Beyond Gradient Descent for Regularized Segmentation Losses

Dmitrii Marin
University of Waterloo
Canada

Meng Tang
University of Waterloo
Canada

Ismail Ben Ayed
ETS Montreal
Canada

Yuri Boykov
University of Waterloo
Canada

Abstract

The simplicity of gradient descent (GD) made it the default method for training ever-deeper and complex neural networks. Both loss functions and architectures are often explicitly tuned to be amenable to this basic local optimization. In the context of weakly-supervised CNN segmentation, we demonstrate a well-motivated loss function where an alternative optimizer (ADM)[1] achieves the state-of-the-art while GD performs poorly. Interestingly, GD obtains its best result for a "smoother" tuning of the loss function. The results are consistent across different network architectures. Our loss is motivated by well-understood MRF/CRF regularization models in "shallow" segmentation and their known global solvers. Our work suggests that network design/training should pay more attention to optimization methods.

1. Motivation and Background

Weakly supervised training of neural networks is often based on *regularized losses* combining an empirical loss with some regularization term, which compensates for lack of supervision [38, 14]. Regularized losses are also useful for CNN segmentation [32, 34] where full supervision is often infeasible, particularly in biomedical applications. Such losses are motivated by regularization energies in *shallow*[2] segmentation, where multi-decade research went into designing robust regularization models based on geometry [24, 7, 5], physics [18, 1], or robust statistics [13]. Such models should represent realistic shape priors compensating for image ambiguities, yet be amenable to efficient solvers. Many robust regularizers commonly used in vision [31, 17] are non-convex and require powerful optimizers to avoid many weak local minima. Basic local optimizers typically fail to produce practically useful results with such models.

Effective weakly-supervised CNN methods for vision should incorporate priors compensating for image data ambiguities and lack of supervision just as in shallow vision methods. For example, recent work [38, 34] formulated

the problems of semi-supervised classification and weakly-supervised segmentation as minimization of regularized losses. This principled approach outperforms common "proposal generation" methods [23, 20] computing "fake" ground truths to mimic standard fully-supervised training. However, we show that the use of regularization models as losses in deep learning is limited by GD, the backbone optimizer in current training methods. It is well-known that GD leads to poor local minima for many regularizers in shallow segmentation and many stronger algorithms were proposed [4, 6, 21, 31, 15]. Similarly, we show better optimization beyond GD for regularized losses in deep segmentation.

One popular general approach applicable to regularized losses is ADMM [3] that splits optimization into two efficiently solvable sub-problems separately focusing on the empirical loss and regularizer. We advocate similar *splitting* to improve optimization of regularized losses in CNN training. In contrast, ADMM-like splitting of network parameters in different layers was used in [35] to improve parallelism.

In our work weakly-supervised CNN segmentation is a context for discussing regularized loss optimization. As a regularizer, we use the common Potts model [6] and consider its nearest- and large-neighborhood variants, a.k.a. sparse *grid CRF* and *dense CRF* models. We show effectiveness of ADMM-like splitting for grid CRF losses due to availability of powerful sub-problem solvers, *e.g.* graph cuts [5]. As detailed in [34, Sec.3], an earlier iterative proposal-generation technique by [20] can be related to regularized loss splitting, but their method is limited to dense CRF and its approximate *mean-field* solver [22]. In fact, given such weak sub-problem solvers, splitting is inferior to basic GD over the regularized loss [34]. More insights on grid and dense CRF are below.

1.1. Pairwise CRF for Shallow Segmentation

Robust pairwise Potts model and its binary version (Ising model) are used in many application such as stereo, reconstruction, and segmentation. One can define this model as a cost functional over integer-valued labeling $S := (S_p \in Z^+ \mid p \in \Omega)$ of image pixels $p \in \Omega$ as follows

$$E_P(S) \;=\; \sum_{pq \in \mathcal{N}} w_{pq} \cdot [S_p \neq S_q] \qquad (1)$$

---

[1] https://github.com/dmitrii-marin/adm-seg
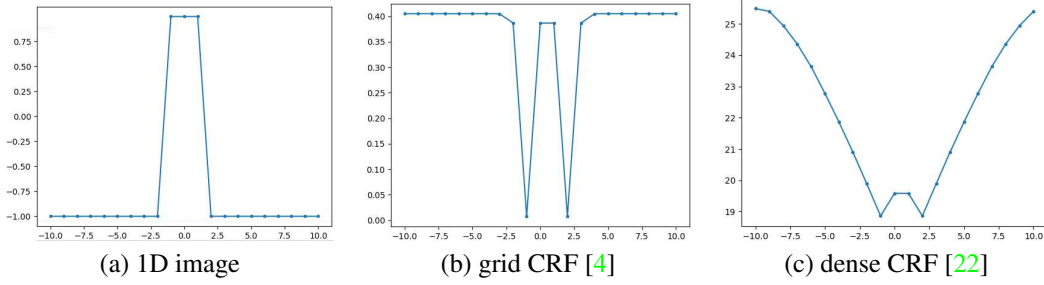[2] In this paper, "shallow" refers to methods unrelated to deep learning.

Figure 1. *Synthetic segmentation example* for grid and dense CRF (Potts) models: (a) intensities $I(x)$ on 1D image. The cost of segments $S^t = \{x \mid x < t\}$ with different discontinuity points t according to (b) nearest-neighbor (grid) Potts and (c) larger-neighborhood (dense) Potts. The latter gives smoother cost function, but its flatter minimum may complicate discontinuity localization.
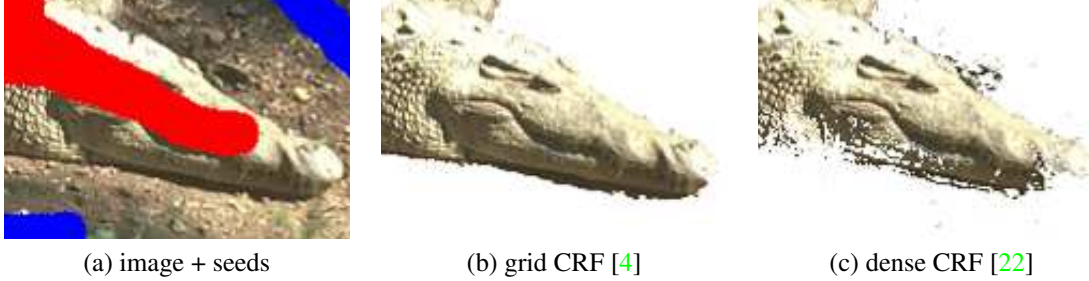

Figure 2. *Real "shallow" segmentation example* for sparse (b) and dense (c) CRF (Potts) models for image with seeds (a). Sparse Potts gives smoother segment boundary with better edge alignment, while dense CRF inference often gives noisy boundary.

where $\mathcal{N}$ is a given neighborhood system, w_{pq} is a discontinuity penalty between neighboring pixels $\{p, q\}$, and $[\cdot]$ is *Iverson bracket*. The nearest-neighbor version over k-connected grid $\mathcal{N}_k$, as well as its popular variational analogues, *e.g.* geodesic active contours [7], convex relaxations [27, 9], or continuous max-flow [39], are particularly well-researched. It is common to use contrast-weighted discontinuity penalties [6, 4] between the neighboring points, as emphasized by the condition $\{pq\} \in \mathcal{N}_k$ below

$$w_{pq} = \lambda \cdot \exp \frac{-||I_p - I_q||^2}{2\sigma^2} \cdot [\{pq\} \in \mathcal{N}_k]. \quad (2)$$

Nearest neighbor Potts models minimize the contrast-weighted length of the segmentation boundary preferring shorter perimeter aligned with image edges, *e.g.* see Fig. 2(b). The popularity of this model can be explained by generality, robustness, well-established foundations in geometry, and a large number of efficient discrete or continuous solvers that guarantee global optimum in binary problems [4] or some quality bound in multi-label settings, *e.g.* α-expansion [6].

Dense CRF [22] is a Potts model where pairwise interactions are active over significantly bigger neighborhoods defined by a Gaussian kernel with a relatively large bandwidth Δ over pixel locations

$$w_{pq} = \lambda \cdot \exp \frac{-||I_p - I_q||^2}{\sigma^2} \cdot \exp \frac{-||p - q||^2}{\Delta^2}. \quad (3)$$

Its use in shallow vision is limited as it often produces noisy boundaries [22], see also Fig. 2(c). Also, global optimization

methods mentioned above do not scale to dense neighborhoods. Yet, dense CRF model is popular in the context of CNNs where it can be used as a differentiable regularization layer [41, 29]. Larger bandwidth yields smoother objective (1), see Fig. 1(c), amenable to gradient descent or other local linearization methods like mean-field inference that are easy to parallelize. Note that existing efficient inference methods for dense CRF require *bilateral filtering* [22], which is restricted to Gaussian weights as in (3). This is in contrast to global Potts solvers, *e.g.* α-expansion, that can use arbitrary weights, but become inefficient for dense neighborhoods.

Noisier dense CRF results, *e.g.* in Fig. 2(c), imply weaker regularization. Indeed, as discussed in [36], for larger neighborhoods the Potts model gets closer to *cardinality potentials*. Bandwidth Δ in (3) is a resolution scale at which the model sees the segmentation boundary. Weaker regularization in dense CRF may preserve some thin structures smoothed out by fine-resolution boundary regularizers, *e.g.* nearest-neighbor Potts. However, this is essentially the same "noise preservation" effect shown in Fig. 2(c). For consistency, the rest of the paper refers to the nearest-neighbor Potts model as *grid CRF*, and large-neighborhood Potts as *dense CRF*.

1.2. Summary of Contributions

Any motivation for standard regularization models in shallow image segmentation, as in the previous section, directly translates into their motivation as regularized loss functions in weakly supervised CNN segmentation [32, 34]. The main

issue is how to optimize these losses. Standard training techniques based on gradient descent may not be appropriate for many powerful regularization models, which may have many local minima. Below is the list of our main contributions:

- As an alternative to gradient descent (GD), we propose a splitting technique, *alternating direction method* (ADM)[3], for minimizing *regularized losses* during network training. ADM can directly employ efficient regularization solvers known in shallow segmentation.

- Compared to GD, our ADM approach with α-expansion solver significantly improves optimization quality for the *grid CRF* (nearest-neighbor Potts) loss in weakly supervised CNN segmentation. While each iteration of ADM is slower than GD, the loss function decreases at a significantly larger rate with ADM. In one step it can reach lower loss values than those where GD converges. Grid CRF has never been investigated as loss for CNN segmentation and is largely overlooked.

- The training quality with grid CRF loss achieves the-state-of-the-art in weakly supervised CNN segmentation. We compare dense CRF and grid CRF losses.

Our results may inspire more research on regularized segmentation losses and their optimization.

2. ADM for Regularized Loss Optimization

Assume there is a dataset of pairs of images and partial ground truth labelings. For simplicity of notation, we implicitly assume summation over all pairs in dataset for all expressions of loss functions. For each pixel $p \in \Omega$ of each image I there is an associated color or intensity I_p of that pixel. The labeling $Y = (Y_p | p \in \Omega_{\mathcal{L}})$ where $\Omega_{\mathcal{L}} \subset \Omega$ is a set of labeled pixels, each $Y_p \in \{0,1\}^K$ is a one-hot distribution and K is the number of labels. We consider a regularized loss for network ϕ_θ of the following form

$$\ell(S_\theta, Y) \; + \; \lambda \cdot R(S_\theta) \quad \rightarrow \quad \min_\theta \qquad (4)$$

where $S_\theta \in [0,1]^{|\Omega| \times K}$ is a K-way softmax segmentation generated by the network $S_\theta := \phi_\theta(I)$, and $R(\cdot)$ is a regularization term, *e.g. relaxed* sparse Potts or dense CRF, and $\ell(\cdot, \cdot)$ is a partial ground-truth loss, for instance:

$$\ell(S_\theta, Y) \; = \; \sum_{p \in \Omega_{\mathcal{L}}} H(Y_p, S_{p,\theta}),$$

where $H(Y_p, S_{p,\theta}) = -\sum_k Y_p^k \log S_{p,\theta}^k$ is the cross entropy between predicted segmentation $S_{p,\theta}$ (a row of matrix S_θ corresponding to pixel p) and ground truth labeling Y_p.

---

[3]Standard ADM*M* [3] casts a problem $\min_x f(x) + g(x)$ into $\min_{x,y} \max_\lambda f(x) + g(y) + \lambda^{\mathrm{T}}(x - y) + \rho\|x - y\|^2$ and alternates updates over x, y and λ optimizing f and g in parallel. Our ADM uses a different form of splitting and can be seen as a *penalty method*, see Sec. 2.

We present a general alternating direction method (ADM) to optimize neural network regularized losses of the general form in (4) using the following splitting of the problem:

$$\begin{aligned} \underset{\theta, X}{\text{minimize}} \quad & \ell(S_\theta, Y) \; + \; \lambda R(X) \\ \text{subject to} \quad & \sum_{p \in \Omega_{\mathcal{U}}} D(X_p | S_{p,\theta}) = 0, \\ & X_p = Y_p \qquad \forall p \in \Omega_{\mathcal{L}} \end{aligned} \qquad (5)$$

where we introduced one-hot distributions $X_p \in \{0,1\}^K$ and some divergence measure D, *e.g.* the Kullback-Leibler divergence. $R(X)$ can now be a *discrete* classic MRF regularization, *e.g.* (1). This equates to the following Lagrangian

$$\min_{\theta, X} \max_\gamma \quad \ell(S_\theta, Y) + \lambda R(X) + \gamma \sum_{p \in \Omega_{\mathcal{U}}} D(X_p | S_{p,\theta})$$
$$\text{subject to} \quad X_p = Y_p \quad \forall p \in \Omega_{\mathcal{L}}. \qquad (6)$$

We alternate optimization over X and θ in (6). The maximization over γ increases its value at every update resulting in a variant of simulated annealing. We have experimented with variable multiplier γ but found no advantage compared to fixed γ. So, we fix γ and do not optimize for it. In summary, instead of optimizing the regularization term with gradient descent, our approach splits regularized-loss problem (4) into two sub-problems. We replace the softmax outputs $S_{p,\theta}$ in the regularization term by latent discrete variables X_p and ensure consistency between both variables (*i.e.*, S_θ and X) by minimizing divergence D.

This is similar conceptually to the general principles of ADMM [3, 37]. Our ADM splitting accommodates the use of powerful and well-established discrete solvers for the regularization loss. As we show in Sec. 3, the popular α-expansion solver [6] significantly improves optimization of grid CRF losses yielding state-of-the-art training quality. Such efficient solvers guarantee global optimum in binary problems [4] or a quality bound in multi-label case [6].

Our discrete-continuous ADM method alternates two steps, each decreasing (6), until convergence. Given fixed discrete latent variables X_p computed at the previous iteration, the first step learns the network parameters θ by minimizing the following loss via standard back-propagation and a variant of stochastic gradient descent (SGD):

$$\underset{\theta}{\text{minimize}} \; \ell(S_\theta, Y) \; + \; \gamma \sum_{p \in \Omega_{\mathcal{U}}} D(X_p | S_{p,\theta}) \qquad (7)$$

The second step fixes the network output S_θ and finds the next latent binary variables X by minimizing the following objective over X via any suitable discrete solver:

$$\begin{aligned} \underset{X \in \{0,1\}^{|\Omega| \times K}}{\text{minimize}} \quad & \lambda R(X) \; + \; \gamma \sum_{p \in \Omega_{\mathcal{U}}} D(X_p | S_{p,\theta}) \\ \text{subject to} \quad & X_p = Y_p \quad \forall p \in \Omega_{\mathcal{L}}. \end{aligned} \qquad (8)$$

Because X_p is a discrete variable with only K possible values, the second term in (8) is a basic unary term. Similarly, the equality constraints could be implemented as unary terms using prohibitive values of unary potentials. Unary terms are simplest possible energy potentials that can be handled by any general discrete solver. On the other hand, the regularization term $R(X)$ usually involves interactions of two or more variables introducing new properties of solution together with optimization complexity. In case of grid CRF one can use graph cut [4], α-expansion [6], QPBO [2, 28], TRWS [21], LBP [26], LSA-TR [15] *etc*.

In summary, our approach alternates the two steps described above. For each minibatch we compute network prediction, then compute hidden variables X optimizing (8), then compute gradients of loss (7) and update the parameters of the network using a variant of SGD. The outline of our ADM scheme is shown in Alg. 1.

Algorithm 1 ADM for regularized loss (4).

Require: sequence of minibatches

  $i \leftarrow 0$;

  initialize network parameters $\theta^{(0)}$;

  for each minibatch B **do**

    for each image-labeling pair $(I, Y) \in B$ **do**

      Compute segmentation prediction $S_\theta \leftarrow \phi_{\theta^{(i)}}(I)$;

      Solve energy (8) for X with *e.g.* α-expansion;

      Compute gradients g w.r.t. parameters θ of (7);

    end for

    Compute average over the batch $g^{(i)} \leftarrow \frac{1}{|B|} \sum g$;

    Update network parameters $\theta^{(i+1)}$ using gradient $g^{(i)}$;

    $i \leftarrow i + 1$;

  end for

---

3. Experimental Results

We conduct experiments for weakly supervised CNN segmentation with scribbles as supervision [23]. The focus is on regularized loss approaches [32, 34] yet we also compare our results to proposal generation based method, *e.g.* ScribbleSup [23]. We test both the grid CRF and dense CRF as regularized losses. Such regularized loss can be optimized by stochastic gradient descent (GD) or alternative direction method (ADM), as discussed in Sec. 2. We compare three training schemes, namely dense CRF with GD [34], grid CRF with GD and grid CRF with ADM for weakly supervised CNN segmentation.

Before comparing segmentations, in Sec. 3.1 we test if using ADM gives better regularization losses than that using standard GD. Our plots of training losses (CRF energy) vs training iterations show how fast the losses converge when minimized by ADM or GD. Our experiment confirms that first order approach like GD leads to a poor local minimum

for the grid CRF loss. There are clear improvements of ADM over GD for minimization of the grid CRF loss. In Sec. 3.2, rather than comparing in terms of optimization, we compare ADM and GD in terms of segmentation quality. We report both mIOU (mean intersection over union) and accuracy in particular for boundary regions. In Sec. 3.3, we also study these variants of regularized loss method in a more challenging setting of shorter scribbles [23] or clicks in the extreme case. With ADM as the optimizer, our approach of grid CRF regularized loss compares favorably to dense CRF based approach [34].

Dataset and implementation details Following recent work [10, 23, 20, 32] on CNN semantic segmentation, we report our results on PASCAL VOC 2012 segmentation dataset. We train with scribbles from [23] on the augmented datasets of 10,582 images and test on the *val* set of 1,449 images. We report mIOU (mean intersection over union) and pixel-wise accuracy. In particular, we are interested in how good is the segmentation in the boundary region. So we compute accuracy for those pixels close to the boundary, for example within 8 or 16 pixels from semantic boundaries. Besides mIOU and accuracy, we also measure the regularization losses, *i.e.* the grid CRF. Our implementation is based on DeepLabv2[4] and we show results on different networks including deeplab-largeFOV, deeplab-msc-largeFOV, deeplab-vgg16 and resnet-101. We do not apply any post-processing to network output segmentation.

The networks are trained in two phases. Firstly, we train the network to minimize partial cross entropy (pCE) loss w.r.t scribbles. Then we train with a grid CRF or dense CRF regularization term. To implement gradient descent for the discrete grid CRF loss, we first take its quadratic relaxation,

$$\langle \mathbf{1}, S_{p,\theta} \rangle + \langle \mathbf{1}, S_{q,\theta} \rangle - 2 \langle S_{p,\theta}, S_{q,\theta} \rangle. \tag{9}$$

where $S_{p,\theta}, S_{q,\theta} \in [0,1]^K$ and $\langle \cdot, \cdot \rangle$ is the dot product. Then we differentiate w.r.t. S_θ during back-propagation. While there are ways, *e.g.* [11, 8], to relax discrete Pott's model, we focus on this simple and standard relaxation [40, 22, 34].

For our proposed ADM algorithm, which requires inference in the grid CRF, we use a public implementation of α-expansion[5]. The CRF inference and loss are implemented and integrated as Caffe [16] layers. We run α-expansion for five iterations, which in most cases gives convergence. Our dense CRF loss does not include the Gaussian kernel on locations XY, since ignoring this term does not change the mIOU measure [22]. The bandwidth for the dense Gaussian kernel on $RGBXY$ is validated to give the best mIOU. For the grid CRF, the kernel bandwidth selection in (2) follows

---

[4]https://bitbucket.org/aquariusjay/deeplab-public-ver2

[5]http://mouse.cs.uwaterloo.ca/code/gco-v3.0.zip

network	training set†		validation set	
	GD	ADM	GD	ADM
Deeplab-LargeFOV	2.52	2.41	2.51	2.33
Deeplab-MSc-largeFOV	2.51	2.40	2.49	2.33
Deeplab-VGG16	2.37	2.10	2.42	2.14
Resnet-101	2.66	2.49	2.61	2.42

Table 1. ADM gives better grid CRF losses than gradient descend (GD). †We randomly selected 1,000 training examples.

standard Boykov-Jolly [4]

$$\sigma^2 \;\; = \;\; \frac{1}{|N|} \sum_{pq \in N} \|I_p - I_q\|^2.$$

In general, our ADM optimization for regularized loss is slower than GD due to the inference of grid CRF. However, for inference algorithms, *e.g.* α-expansion, that cannot be easily parallelized, we utilize simple multi-core parallelization for all images in a batch to accelerate training. Note that we do not use CRF inference during testing.

3.1. Loss Minimization

In this section, we show that for grid CRF losses the ADM approach employing α-expansion [6], a powerful discrete optimization method, outperforms common gradient descend methods for regularized losses [32, 34] in terms of finding a lower minimum of regularization loss. Tab. 1 shows the grid CRF losses on both training and validation sets for different network architectures. Fig. 3(a) shows the evolution of the grid CRF loss over the number of iterations of training. ADM requires fewer iterations to achieve the same CRF loss. The networks trained using ADM scheme give lower CRF losses for both training and validation sets.

The gradients with respect to the soft-max layer's input of the network are visualized in Fig. 4. Clearly, our ADM approach with the grid CRF enforces better edge alignment. Despite different formulations of regularized losses and their optimization, the gradients of either (4) or (7) w.r.t. network output S_θ are the driving force for training. In most of the cases, GD produces significant gradient values only in the vicinity of the current model prediction boundary as in Fig. 4(c,d). If the actual object boundary is sufficiently distant the gradient methods fail to detect it due to the sparsity of the grid CRF model, see Fig. 1 for an illustrative "toy" example. On the other hand, the ADM method is able to predict a good latent segmentation allowing gradients leading to a good solution more effectively, see Fig. 4(e).

Thus, in the context of grid CRFs, the ADM approach coupled with α-expansion shows drastic improvement in the optimization quality. In the next section, we further compare ADM with GD to see which gives better segmentation.
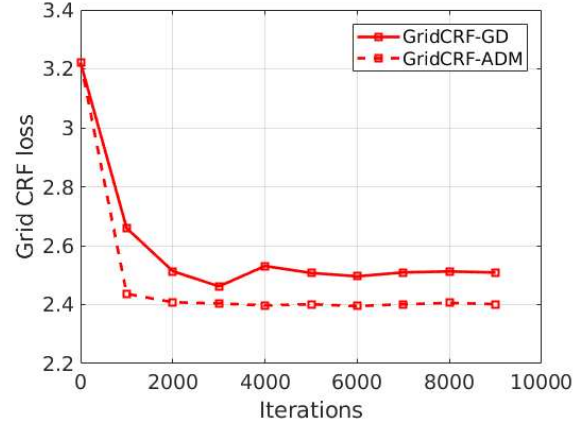

Figure 3. Training progress of ADM and gradient descend (GD) on Deeplab-MSc-largeFOV. Our ADM for the grid CRF loss with α-expansion significantly improves convergence and achieves lower training loss. For example, first 1,000 iterations of ADM give grid CRF loss lower than GD's best result.

3.2. Segmentation Quality

The quantitative measures for segmentation by different methods are summarized in Tab. 2 and Tab. 3. The mIOU and segmentation accuracy on the *val* set of PASCAL 2012 [12] are reported for various networks. The supervision is scribbles [23]. The quality of weakly supervised segmentation is bounded by that with full supervision and we are interested in the gap for different weakly supervised approaches.

The baseline approach is to train the network using proposals generated by GrabCut style interactive segmentation with such scribbles. Besides the baseline (train w/ proposals), here we compare variants of regularized losses optimized by gradient descent or ADM. The regularized loss is comprised of the partial cross entropy (pCE) w.r.t. scribbles and grid/dense CRF. Other losses *e.g.* normalized cut [30, 32] may give better segmentation, but the focus is to compare gradient descent vs ADM optimization for the grid CRF.

It is common to apply dense CRF post-processing [10] to the network's output during testing. However, for the sake of clear comparison, we show results without it.

As shown in Tab. 2, all regularized approaches work better than the non-regularized approach that only minimizes the partial cross entropy. Also, the regularized loss approaches are much better than proposal generation based method since erroneous proposals may mislead training.

Among regularized loss approaches, grid CRF with GD performs the worst due to the fact that a first-order method like gradient descent leads to the poor local minimum for the grid CRF in the context of energy minimization. Our ADM for the grid CRF gives much better segmentation competitive with the dense CRF with GD. The alternative grid CRF based method gives good quality segmentation approaching that for

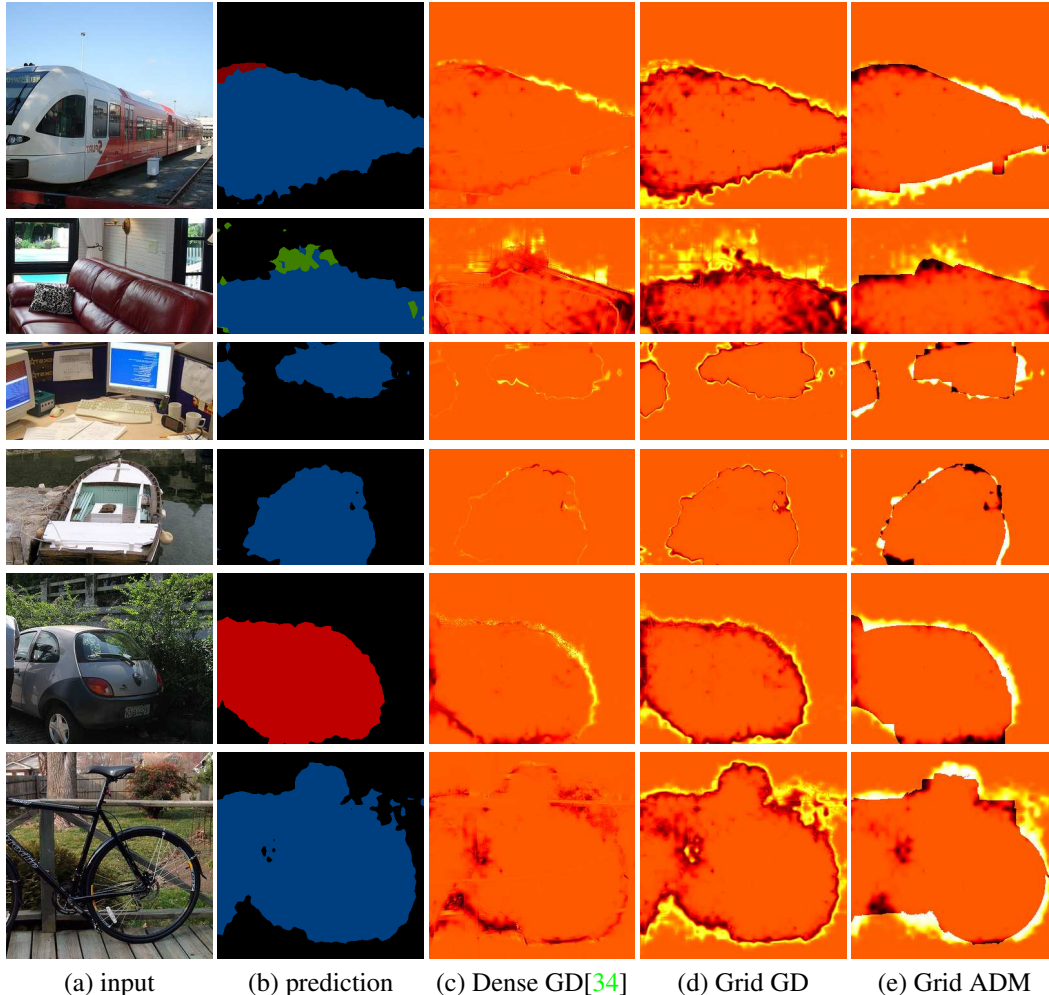| (a) input | (b) prediction | (c) Dense GD[34] | (d) Grid GD | (e) Grid ADM |

Figure 4. The gradients with respect to scores of the deeplab_largeFOV network with the dense CRF (c) and grid CRF (d and e for using either the plain stochastic gradient descent or our ADM scheme). Latent segmentation in ADM with the grid CRF loss produces gradients more directly pointing to a good solution (e). Note, the object boundaries are more prominent in (e).

full supervision. Tab. 3 shows *accuracy* of different methods for pixels close to the semantic boundaries. Such measure tells the quality of segmentation in boundary regions.

Fig. 5 shows a few qualitative segmentation results.

3.3. Shortened Scribbles

Following the evaluation protocol in ScribbleSup [23], we also test our regularized loss approaches training with shortened scribbles. We shorten the scribbles from the two ends at certain ratios of length. In the extreme case, scribbles degenerate to clicks for semantic objects. We are interested in how the weakly-supervised segmentation methods degrade as we reduce the length of the scribbles. We report both mIOU and pixel-wise accuracy. As shown in Fig. 6, our ADM for the grid CRF loss outperforms all competitors giving significantly better mIOU and accuracy than GD for the grid CRF loss. ADM degrades more gracefully than the dense CRF as the supervision weakens.

The grid CRF has been overlooked in regularized CNN segmentation currently dominated by the dense CRF as either post-processing or trainable layers. We show that for weakly supervised CNN segmentation, the grid CRF as the regularized loss can give segmentation at least as good as that with the dense CRF. The key to minimizing the grid CRF loss is better optimization via ADM rather than gradient descent. Such competitive results for the grid CRF loss confirm that it has been underestimated as a loss regularizer for neural network training, as discussed in Sec. 1.

It has not been obvious whether the grid CRF as a loss is beneficial for CNN segmentation. We show that straightforward gradient descent for the grid CRF does not work well. Our technical contribution on optimization helps to reveal the limitation and advantage of the grid CRF vs dense CRF models. The weaker regularization properties, as discussed in Sec. 1.1, of the dense CRF and our experiments favors the grid CRF regularizer compared to the dense CRF.

Network	Full sup.	Weak supervision				
		train w/ proposals	pCE loss	+dense CRF loss GD [34]	+grid CRF loss GD	ADM
Deeplab-largeFOV	63.0	54.8	55.8	**62.2**	60.4	61.7
Deeplab-MSc-largeFOV	64.1	55.5	56	**63.1**	61.2	62.9
Deeplab-VGG16	68.8	59.0	60.4	64.4	63.3	**65.2**
ResNet-101	75.6	64.0	69.5	**72.9**	71.7	72.8

Table 2. Weakly supervised segmentation results for different choices of network architecture, regularized losses and optimization via gradient descent or ADM. We show mIOU on *val* set of PASCAL 2012. ADM consistently improves over GD for different networks for grid CRF. Our grid CRF with ADM is competitive to previous state-of-the-art dense CRF (with GD) [34].

	Network	Full sup.	Weak supervision				
			train w/ proposals	pCE loss	+dense CRF loss GD [34]	+grid CRF loss GD	ADM
all pixels	Deeplab-MSc-largeFOV	90.9	86.4	86.5	**90.6**	89.9	90.5
	Deeplab-VGG16	91.6	88.6	88.9	91.1	90.5	**91.3**
	ResNet-101	94.5	90.2	92	93.1	92.9	**93.4**
trimap 16 pixels	Deeplab-MSc-largeFOV	80.1	73.9	66.7	**77.8**	74.8	76.7
	Deeplab-VGG16	81.9	75.5	70.9	77.8	75.6	**78.1**
	ResNet-101	85.7	78.4	77.7	82.0	80.6	**82.2**
trimap 8 pixels	Deeplab-MSc-largeFOV	75.0	69.5	60.3	**72.5**	68.4	71.4
	Deeplab-VGG16	76.9	70.4	64.1	72.0	69.0	**72.4**
	ResNet-101	81.5	73.8	71.2	76.7	74.6	**77.0**

Table 3. Pixel-wise accuracy on *val* set of PASCAL 2012. Top 3 rows: accuracy over all pixels. Middle 3 rows: accuracy for pixels within 16 pixels away from semantic boundaries. Bottom 3 rows: accuracy for pixels within 8 pixels aways from semantic boundaries. Pixels closer to boundaries are more likely to be mislabeled. Our ADM scheme improves over GD for grid CRF loss consistently for different networks. Note that weak supervision with our approach is almost as good as full supervision.

4. Conclusion

Gradient descent (GD) is the default method for training neural networks. Often, loss functions and network architectures are designed to be amenable to GD. The top-performing weakly-supervised CNN segmentation [32, 34] is trained via regularized losses, as common in weakly-supervised deep learning [38, 14]. In general, GD allows any differentiable regularizers. However, in shallow image segmentation it is know that generic GD is a substandard optimizer for (relaxations of) standard robust regularizers, *e.g.* grid CRF.

Here we propose a general splitting technique, ADM, for optimizing regularized losses. It can take advantage of many existing efficient regularization solvers known in shallow segmentation. In particular, for grid CRF our ADM approach using α-expansion solver achieves significantly better optimization quality compared to GD. With such ADM optimization, training with grid CRF loss achieves the-state-of-the-art in weakly supervised CNN segmentation. We systematically

compare grid CRF and dense CRF losses from modeling and optimization perspectives. Using ADM optimization, the grid CRF loss achieves CNN training favourably comparable to the best results with the dense CRF loss. Our work suggests that in the context of network training more attention should be paid to optimization methods beyond GD.

In general, our ADM approach applies to many regularized losses, as long as there are efficient solvers for the corresponding regularizers. This work is focused on ADM in the context of common pairwise regularizers. Interesting future work is to investigate losses with non-Gaussian pairwise CRF potentials and higher-order segmentation regularizers, *e.g.* P^n Potts model [19], curvature [25], and kernel clustering [30, 33]. Also with ADM framework, we can explore other optimization methods [17] besides α-expansion for various kinds of regularized losses in segmentation. Our work bridges optimization method in "shallow" segmentation and loss minimization in deep CNN segmentation.

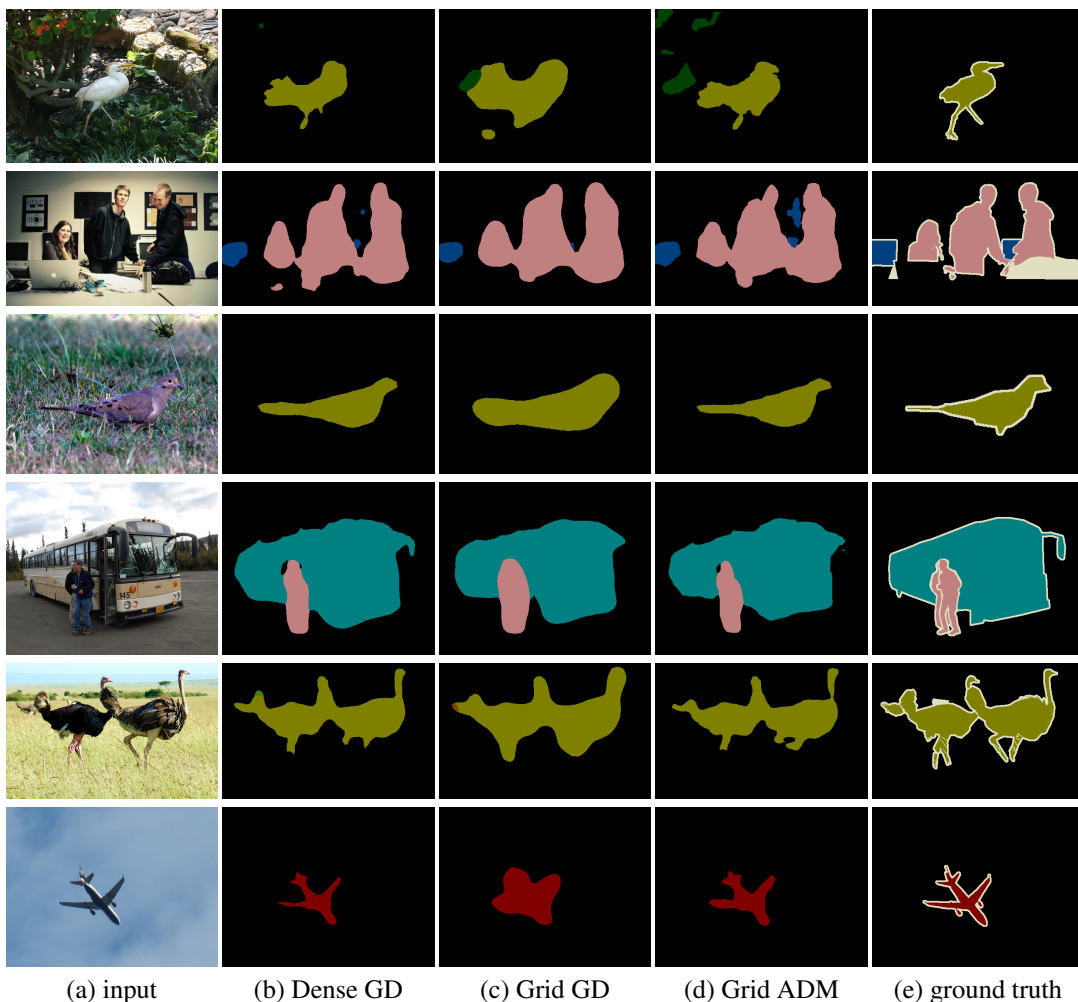|				
(a) input	(b) Dense GD	(c) Grid GD	(d) Grid ADM	(e) ground truth

Figure 5. Example segmentations (Deeplab-MSc-largeFOV) by variants of regularized loss approaches. Gradient descent (GD) for grid CRF gives segmentation of poor boundary alignment though grid CRF is part of the regularized loss. ADM for grid CRF significantly improves edge alignment and compares favorably to dense CRF based method.
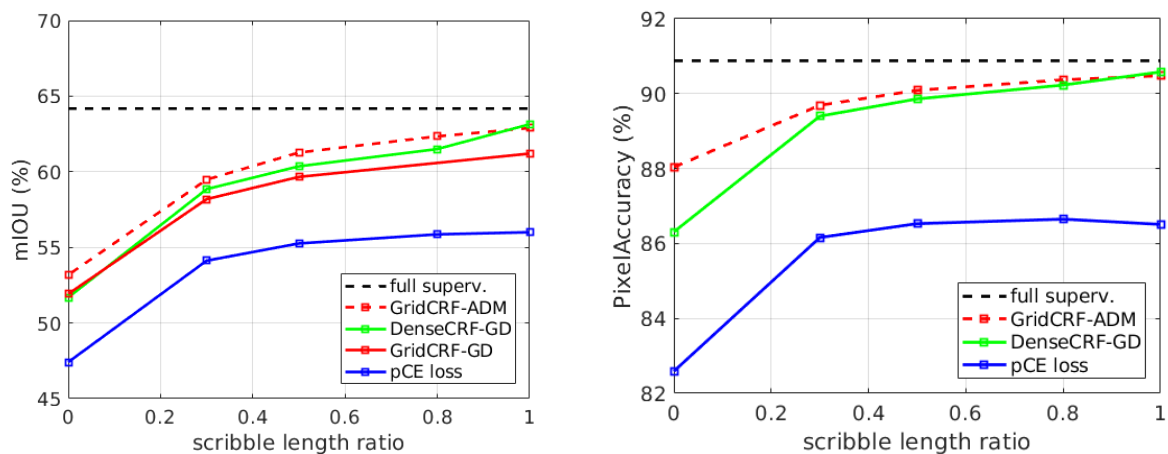

Figure 6. Experiment results of training with shorter scribbles with variants of regularized loss approaches. The results are for Deeplab-MSc-largeFOV. We report mIOU (left) and pixel-wise accuracy (right).

References

[1] A. Blake and A. Zisserman. *Visual Reconstruction*. Cambridge, 1987. 1

[2] Endre Boros, PL Hammer, and X Sun. Network flows and minimization of quadratic pseudo-boolean functions. Technical report, Technical Report RRR 17-1991, RUTCOR, 1991. 4

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 1, 3

[4] Yuri Boykov and Marie-Pierre Jolly. *Interactive graph cuts* for optimal boundary & region segmentation of objects in N-D images. In *ICCV*, volume I, pages 105–112, July 2001. 1, 2, 3, 4, 5

[5] Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *International Conference on Computer Vision*, volume I, pages 26–33, 2003. 1

[6] 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, November 2001. 1, 2, 3, 4, 5

[7] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997. 1, 2

[8] Antonin Chambolle, Daniel Cremers, and Thomas Pock. A convex approach to minimal partitions. *SIAM Journal on Imaging Sciences*, 5(4):1113–1158, 2012. 4

[9] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. 2

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016. 4, 5

[11] C. Couprie, L. Grady, L. Najman, and H. Talbot. Power watershed: A unifying graph-based optimization framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1384–1399, July 2011. 4

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascalnetwork.org/challenges/VOC/voc2012/workshop/index.html. 5

[13] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984. 1

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 1, 7

[15] L Gorelick, O Veksler, Y Boykov, I Ben Ayed, and A Delong. Local submodular approximations for binary pairwise energies. In *Computer Vision and Pattern Recognition*, 2014. 1, 4

[16] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 4

[17] Jörg H Kappes, Bjoern Andres, Fred A Hamprecht, Christoph Schnörr, Sebastian Nowozin, Dhruv Batra, Sungwoong Kim, Bernhard X Kausler, Thorben Kröger, Jan Lellmann, et al. A comparative study of modern inference techniques for structured discrete energy minimization problems. *International Journal of Computer Vision*, 115(2):155–184, 2015. 1, 7

[18] M. Kass, A. Witkin, and D. Terzolpoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988. 1

[19] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. 7

[20] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016. 1, 4

[21] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1568–1583, 2006. 1, 4

[22] Philipp Krahenbuhl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011. 1, 2, 4

[23] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016. 1, 4, 5, 6

[24] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42:577–685, 1989. 1

[25] Claudia Nieuwenhuis, Eno Toeppe, Lena Gorelick, Olga Veksler, and Yuri Boykov. Efficient squared curvature. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 7

[26] Judea Pearl. *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California, Los Angeles, 1982. 4

[27] Thomas Pock, Antonine Chambolle, Daniel Cremers, and Horst Bischof. A convex relaxation approach for computing minimal partitions. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[28] Carsten Rother, Vladimir Kolmogorov, Victor Lempitsky, and Martin Szummer. Optimizing binary mrfs via extended roof duality. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 4

[29] Alexander G Schwing and Raquel Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015. 2

[30] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:888–905, 2000. 5, 7

[31] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, 2008. 1

[32] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized Cut Loss for Weakly-supervised CNN Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 4, 5, 7

[33] Meng Tang, Dmitrii Marin, Ismail Ben Ayed, and Yuri Boykov. Kernel cuts: Kernel and spectral clustering meet regularization. *International Journal of Computer Vision (IJCV)*, 127(5):477–511, May 2019. 7

[34] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On Regularized Losses for Weakly-supervised CNN Segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 4, 5, 6, 7

[35] Gavin Taylor, Ryan Burmeister, Zheng Xu, Bharat Singh, Ankit Patel, and Tom Goldstein. Training neural networks without gradients: A scalable admm approach. In *International conference on machine learning*, pages 2722–2731, 2016. 1

[36] Olga Veksler. Efficient graph cut optimization for full CRFs with quantized edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, doi 10.1109/TPAMI.2019.2906204, 2019 (accepted). 2

[37] Huahua Wang and Arindam Banerjee. Bregman alternating direction method of multipliers. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2816–2824. Curran Associates, Inc., 2014. 3

[38] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012. 1, 7

[39] Jing Yuan, Egil Bae, and Xue-Cheng Tai. A study on continuous max-flow and min-cut approaches. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[40] Alan Yuille. Belief propagation, mean-field, and bethe approximations. *Dept. Stat., Univ. California, Los Angeles, Los Angeles, CA, USA, Tech. Rep*, 2010. 4

[41] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015. 2