ADM for grid CRF loss in CNN segmentation

Dmitrii Marin  
Western University  
Canada

Meng Tang  
University of Waterloo  
Canada

Ismail Ben Ayed  
ETS Montreal  
Canada

Yuri Boykov  
University of Waterloo  
Canada

Abstract

Variants of gradient descent (GD) dominate CNN loss minimization in computer vision. But, as we show, some powerful loss functions are practically useless only due to their poor optimization by GD. In the context of weakly-supervised CNN segmentation, we present a general ADM approach to regularized losses, which are inspired by well-known MRF/CRF models in “shallow” segmentation. While GD fails on the popular nearest-neighbor Potts loss, ADM splitting with $\alpha$-expansion solver significantly improves optimization of such grid CRF losses yielding state-of-the-art training quality. Denser CRF losses become amenable to basic GD, but they produce lower quality object boundaries in agreement with known noisy performance of dense CRF inference in shallow segmentation.

1 Motivation and background

Regularized loss functions are widely used for weakly supervised training of neural networks [23, 10]. In particular, they are useful for weakly supervised CNN segmentation [21, 22] where full supervision is often infeasible, particularly in biomedical applications. Such losses are motivated by regularization energies in shallow segmentation, where multi-decade research efforts went into designing robust regularization models based on geometry [17, 6, 3], physics [13, 1], or robust statistics [9]. Such models should represent realistic shape priors compensating for significant image data ambiguities, yet be amenable to efficient solvers. Many robust regularization models commonly used in vision [20, 12] 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. However, the use of regularization models as losses in deep learning is limited by the ability to optimize them via gradient descent, the backbone of current training methods.

This paper uses weakly supervised CNN segmentation as a representative example to discuss optimization of a general class of regularized losses based on pairwise Potts model, one of the most basic robust models in vision. We consider two common variants, the nearest-neighbor and large-neighborhood Potts, also known as sparse grid CRF and dense CRF models in shallow segmentation.

1.1 Pairwise CRF regularization 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 $S := (S_p \in \mathbb{Z}^+ | p \in \Omega)$ of image pixels $p \in \Omega$ as follows

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

In this paper, “shallow” refers to methods unrelated to deep learning.
where $N$ 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 $N_k$, as well as its popular variational analogues, e.g. geodesic active contours [6], convex relaxations [18, 7], or continuous max-flow [24], are particularly well-researched. It is common to use contrast-weighted discontinuity penalties \([5, 4]\) between the neighboring points, as emphasized by the condition \(\{pq\} \in N_k\) below

\[
w_{pq} = \lambda \cdot \exp \left(-\frac{||I_p - I_q||^2}{\sigma^2}\right) \cdot \left[\{pq\} \in N_k\right]. \tag{2}
\]

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. $\alpha$-expansion [5].

Dense CRF [15] is a Potts model where pairwise interactions are active over significantly bigger neighborhoods defined by a Gaussian kernel with a relatively large bandwidth $\Delta$ over pixel locations

\[
w_{pq} = \lambda \cdot \exp \left(-\frac{||I_p - I_q||^2}{\sigma^2}\right) \cdot \exp \left(-\frac{|p - q|^2}{\Delta^2}\right). \tag{3}
\]

Its use in shallow vision is limited as it often produces noisy boundaries [15], see also Fig.2(c). Also, global optimization methods mentioned above do not scale to dense neighborhoods. Yet, dense CRF model is very popular in the context of CNNs where it can be used as a trainable regularization layer [25]. 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 [15], which is restricted to Gaussian weights exactly as in (3). This is in contrast with global Potts solvers, e.g. $\alpha$-expansion, that can use arbitrary weights, but become inefficient for dense neighborhoods.

Noise in dense CRF inference results suggests weaker regularization properties. Indeed, it is easy to check that for increasingly larger neighborhoods the Potts model gets closer and closer to cardinality...
We present a general alternating direction method (ADM) to optimizing regularized losses of the form
\[ \min_{\theta} \ell(S_\theta, Y) + \lambda \cdot E_p(S_\theta) \]
where \( S_\theta \in [0,1]^{\Omega \times K} \) is a \( K \)-way softmax segmentation generated by the network, with \( K \) the number of labels and \( \theta \) the set of parameters. \( E_p(S_\theta) \) is a regularization term, e.g., sparse Potts or dense CRF, and \( \ell(S_\theta, Y) \) is a partial ground-truth loss, for instance:
\[ \ell(S_\theta, Y) = \sum_{p \in \Omega_L} H(Y_p, S_{p,\theta}), \]
where \( \Omega_L \subset \Omega \) is the set of labeled pixels and \( H(Y_p, S_{p,\theta}) = -\sum_k Y_p^k \log S_{p,\theta}^k \) is the cross entropy between network predicted segmentation \( S_{p,\theta} \in [0,1]^K \) (a row of matrix \( S_\theta \) corresponding to point \( p \)) and ground truth labeling \( Y_p \in \{0,1\}^K \). \( Y \) is the matrix whose rows are given by the \( Y_p \)'s.

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 \[21\] \[22\]. 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 general ADM-based framework for minimizing regularized losses during network training that can directly employ known efficient solvers for the corresponding shallow regularizers.
- Compared to GD, our ADM approach with \( \alpha \)-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.
- The training quality with grid CRF loss achieves the-state-of-the-art in weakly supervised CNN segmentation. We compare dense CRF loss and (sparse) grid CRF losses.

Our results may inspire more research on loss functions and their optimization.

2 ADM for loss optimization

Given an image \( I \) and its partial ground-truth labeling or mask \( Y \), we consider a CNN regularized loss of the following form:
\[ \min_{\theta} \ell(S_\theta, Y) + \lambda \cdot E_p(S_\theta) \]
where \( S_\theta \in [0,1]^{\Omega \times K} \) is a \( K \)-way softmax segmentation generated by the network, with \( K \) the number of labels and \( \theta \) the set of parameters. \( E_p(S_\theta) \) is a regularization term, e.g., sparse Potts or dense CRF, and \( \ell(S_\theta, Y) \) is a partial ground-truth loss, for instance:
\[ \ell(S_\theta, Y) = \sum_{p \in \Omega_L} H(Y_p, S_{p,\theta}), \]
where \( \Omega_L \subset \Omega \) is the set of labeled pixels and \( H(Y_p, S_{p,\theta}) = -\sum_k Y_p^k \log S_{p,\theta}^k \) is the cross entropy between network predicted segmentation \( S_{p,\theta} \in [0,1]^K \) (a row of matrix \( S_\theta \) corresponding to point \( p \)) and ground truth labeling \( Y_p \in \{0,1\}^K \). \( Y \) is the matrix whose rows are given by the \( Y_p \)'s.

We present a general alternating direction method (ADM) to optimizing regularized losses of the general form in (4) using the following decomposition of the problem:
\[ \min_{\theta, X_p \in \{0,1\}^{\Omega_L \times K}} \ell(S_\theta, Y) + \lambda E_p(X) + \gamma \sum_{p \in \Omega_L} D(X_p|S_{p,\theta}) \]
where \( D \) denotes some divergence measure, e.g., the Kullback-Leibler divergence, and \( \gamma \) a Lagrange multiplier for constraints \( X_p = S_{p,\theta} \). In (5), we introduced latent discrete (binary) variables \( X_p \in \{0,1\}^K \), which are unknown of unlabeled pixels \( p \in \Omega_U \) and constrained to be equal to \( Y_p \) for labeled pixels \( p \in \Omega_L \) (\( X \) is the binary matrix whose rows are given by the \( X_p \)'s). Therefore, instead of optimizing the regularization term with gradient descent, our approach splits regularized-loss problem (4) into two sub-problems. We replace the network softmax outputs \( S_{p,\theta} \) in the regularization term by latent discrete variables \( X_p \) and ensures consistency between both variables (i.e., \( S_\theta \) and \( X \) are minimized simultaneously) in each iteration of ADM.
The networks are trained in two phases. First we train minimizing (partial) cross entropy loss w.r.t scribbles. Then we train with extra GridCRF or DenseCRF regularization loss. For inference of the focus is on regularized loss approaches [21, 22] yet we also compare our results to proposal generation based method, e.g. ScribbleSup [16]. We test both GridCRF 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. We compare four variants, namely DenseCRF-GD, DenseCRF-ADM, GridCRF-GD and GridCRF-ADM for weakly supervised CNN segmentation. Before comparing segmentations, in Sec. we investigate if ADM optimization achieves better regularized losses than standard GD. Our plots of training losses (CRF energy) vs training iterations show how fast ADM or GD converges. Our experiment confirms that first order approach like GD leads to poor local minimum for GridCRF. There is clear advantage of ADM over GD for minimization of GridCRF loss. In Sec. rather than comparing optimizers, we focus on the modeling, i.e. GridCRF vs DenseCRF. With ADM as the optimizer, our approach of GridCRF regularized loss gives very comparable segmentations to that of DenseCRF based approach. We also study these variants of regularized loss method in more challenging setting of shorter scribbles [16] or clicks in the extreme case.

Dataset and Implementation details Following recent work [8, 16, 14, 21] on CNN semantic segmentation, we report our results on PASCAL VOC 2012 segmentation dataset. We train with scribbles from [16] on the augmented datasets of 10,582 images and test on the val set. Besides standard mIOU (mean intersection over union), we also measure the regularization losses, i.e. GridCRF or DenseCRF. Our implementation is based on DeepLabv2 [2] and we show results on different networks including deeplab-largeFOV, deeplab-msc-largeFOV, deeplab-vgg16 and resnet-101. The networks are trained in two phases. First we train minimizing (partial) cross entropy loss w.r.t scribbles. Then we train with extra GridCRF or DenseCRF regularization loss. For inference of GridCRF and DenseCRF, we use the public implementation of \( \alpha \)-expansion [2] and mean-field [15] respectively. The CRF inference and loss is implemented and integrated as Caffe [11] layers. We run \( \alpha \)-expansion for five iterations, which in most cases gives convergence. Our DenseCRF loss does not include the Gaussian kernel on locations \( XY \), since ignoring this term does not change the mIOU measure [15]. The bandwidth for dense Gaussian kernel on \( RGBXY \) is validated to give the best mIOU. For GridCRF, the kernel bandwidth selection follows standard Boykov-Jolly [4].

In general, our ADM optimization for regularized loss is slower than GD due to Grid/Dense CRF inference. However, for inference algorithms e.g. graph cuts that cannot be easily parallelized, we utilize simple multi-core parallelization for images in a batch to accelerate training.

\[
\min_{\theta} \ell(S_\theta, Y) + \sum_{p \in \Omega_u} D(X_p|S_{p,\theta})
\]

The second step fixes the network output \( S_\theta \) and finds the next latent binary variables \( X \) by minimizing the following objective over \( X \) via \( \alpha \)-expansion:

\[
\min_{X \in \{0,1\}^{d|U} \times K} \lambda E_p(X) + \gamma \sum_{p \in \Omega_u} D(X_p|S_{p,\theta})
\]

3 Experimental results

We conduct experiments for weakly supervised CNN segmentation with scribbles as supervision. The focus is on regularized loss approaches [21, 22] yet we also compare our results to proposal generation based method, e.g. ScribbleSup [16]. We test both GridCRF 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. We compare four variants, namely DenseCRF-GD, DenseCRF-ADM, GridCRF-GD and GridCRF-ADM for weakly supervised CNN segmentation.

Before comparing segmentations, in Sec. we investigate if ADM optimization achieves better regularized losses than standard GD. Our plots of training losses (CRF energy) vs training iterations show how fast ADM or GD converges. Our experiment confirms that first order approach like GD leads to poor local minimum for GridCRF. There is clear advantage of ADM over GD for minimization of GridCRF loss. In Sec. rather than comparing optimizers, we focus on the modeling, i.e. GridCRF vs DenseCRF. With ADM as the optimizer, our approach of GridCRF regularized loss gives very comparable segmentations to that of DenseCRF based approach. We also study these variants of regularized loss method in more challenging setting of shorter scribbles [16] or clicks in the extreme case.

Dataset and Implementation details Following recent work [8, 16, 14, 21] on CNN semantic segmentation, we report our results on PASCAL VOC 2012 segmentation dataset. We train with scribbles from [16] on the augmented datasets of 10,582 images and test on the val set. Besides standard mIOU (mean intersection over union), we also measure the regularization losses, i.e. GridCRF or DenseCRF. Our implementation is based on DeepLabv2 [2] and we show results on different networks including deeplab-largeFOV, deeplab-msc-largeFOV, deeplab-vgg16 and resnet-101. The networks are trained in two phases. First we train minimizing (partial) cross entropy loss w.r.t scribbles. Then we train with extra GridCRF or DenseCRF regularization loss. For inference of GridCRF and DenseCRF, we use the public implementation of \( \alpha \)-expansion [2] and mean-field [15] respectively. The CRF inference and loss is implemented and integrated as Caffe [11] layers. We run \( \alpha \)-expansion for five iterations, which in most cases gives convergence. Our DenseCRF loss does not include the Gaussian kernel on locations \( XY \), since ignoring this term does not change the mIOU measure [15]. The bandwidth for dense Gaussian kernel on \( RGBXY \) is validated to give the best mIOU. For GridCRF, the kernel bandwidth selection follows standard Boykov-Jolly [4].

In general, our ADM optimization for regularized loss is slower than GD due to Grid/Dense CRF inference. However, for inference algorithms e.g. graph cuts that cannot be easily parallelized, we utilize simple multi-core parallelization for images in a batch to accelerate training.
Figure 3: Training progress of ADM and GD on Deeplab-MSc-largeFOV. (a) ADM for sparse CRF with \( \alpha \)-expansion significantly improves convergence. For example, first 1000 iterations of ADM give grid CRF loss lower than converged GD. (b) In contrast, ADM for dense CRF with mean field by [14] performs much worse than simple stochastic GD [21].

3.1 ADM vs gradient descent for CRF losses

In this section we show that for sparse grid CRF losses the ADM approach employing \( \alpha \)-expansion [5], a powerful discrete optimization method, outperforms common gradient descend methods for regularized losses [21, 22] in terms of finding a lower minimum of regularization loss. Table 1 shows grid CRF losses on both training and validation sets for different network architectures. Figure 3(a) shows the evolution of the sparse CRF loss over the number of iterations of training. ADM requires fewer iterations to achieve the same CRF loss.

| network             | training set | validation set |
|---------------------|--------------|----------------|
|                     | GD       | ADM     | GD       | ADM     |
| Deeplab-LargeFOV    | 2.518    | 2.408   | 2.509    | 2.334   |
| Deeplab-MSc-largeFOV| 2.509    | 2.401   | 2.494    | 2.326   |
| Deeplab-VGG16       | 2.374    | 2.098   | 2.421    | 2.138   |
| Resnet-101          | 2.661    | 2.488   | 2.605    | 2.419   |

Table 1: ADM gives better GridCRF losses than gradient descend (GD).

In contrast to Grid CRF, Figure 3(b) shows that the ADM approach by [14, 21] using Dense CRF is worse than simple gradient descend [21] due to limitations of the mean-field optimizer. There is no practical benefits of ADM over gradient descend for Dense CRF. The reason is that both mean-field and gradient descent are first-order approach for approximating Gibbs distribution or the original energy.

We also visualize the gradients with respect to the soft-max layer’s input of the network in Figure 4. Despite different formulations of regularized losses and their optimization, the gradients w.r.t network output are the driving force for training. In most of the cases the gradient based method produces significant gradient values only in a vicinity of the current model prediction boundary. 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 Figure 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.

Thus, in the context of sparse CRFs the ADM approach coupled with \( \alpha \)-expansion shows drastic improvement in the optimization quality.

---

\(^5\) we have sampled 1000 training examples.
3.2 Grid CRF vs Dense CRF Loss for Weakly Supervised CNN Segmentation

The main results of this paper are summarized in Tab. 2. The mIOU measures on the val set of PASCAL 2012 are reported for various networks. To see more clearly the effects of Grid/Dense CRF losses for network training, we show results both with or without DenseCRF post-processing, which is popularized by Chen et al. [8]. 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. Here we compare variants of regularized losses optimized by gradient descent (GD) or ADM. The regularized loss is comprised of partial cross entropy (pCE) w.r.t. scribbles and other regularizers or clustering criteria, e.g. Grid/Dense CRF or normalized cut (NC) [19, 21]. The focus of comparison is on Grid CRF vs Dense CRF via gradient descent or ADM optimization.

As shown in Tab. 2, all regularized approaches work better than non-regularized loss approach that only minimizes empirical loss w.r.t. scribbles. GridCRF-GD performs relatively the worst among other regularized loss method. This is due to the fact that first-order method like gradient descent leads to poor local minimum for Grid-CRF in the context of energy minimization. Better optimization via ADM for GridCRF gives much better segmentation. Indeed, our Grid-ADM compares favorably to DenseCRF-GD and DenseCRF-ADM. The alternative GridCRF based method gives good quality segmentation approaching that for full supervision. Some qualitative segmentation examples are shown in Fig. 5.

GridCRF has been overlooked in deep CNN segmentation currently dominated by DenseCRF as post-processing or trainable layers in fully supervised setting. We show that for weakly supervised CNN segmentation, GridCRF as regularized loss can give segmentation as good as that with DenseCRF. The key of minimizing GridCRF as loss is better optimization via ADM rather than gradient descent. Such competitive results for GridCRF loss confirms that GridCRF is better than DenseCRF in terms of regularization of segmentation, as discussed in Sec. 1.
Table 2: Weakly supervised segmentation results for different choices of network architecture, post-processing, regularized losses and optimization vis gradient descent or ADM. Here shows mIOU on val set of PASCAL 2012.

| Network         | Post processing | Full | Weak Gradient Descent | ADM |
|-----------------|-----------------|------|-----------------------|-----|
|                 |                 |      | pCE [21] | NC [21] | Dense | Grid | Dense | Grid |
| Deeplab-largeFOV | No              | 63.0 | 55.8   | 59.7   | **62.2** | 60.4 | 61.0  | 61.7 |
| Deeplab-MSc-largeFOV | No          | 64.1 | 56     | 60.5   | **63.1** | 61.2 | 61.3  | 62.9 |
| Deeplab-MSc-largeFOV | Yes         | 68.7 | 62.0   | 65.1   | 65.9  | 62.4 | 65.4  | **66.5** |
| Deeplab-VGG16    | No              | 68.8 | 60.4   | 62.4   | 64.3  | 63.3 | 63.4  | **65.2** |
| Deeplab-VGG16    | Yes             | 71.5 | 64.3   | 65.2   | 66.4  | 64.4 | 65.7  | **67.7** |
| ResNet-101       | No              | 75.6 | 69.5   | 72.8   | **72.9** | 71.7 | 72.5  | 72.8 |
| ResNet-101       | Yes             | 76.8 | 72.8   | 74.5   | **75.0** | 74.1 | 74.5  | **75.0** |

It is not obvious whether GridCRF as loss is beneficial for CNN segmentation and we show that straightforward gradient descent for GridCRF didn’t work well. We are the first to systematically compare GridCRF and DenseCRF loss for weakly supervised CNN segmentation and discuss thoroughly their corresponding optimization via gradient descent or ADM. Our technical contribution on optimization helps to reveal the limitation and advantage of GridCRF vs DenseCRF models. Note that our formulation of ADM for regularized loss in Sec. 2 is general and allow any regularization for which there is good inference/optimization algorithm. This connects the abundant literature on energy minimization for various geometric or high-order terms and loss minimization for weakly supervised CNN segmentation.

Following the evaluation protocol in ScribbleSup [16], we also test our regularized loss approaches training with gradually shortened scribbles. In the extreme case, scribbles degenerates to clicks for semantic objects and we are interested in how much weakly supervised segmentation degrades. As shown in Fig. 6, the performance of ScribbleSup [16] drops drastically with shorter scribbles. Our GridCRF-ADM approach gives competitive performance.

4 Conclusion

The top-performing supervised CNN segmentation [21, 22] is based on regularized loss framework for deep learning [23, 10]. While this framework allows any differentiable regularization, gradient descent is known not to be a good optimization method for many regularization terms in shallow image segmentation, e.g. standard Grid CRF. In this paper, we propose a general ADM-based optimization framework for minimizing regularized losses that can take advantage of existing efficient solvers for the corresponding shallow regularizers. In particular, our ADM approach with $\alpha$ -expansion solver achieves significantly better optimization quality for Grid CRF compared to that with gradient descent. With such ADM optimization, training with grid CRF loss achieves the-state-of-the-art in weakly supervised CNN segmentation. We systematically investigated grid CRF and dense CRF losses from modeling and optimization perspectives. With the proposed ADM optimization strategy, we find the largely overlooked grid CRF to compare favorably to popular dense CRF. Our general ADM optimization framework allows further integration of other segmentation regularizers and their efficient solvers to weakly supervised CNN segmentation.
Figure 5: Example segmentations (Deeplab-MSc-largeFOV) by variants of regularized loss approaches. Gradient descent (GD) for GridCRF gives segmentation of poor boundary alignment though GridCRF is part of the regularized loss. ADM for GridCRF significantly improves edge alignment and compares favorably to DenseCRF based method.

Figure 6: Experiment results of training with shorter scribbles with different regularized loss approaches and proposal generation approach [16]. The plots are for Deeplab-MSc-largeFOV without or with DenseCRF post-processing respectively.
References

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

[2] 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.

[3] 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.

[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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

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

[15] 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.
[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.