results were produced using the public code and model of [53].
\((X_d, Y_d)\) as follows:

\[
\psi_{d_\text{Det}}(X_d = x_d, Y_d = y_d) = \begin{cases} 
  w_{d_\text{Det}} \frac{s_d}{n_d} \sum_{i=1}^{n_d} [x_d^{(i)} = l_d] & \text{if } y_d = 0, \\
  w_{d_\text{Det}} \frac{s_d}{n_d} \sum_{i=1}^{n_d} [x_d^{(i)} \neq l_d] & \text{if } y_d = 1,
\end{cases}
\]

(2)

where \(n_d = |F_d|\) is the number of foreground pixels in the \(d\)th detection, \(x_d^{(i)}\) is the \(i\)th element of the vector \(x_d\), \(w_{d_\text{Det}}\) is a learnable weight parameter, and \([\cdot]\) is the Iverson bracket. Note that this potential encourages \(X_d^{(i)}\) s to take the value \(l_d\) when \(Y_d\) is 1, and at the same time encourages \(Y_d\) to be 0 when many \(X_d^{(i)}\) s do not take \(l_d\). In other words, it enforces the consistency among \(X_d^{(i)}\) s and \(Y_d\).

An important property of the above definition of \(\psi_{d_\text{Det}}(\cdot)\) is that it can be simplified as a sum of pairwise potentials between \(Y_d\) and each \(X_d^{(i)}\) for \(i = 1, 2, \ldots, n_d\). That is,

\[
\psi_{d_\text{Det}}(X_d = x_d, Y_d = y_d) = \sum_{i=1}^{n_d} f_d(x_d^{(i)}, y_d), \quad \text{where,}
\]

\[
f_d(x_d^{(i)}, y_d) = \begin{cases} 
  w_{d_\text{Det}} \frac{s_d}{n_d} [x_d^{(i)} = l_d] & \text{if } y_d = 0, \\
  w_{d_\text{Det}} \frac{s_d}{n_d} [x_d^{(i)} \neq l_d] & \text{if } y_d = 1.
\end{cases}
\]

(3)

We make use of this simplification in Section 5 when deriving the mean field updates associated with this potential.

For the latent \(Y\) variables, in addition to the joint potentials with \(X\) variables, described in Eq. (2) and (3), we also include unary potentials, which are initialised from the score \(s_d\) of the object detection. The underlying idea is that if the object detector detects an object with high confidence, the CRF in turn starts with a high initial confidence about the validity of that detection. This confidence can, of course, change during the CRF inference depending on other information (e.g. segmentation unary potentials) available to the CRF.

Examples of input images with multiple detections and GrabCut foreground masks are shown in Figure 3. Note how false detections are ignored and erroneous parts of the foreground mask are also largely ignored.

### 4.2. Superpixel Based Potentials

The next type of higher order potential we use is based on the idea that superpixels obtained from oversegmentation [1, 12] quite often contain pixels from the same visual object. It is therefore natural to encourage pixels inside a superpixel to have the same semantic label. Once again, this should not be a hard constraint in order to keep the algorithm robust to initial superpixel segmentation errors and to violations of this key assumption.

We use two types of energies in the CRF to encourage superpixel consistency in semantic segmentation. Firstly, we use the \(P^n\)-Potts model type energy [20], which is described by,

\[
\psi_{s_\text{Det}}(X_s = x_s) = \begin{cases} 
  w_{s_\text{Low}}(l) & \text{if all } x_s^{(i)} = l, \\
  w_{s_\text{High}} & \text{otherwise},
\end{cases}
\]

(4)

where \(w_{s_\text{Low}}(l) < w_{s_\text{High}}\) for all \(l\), and \(\{X_s\} \subset \{X\}\) is a clique defined by a superpixel. The primary idea is that assigning different labels to pixels in the same superpixel incurs a higher cost, whereas one obtains a lower cost if the labelling is consistent throughout the superpixel. Costs \(w_{s_\text{Low}}(l)\) and \(w_{s_\text{High}}\) are learnable during the end-to-end training of the network.

Secondly, to make this potential stronger, we average initial unary potentials from the classifier (the CNN in our case), across all pixels in the superpixel and use the average as an additional unary potential for those pixels. During experiments, we observed that superpixel based higher order energy helps in getting rid of small spurious regions of wrong labels in the segmentation output, as shown in Fig. 4.

### 5. Mean Field Updates and Their Differentials

This section discusses the mean field updates for the higher order potentials previously introduced. These update operations are differentiable with respect to the \(Q_c(X_c)\) distribution inputs at each iteration, as well as the parameters of our higher order potentials. This allows us to train our CRF end-to-end as another layer of a neural network.

Take a CRF with random variables \(V_1, V_2, \ldots, V_N\) and a set of cliques \(\mathcal{C}\), which includes unary, pairwise and higher order cliques. Mean field inference approximates the joint distribution \(\Pr(V)\) with the product of marginals \(\prod_c Q_c(V_c)\). We use \(Q_c(V_c = v_c)\) to denote the marginal probability mass for a subset \(\{V_c\}\) of these variables. Where there is no ambiguity, we use the short-hand notation \(Q(v_c)\) to represent \(Q(V_c = v_c)\). General mean field updates of such a CRF take the form [23]

\[
Q_{t+1}(V_i = v) = \frac{1}{Z_i} \exp \left( - \sum_{c \in \mathcal{C}} \sum_{v_{c-1}} Q^t(v_{c-1}) \psi_c(v_c) \right),
\]

(5)

where \(Q^t\) is the marginal after the \(t\)th iteration, \(v_c\) an assignment to all variables in clique \(c\), \(v_{c-1}\) an assignment to all variables in \(c\) except for \(V_i\), \(\psi_c(v_c)\) is the cost of assigning \(v_c\) to the clique \(c\), and \(Z_i\) is the normalisation constant that makes \(Q(V_i = v)\) a probability mass function after the update.

### 5.1. Updates from Detection Based Potentials

Following Eq. (3) above, we now use Eq. (5) to derive the mean field updates related to \(\psi_{d_\text{Det}}\). The contribution from
Figure 3: Effects of imperfect foreground segmentation (a,b) Detected objects, as well as the foreground masks obtained from GrabCut. (c,d) Output using detection potentials. Incorrect parts of the foreground segmentation of the main aeroplane, and entire TV detection have been ignored by CRF inference as they did not agree with the other energy terms. The person is a failure case though as the detection has caused part of the sofa to be erroneously labelled.

Figure 4: Segmentation enhancement from superpixel based potentials (a) The output of our system without any superpixel potentials. (b) Superpixels obtained from the image using the method of [12]. Only one “layer” of superpixels is shown. In practice, we used four. (c) The output using superpixel potentials. The result has improved as we encourage consistency over superpixel regions. This removes some of the spurious noise that was present previously.

\[ \psi_{\text{Det}} \] to the update of \( Q(X_d^{(i)} = l) \) takes the form
\[
\sum_{\{(x_d, y_d) | x_d^{(i)} = l\}} Q(x_d, y_d) \psi_{\text{Det}}(x_d, y_d) = \\
\begin{cases} 
    w_{\text{Det}} \frac{a_d}{n_d} Q(Y_d = 0) & \text{if } l = l_d, \\
    w_{\text{Det}} \frac{a_d}{n_d} Q(Y_d = 1) & \text{otherwise}, 
\end{cases}
\]
where \( x_{d-i} \) is an assignment to \( X_d \) with the \( i \)th element deleted. Using the same equations, we derive the contribution from the energy \( \psi_{\text{Det}} \) to the update of \( Q(Y_d = b) \) to take the form
\[
\sum_{\{(x_d, y_d) | y_d = b\}} Q(x_d) \psi_{\text{Det}}(x_d, y_d) = \\
\begin{cases} 
    w_{\text{Det}} \frac{a_d}{n_d} \sum_{i=1}^{n_d} Q(X_d^{(i)} = l_d) & \text{if } b = 0, \\
    w_{\text{Det}} \frac{a_d}{n_d} \sum_{i=1}^{n_d} (1 - Q(X_d^{(i)} = l_d)) & \text{otherwise}. 
\end{cases}
\]

Note that due to the presence of detection potentials in the CRF, error differentials calculated with respect to the \( X \) variable unary potentials and pairwise parameters will no longer be valid in the forms described in [53]. The error differentials with respect to the \( X \) and \( Y \) variables, as well as class-specific detection potential weights \( w_{\text{Det}}(l_d) \) are included in the supplementary material.

5.2. Updates for Superpixel Based Potentials

The contribution from the \( P^n \)-Potts type potential to the mean field update of \( Q(x_i = l) \), where pixel \( i \) is in the
superpixel clique $s$, was derived in [48] as
\[
\sum_{\{x_s | x^{(l)}_s = l\}} Q(x_s - i) \psi_{\text{SP}}(x_s) = w_{\text{Low}}(l) \prod_{j \in s, j \neq i} Q(X_j = l) + w_{\text{High}} \left( 1 - \prod_{j \in s, j \neq i} Q(X_j = l) \right). \tag{8}
\]
This update operation is differentiable with respect to the parameters $w_{\text{Low}}(l)$ and $w_{\text{High}}$, allowing us to optimise them via backpropagation, and also with respect to the $Q(X)$ values enabling us to optimise previous layers in the network.

6. Experiments

We evaluate our new CRF formulation on two different datasets using the CRF-RNN network [53] as the main baseline, since we are essentially enriching the CRF model of [53]. We then present ablation studies on our models.

6.1. Experimental set-up and results

Our deep network consists of two conceptually different, but jointly trained stages. The first, “unary” part of our network is formed by the FCN-8s architecture [35]. It is initialised from the Imagenet-trained VGG-16 network [45], and then fine-tuned with data from the VOC 2012 training set [10], extra VOC annotations of [16] and the MS COCO [31] dataset.

The output of the first stage is fed into our CRF inference network. This is implemented using the mean field update operations and their differentials described in Section 5. Five iterations of mean field inference were performed during training. Our CRF network has two additional inputs in addition to segmentation unaries obtained from the FCN-8s network: data from the object detector and superpixel oversegmentations of the image.

We used the publicly available code and model of the Faster R-CNN [40] object detector. The fully automated version of GrabCut [42] was then used to obtain foregrounds from the detection bounding boxes. These choices were made after conducting preliminary experiments with alternate detection and foreground segmentation algorithms.

Four levels of superpixel oversegmentations were used, with increasing superpixel size to define the cliques used in this potential. Four levels were used since performance on the VOC validation set stopped increasing after this number. We used the superpixel method of [12] as it was shown to adhere to object boundaries the best [1], but our method generalises to any oversegmentation algorithm.

We trained the full network end-to-end, optimising the weights of the CNN classifier (FCN-8s) and CRF parameters jointly. We initialised our network using the publicly available weights of [53], and trained with a learning rate of $10^{-10}$. The learning rate is low because the loss was not normalised by the number of pixels in the training image. When training our CRF, we only used VOC 2012 data [10] as it has the most accurate labelling, particularly around boundaries.

6.1.1 PASCAL VOC 2012 Dataset

The improvement obtained by each higher order potential was evaluated on the same reduced validation set [35] used by our baseline [53]. As Table 1 shows, each new higher order potential improves the mean IoU over the baseline. We only report test set results for our best method since the VOC guidelines discourage the use of the test set for ablation studies. On the test set (Table 2), we outperform our baseline by 3.2% which equates to a 12.6% reduction in the error rate. This sets a new state-of-the-art on the VOC dataset. Qualitative results highlighting positive examples and failure cases of our algorithm are shown in our supplementary material.

6.1.2 PASCAL Context

Table 3 shows our state-of-the-art results on the recently released PASCAL Context dataset [36]. We trained on the provided training set of 4998 images, and evaluated our method on the validation set of 5105 images. This dataset augments VOC with annotations for all objects in the scene. As a result, there are 59 classes as opposed to the 20 in the VOC dataset. Many of these new labels are “stuff” classes such as “grass” and “sky”. Our object detectors are therefore only trained for 20 of the 59 labels in this dataset. Nevertheless, we improve by 0.8% over the previous state-of-the-art [6] and beat our baseline [53] by 2%.

6.2. Ablation Studies

We perform additional experiments to determine the errors made by our system, show the benefits of end-to-end training and compare our detection potentials to a simpler baseline. Unless otherwise stated, these experiments are performed on the VOC 2012 reduced validation set.

6.2.1 Error Analysis

To analyse the improvements made by our higher order potentials, we separately evaluate the performance on the “boundary” and “interior” regions in a similar manner to [6]. As shown in Fig. 5 c) and d), we consider a narrow band (trimap [21]) around the “void” labels annotated in the VOC 2012 reduced validation set. The mean IoU of pixels lying within this band is termed the “Boundary IoU” whilst the “Interior IoU” is evaluated outside this region.

Fig. 5 shows our results as the trimap width is varied. Adding the detection potentials improves the Interior IoU over our baseline (only pairwise potentials [53]) as the object
Table 1: Comparison of each higher order potential with baseline on VOC 2012 reduced validation set

| Method                          | Reduced val set(%) |
|--------------------------------|--------------------|
| Baseline (unary + pairwise) [53]| 72.9               |
| Superpixels only                | 74.0               |
| Detections only                 | 74.9               |
| Detections and Superpixels      | 75.8               |

Table 2: Mean IoU accuracy on VOC 2012 test set. All methods are trained with MS COCO [31] data

| Method                  | Test set(%) |
|-------------------------|-------------|
| Ours                    | 77.9        |
| DPN [34]                | 77.5        |
| Centrale Super Boundaries [22] | 75.7 |
| Dilated Convolutions [52] | 75.3        |
| BoxSup [6]              | 75.2        |
| DeepLab Attention [5]   | 75.1        |
| CRF-RNN (baseline) [53] | 74.7        |
| DeepLab WSSL [37]       | 73.9        |
| DeepLab [4]             | 72.7        |

Table 3: Mean Intersection over Union (IoU) results on PASCAL Context validation set compared to other current methods.

| Method                  | Ours | BoxSup [6] | ParseNet [33] | CRF-RNN [53] | FCN-8s [35] | CFM [7] |
|-------------------------|------|------------|---------------|--------------|-------------|--------|
| Mean IoU (%)            | 41.3 | 40.5       | 40.4          | 39.3         | 37.8        | 34.4   |

detector may recognise objects in the image which the pixel-wise classification stage of our network may have missed out. However, the detection potentials also improve the Boundary IoU for all tested trimap widths as well. Improving the recognition of pixels in the interior of an object also helps with delineating the boundaries since the strength of the pairwise potentials exerted by the $Q$ distributions at each of the correctly-detected pixels increase.

Our superpixel priors also increase the Interior IoU with respect to the baseline. Encouraging consistency over regions helps to get rid of spurious regions of wrong labels (as shown in Fig. 4). Fig. 5 suggests that most of this improvement occurs in the interior of an object. The Boundary IoU is slightly lower than the baseline, and this may be due to the fact that superpixels do not always align correctly with the edges of an object (the “boundary recall” of various superpixel methods are evaluated in [1]).

We can see that the combination of detection and superpixel potentials results in a substantial improvement in our Interior IoU. This is the primary reason our overall IoU on the VOC benchmark increases with higher order potentials.

6.2.2 Benefits of end-to-end training

Table 4 shows how end-to-end training outperforms piecewise training. We trained the CRF piecewise by freezing the weights of the unary part of the network, and only learning the CRF parameters.

Our results in Table 2 used the FCN-8s [35] architecture to generate unaries. To show that our higher order potentials improve performance regardless of the underlying CNN used for producing unaries, we also perform an experiment using our reimplementation of the “front-end” module proposed in the Dilated Convolution Network (DCN) of [52] instead of FCN-8s.

Table 4 shows that end-to-end training of the CRF yields considerable improvements over piecewise training. This was the case when using either FCN-8s or DCN for obtaining the initial unaries before performing CRF inference with higher order potentials. This suggests that our CRF network module can be plugged into different architectures and achieve performance improvements.

6.2.3 Baseline for detections

To evaluate the efficacy of our detection potentials, we formulate a simpler baseline since no other methods use detection information at inference time (BoxSup [6] derives ground truth for training using ground-truth bounding boxes).

Our baseline is similar to CRF-RNN [53], but prior to CRF inference, we take the segmentation mask from the
object detection and add a unary potential proportional to the detector’s confidence to the unary potentials for those pixels. We then perform mean-field inference (with only pairwise terms) on these “augmented” unaries. Using this method, the mean IoU increases from 72.9% to 73.6%, which is significantly less than the 74.9% which we obtained using only our detection potentials without superpixels (Table 1).

Our detection potentials perform better since our latent $Y$ detection variables model whether the detection hypothesis is accepted or not. Our CRF inference is able to evaluate object detection inputs in light of other potentials. Inference increases the relative score of detections which agree with the segmentation, and decreases the score of detections which do not agree with other energies in the CRF. Figures 2 b) and d) show examples of false-positive detections that have been ignored and correct detections that have been used to refine our segmentation. Our baseline, on the other hand, is far more sensitive to erroneous detections as it cannot adjust the weight given to them during inference.

7. Conclusion

We presented a CRF model with two different higher order potentials to tackle the semantic segmentation problem. The first potential is based on the intuitive idea that object detection can provide useful cues for semantic segmentation. Our formulation is capable of automatically rejecting false object detections that do not agree at all with the semantic segmentation. Secondly, we used a potential that encourages superpixels to have consistent labelling. These two new potentials can co-exist with the usual unary and pairwise potentials in a CRF.

Importantly, we showed that efficient mean field inference is still possible in the presence of the new higher order potentials and derived the explicit forms of the mean field updates and their differentials. This enabled us to implement the new CRF model as a stack of CNN layers and to train it end-to-end in a unified deep network with a pixel-wise CNN classifier. We experimentally showed that the addition of higher order potentials results in a significant increase in semantic segmentation accuracy allowing us to reach state-of-the-art performance.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. PAMI, 34(11):2274–2282, 2012.

[2] D. Batra, R. Sukthankar, and T. Chen. Learning class-specific affinities for image labelling. In CVPR, pages 1–8. IEEE, 2008.

[3] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In ECCV, pages 430–443. 2012.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. ICLR, 2015.

[5] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. arXiv preprint arXiv:1511.03339, 2015.
[6] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In ICCV, 2015.

[7] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. CVPR, 2015.

[8] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshthkhar, and G. Mori. Deep structured models for group activity recognition. In BMVC, 2015.

[9] J. Domke. Learning graphical model parameters with approximate marginal inference. PAMI, 2013.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010.

[11] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. PAMI, 2013.

[12] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. IJCV, 2004.

[13] R. Girshick. Fast r-cnn. In ICCV, 2015.

[14] J. M. Gonfaus, X. Boix, J. Van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez. Harmony potentials for joint classification and segmentation. In CVPR, pages 3280–3287. IEEE, 2010.

[15] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In NIPS, pages 665–663, 2009.

[16] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In IJCV, pages 991–998. IEEE, 2011.

[17] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In ECCV, pages 297–312. Springer, 2014.

[18] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In CVPR, volume 2, pages II–695. IEEE, 2009.

[19] M. Kiefel and P. V. Gehler. Human pose estimation with fields of parts. In ECCV, pages 331–346. Springer, 2014.

[20] P. Kohli, M. P. Kumar, and P. H. Torr. P3 & beyond: Solving energies with higher order cliques. In CVPR, 2007.

[21] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. IJCV, 82(3):302–324, 2009.

[22] I. Kokkinos. Pushing the boundaries of boundary detection using deep learning. In ICLR, 2016.

[23] D. Koller and N. Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

[24] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In NIPS, 2011.

[25] P. Krähenbühl and V. Koltun. Parameter learning and convergent inference for dense random fields. In ICML, 2013.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1097–1105, 2012.

[27] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In ICCV, pages 739–746, 2009.

[28] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence statistics. In ECCV, pages 239–253, 2010.

[29] L. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and crfs. In ECCV, pages 424–437, 2010.

[30] G. Lin, C. Shen, I. Reid, et al. Efficient piecewise training of deep structured models for semantic segmentation. arXiv preprint arXiv:1504.01013, 2015.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755. Springer, 2014.

[32] B. Liu and X. He. Multiclass semantic video segmentation with object-level active inference. In CVPR, pages 4286–4294, 2015.

[33] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579, 2015.

[34] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In ICCV, 2015.

[35] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[36] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, et al. The role of context for object detection and semantic segmentation in the wild. In CVPR, pages 891–898. IEEE, 2014.

[37] G. Papandreou, L. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. In ICCV, 2015.

[38] X. Qi, J. Shi, S. Liu, R. Liao, and J. Jia. Semantic segmentation with object clique potential. In Proceedings of the IEEE International Conference on Computer Vision, pages 2587–2595, 2015.

[39] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In ICCV, pages 1–8, 2007.

[40] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[41] S. Ross, D. Munoz, M. Hebert, and J. A. Bagnell. Learning message-passing inference machines for structured prediction. In CVPR, 2011.

[42] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. ACM TOG, 2004.

[43] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. arXiv preprint arXiv:1503.02351, 2015.

[44] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. IJCV, 2009.

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[46] M. Sun, B.-s. Kim, P. Kohli, and S. Savarese. Relating things and events by jointly modeling texture, layout, and context. In CVPR, 2016.

[47] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In NIPS, pages 1799–1807, 2014.

[48] V. Vineet, J. Warrell, and P. H. Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. IJCV, 2014.

[49] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In ECCV, pages 733–747. Springer, 2008.

[50] Y. Yang, S. Hallman, D. Ramanan, and C. C. Fowlkes. Layered object models for image segmentation. PAMI, 2012.

[51] J. Yao, S. Fidler, and R. Urtasun. Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation. In CVPR, pages 702–709, 2012.

[52] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In ICLR, 2016.

[53] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In ICCV, 2015.