On Regularized Losses for Weakly-supervised CNN Segmentation

Meng Tang[1], Federico Perazzi[2], Abdelaziz Djelouah[3]
Ismail Ben Ayed[4], Christopher Schroers[3], and Yuri Boykov[1]

[1] Cheriton School of Computer Science, University of Waterloo, Canada
[2] Adobe Research, United States
[3] Disney Research, Zürich, Switzerland
[4] ETS Montreal, Canada

Abstract. Minimization of regularized losses is a principled approach to weak supervision well-established in deep learning, in general. However, it is largely overlooked in semantic segmentation currently dominated by methods mimicking full supervision via "fake" fully-labeled masks (proposals) generated from available partial input. To obtain such full masks the typical methods explicitly use standard regularization techniques for "shallow" segmentation, e.g. graph cuts or dense CRFs. In contrast, we integrate such standard regularizers directly into the loss functions over partial input. This approach simplifies weakly-supervised training by avoiding extra MRF/CRF inference steps or layers explicitly generating full masks, while improving both the quality and efficiency of training. This paper proposes and experimentally compares different losses integrating MRF/CRF regularization terms. We juxtapose our regularized losses with earlier proposal-generation methods. Our approach achieves state-of-the-art accuracy in semantic segmentation with near full-supervision quality.

Keywords: Regularization · Semi-supervised Learning · CNN Segmentation

1 Introduction

We advocate *regularized losses* for weakly-supervised training of semantic CNN segmentation. The use of unsupervised loss terms acting as regularizers on the output of deep-learning architectures is a principled approach to exploit structure similarity of partially labeled data [34, 15]. Surprisingly, this general idea was largely overlooked in weakly-supervised CNN segmentation where current methods often introduce computationally expensive MRF/CRF pre-processing or layers generating "fake" full masks from partial input.

We propose to use (relaxations of) MRF/CRF terms directly inside the loss avoiding explicit guessing of full training masks. This approach follows well-established ideas for weak supervision in deep learning [34, 15] and continues our recent work [30] that proposed the integration of standard objectives in shallow[5] segmentation directly into loss functions. While our prior work [30] is

[5] In this paper "shallow" refers to segmentation methods unrelated to CNNs.

entirely focused on the *normalized cut loss* motivated by a popular balanced segmentation criterion [29], we now study a different class of *regularized losses* including (relaxations of) standard MRF/CRF potentials. Though common as shallow regularizers [6, 5, 27, 20] or jointly trained in CNN [35, 28, 2], CRF/MRF were never used directly as losses in segmentation.

We propose and evaluate several new losses motivated by MRF/CRF potentials and their combination with balanced partitioning criteria [31]. Such losses can be adapted to many forms of weak (or semi-) supervision based on diverse existing MRF/CRF formulations for segmentation. The scope of this paper is focused on training with partial (user scribble) masks where regularized losses combined with cross entropy over the partial masks achieve the state-of-the-art close to full-supervision quality. We also show advantage of regularized loss for semi-supervised segmentation training with both labeled and unlabeled images.

Instead of the basic Potts model [6], we choose popular fully connected pairwise CRF potentials of Krähenbühl and Koltun [20], often referred to as *dense CRF*. In conjunction with CNNs, dense CRFs have become the de-facto choice for semantic segmentation in the contexts of fully [10, 2, 35, 28] and weakly/semi [19, 23, 26] supervised learning. For instance, DeepLab [10] popularized dense CRF as a post-processing step. In fully supervised setting, integrating the unary scores of a CNN classifier and the pairwise potentials of dense CRF achieve competitive performances [35, 2]. This is facilitated by fast mean-field inference techniques for dense CRF based on high-dimensional filtering [1].

Weakly supervised semantic segmentation is commonly addressed by mimicking full supervision via synthesizing fully-labeled training masks (proposals) from the available partial inputs [26, 23, 22]. These schemes typically iterate between two steps: CNN training and proposal generation via regularization-based shallow interactive segmentation, e.g. graph cut [22] for grid CRF or mean-field inference [26, 23] for dense CRF. In contrast, our approach avoids explicit inference by integrating shallow regularizers directly into the loss functions. Section 3 makes some interesting connections between proposal-generation and our regularized losses from optimization perspective.

For simplicity, this paper uses a very basic quadratic relaxation of discrete MRF/CRF potentials, even though there are many alternatives, e.g. TV-based [7] and convex formulations [8, 25], L_p relaxations [12], LP and other relaxations [14, 32]. Evaluation of different relaxations in the context of regularized weak supervision losses is left for future work. Our main contributions are:

- We propose and evaluate several *regularized losses* for weakly supervised CNN segmentation based on dense CRF [20] and kernel cut [31] regularizers (Sec.2). Our approach avoids explicit inference as in proposal-based methods. This continues the study of losses motivated by standard shallow segmentation energies started in [30] with *normalized cut loss*.
- We show that iterative proposal-generation schemes, which alternate CNN learning and mean-field inference for dense CRF, can be viewed as an approximate alternating direction optimization of regularized losses (Sec.3). Alternating schemes (proposal generation) give higher dense CRF loss.

– Comprehensive experiments (Sec.4) with our regularized weakly supervised losses show (1) state-of-the-art performance for weakly supervised CNN segmentation reaching near full-supervision accuracy and (2) better quality and efficiency than proposal generating methods or normalized cut loss [30].

2 Our Regularized Semi-supervised Losses

This section introduces our regularized losses for weakly-supervised segmentation. In general, the use of regularized losses is a well-established approach in semi-supervised deep learning [34, 15]. We advocate this principle for semantic CNN segmentation, propose specific shallow regularizers for such losses, and discuss their properties.

Assuming image I and its *partial* ground truth labeling or mask Y, let $f_\theta(I)$ be the output of a segmentation network parameterized by θ. In general, CNN training with our joint regularized loss corresponds to optimization problem of the following form

$$\min_\theta \ell(f_\theta(I), Y) \quad + \quad \lambda \cdot R(f_\theta(I)) \tag{1}$$

where $\ell(S, Y)$ is a ground truth loss and $R(S)$ is a regularization term or regularization loss. Both losses have argument $S = f_\theta(I) \in [0,1]^{|\Omega| \times K}$, which is K-way softmax segmentation generated by a network. Using cross entropy over partial labeling as the ground truth loss, we have the following joint *regularized semi-supervised loss*

$$\sum_{p \in \Omega_\mathcal{L}} H(Y_p, S_p) \quad + \quad \lambda \cdot R(S) \tag{2}$$

where $\Omega_\mathcal{L} \subset \Omega$ is the set of labeled pixels and $H(Y_p, S_p) = -\sum_k Y_p^k \log S_p^k$ is the cross entropy between network predicted segmentation $S_p \in [0,1]^K$ (a row of matrix S corresponding to point p) and ground truth labeling $Y_p \in \{0,1\}^K$.

In principle, any function $R(S)$ can be used as a loss given its gradient or sub-gradient. This paper studies (relaxations of) regularizers from shallow segmentation as loss functions. Section 2.1 details our MRF/CRF loss and its implementation. In Section 2.2, we propose *kernel cut loss* combining CRF with normalized cut terms and justify this combination.

2.1 Potts/CRF Loss

Assuming that segmentation variables S_p are restricted to binary class indicators $S_p \in \{0,1\}^K$, the standard Potts/CRF model [6] could be represented via Iverson brackets $[\cdot]$, as on the left hand side below

$$\sum_{p,q \in \Omega} W_{pq} \, [S_p \neq S_q] \quad = \quad \sum_{p,q \in \Omega} W_{pq} \, \|S_p - S_q\|^2, \tag{3}$$

where $W = [W_{pq}]$ is a matrix of pairwise discontinuity costs or an *affinity matrix*. The right hand side above is a particularly straightforward quadratic relaxation

of the Potts model that works for relaxed $S_p \in [0,1]^K$ corresponding to a typical soft-max output of CNNs. In fact, this quadratic function is very common in the general context of regularized weakly supervised losses in deep learning [34].

As discussed in the introduction, this relaxation is not unique [7, 8, 25, 12, 14]. We use slightly different quadratic relaxation of the Potts model

$$R_{CRF}(S) = \sum_k S^{k'} W(\mathbf{1} - S^k) \tag{4}$$

expressed in terms of support vectors for each label k, i.e. columns of the segmentation matrix $S^k \in [0,1]^{|\Omega|}$. For discrete segment indicators (4) gives the cost of a cut between segments, same as the Potts model on the left hand side of (3), but it differs from the relaxation of the right hand side of (3).

The affinity matrix W can be sparse or dense. Sparse W commonly appears in the context of boundary regularization and edge alignment in shallow segmentation [5]. With dense Gaussian kernel W_{pq} (4) is a relaxation of DenseCRF [21]. The implementation details including fast computation of the gradient (11) for CRF loss with dense Gaussian kernel is described in Sec. 4.

2.2 Kernel Cut Loss

Besides the CRF loss (4), we also propose its combination with normalized cut loss [30] where each term is a ratio of a segment's cut cost (Potts model) over the segment's weighted size (normalization)

$$R_{NC}(S) = \sum_k \frac{S^{k'} \hat{W}(1 - S^k)}{d' S^k}, \tag{5}$$

where $d = \hat{W}\mathbf{1}$ are node degrees. Note that the affinity matrix $\hat{W}$ for normalized cut can be different from W in CRF (4). The combined *kernel cut loss* is simply a linear combination of (4) and (5)

$$R_{KC}(S) = \sum_k S^{k'} W(\mathbf{1} - S^k) \;\; + \;\; \gamma \sum_k \frac{S^{k'} \hat{W}(1 - S^k)}{d' S^k} \tag{6}$$

which is motivated by *kernel cut* shallow segmentation [31] with complementary benefits of balanced normalized cut clustering and object boundary regularization or edge alignment as in Potts model. While the kernel cut loss is a high-order objective, its gradient (12) can be efficiently implemented, see Sec. 4.

This paper compares experimentally CRF, normalized cut and kernel cut losses for weakly supervised segmentation. In our experiments, the best weakly supervised segmentation is achieved with kernel cut loss.

Note that standard normalized cut and CRF objectives in shallow segmentation require fairly different optimization techniques (e.g. spectral relaxation or graph cuts), but the standard gradient descent approach for optimizing losses during CNN training allows significant flexibility in including different regularization terms, as long as there is a reasonable relaxation.

3 Connecting Proposals Generation & Loss Optimization

The majority of weakly-supervised methods generate segmentation proposals and train with such ''fake" ground truth [22, 33, 18, 19, 23, 13]. In fact, many off-line shallow interactive segmentation techniques can be used to propagate labels and generate such masks, e.g. graph cuts [5, 27], random walker [16, 12], etc. However, training is vulnerable to mistakes in the proposals. While alternating proposal generation and network training [22] may improve the quality of the proposals, errors reinforce themselves in such self-taught learning scheme [9]. Our regularized semi-supervised loss framework avoids training networks to fit potential errors and is based on well-established principles [9, 34].

In this section, we show that some methods using *dense CRF* to generate proposals [19] can be viewed as an *approximation* of alternating direction method (ADM) [4] for optimizing our dense CRF loss. Our experiments suggest that gradient descent in standard back-propagation for this loss is a better optimization technique compared to ADM splitting involving mean-field inference [20], both in the efficiency and quality of the solutions obtained, see the loss plot in Fig. 3 and the training times in Table 3. However, there is room for exploring our ADM optimization insights for regulared losses other than dense CRF and more powerful inference than mean-field. This is left for future work.

We consider proposal-generation schemes iterating between two steps, **network training** and **proposal generation**. Then alternation can happen either when CNN training converges or online for each batch. At each iteration, the first step learns the network parameters θ from a given (fixed) ground-truth proposal $\tilde{X}$ computed at the previous iteration. This amounts to updating the K-way softmax segmentation S to $\tilde{S} \equiv f_{\tilde{\theta}}(I)$ by minimizing the following proposal-based cross entropy with respect to parameters θ via standard back-propagation:

$$\tilde{\theta} = \arg\min_{\theta} \sum_{p \in \Omega_{\mathcal{L}}} H(Y_p, S_p) + \sum_{p \in \Omega_{\mathcal{U}}} H(\tilde{X}_p, S_p) \quad \text{for} \quad S \equiv f_\theta(I) \tag{7}$$

where $\tilde{X}_p \in [0,1]^K$ are the ground-truth proposals for unlabeled pixels $p \in \Omega_{\mathcal{U}}$. Mask $\tilde{X}_p$ is constrained to be equal to Y_p for labeled pixels $p \in \Omega_{\mathcal{L}}$. The second step fixes the network output $\tilde{S}$ and finds the next ground-truth proposal by minimizing regularization functionals that are standard in shallow segmentation:

$$\min_{X \in [0,1]^{|\Omega| \times K}} \sum_{p \in \Omega_{\mathcal{U}}} H(X_p, \tilde{S}_p) + \lambda R(X) \tag{8}$$

where $X_p \in [0,1]^k$ denotes latent pixel labels within the probability simplex. Note that for fixed $\tilde{S}$ the cross entropy terms $H(X_p, \tilde{S}_p)$ in (8) are unary potentials for X. When R corresponds to dense CRF, optimization of (8) is facilitated by fast mean-field inference techniques [20, 3] significantly reducing the computational times via parallel updates of variables X_p and high-dimensional filtering [1]. Supplementary material shows that mean-field algorithms can be

equivalently interpreted as a *convex-concave* approach to optimizing the following objective

$$\min_{X\in[0,1]^{|\Omega|\times K}} \sum_{p\in\Omega_{\mathcal{U}}} H(X_p,\tilde{S}_p) + \lambda R(X) - \sum_{p\in\Omega_{\mathcal{U}}} H(X_p) \tag{9}$$

combining (8) and negative entropies $H(X_p) = -\sum_k X_p^k \log X_p^k$ that act as a simplex *barrier* for variables X_p. This yields closed-form independent (parallel) updates of variables X_p, while ensuring convergence under some conditions[6].

Proposition 1. *Proposal methods alternating steps* (9) *and* (7) *can be viewed as approximate* alternating direction method *(ADM)*[7] *[4] for optimizing our* regularized loss (2) *using the following decomposition of the problem:*

$$\min_{\theta, X\in[0,1]^{|\Omega|\times K}} \sum_{p\in\Omega_{\mathcal{L}}} H(Y_p,S_p) + \lambda R(X) + \sum_{p\in\Omega_{\mathcal{U}}} KL(X_p|S_p) \tag{10}$$

where KL denotes the Kullback-Leibler divergence.

Proof. The link between (10) and (9) comes directly from the following relation between the KL divergence and the entropies: $KL(X_p|S_p) = H(X_p,S_p) - H(X_p)$.

Instead of optimizing directly regularized loss (2) with respect to network parameters, proposal methods splits the optimization problem into two easier subproblems in (10). This is done by replacing the network softmax outputs S_p in the regularization by latent distributions X_p (the proposals) and minimizing a divergence between S_p and X_p, which is KL in this case. This is conceptually similar to the general principles of ADM [4], except that the splitting is not done directly with respect the variables of the problem (i.e., parameters θ) but rather with respect to network outputs S. This can be viewed as an *approximate* ADM scheme, which does not account directly for variables θ in the ADM splitting.

In this paper we focus on optimization of dense CRF loss via gradient descent or ADM. The method in [19] generates proposals via dense CRF layer, but their approach slightly deviates from the described ADM scheme since they also back-propagate through this layer[8]. But, as we show in Table 3, such back-propagation does not help in practice and can be dropped. Moreover, our gradient descent optimization of dense CRF loss makes such proposal generation layers (or procedures) redundant. Our approach gives simpler and more efficient training without expensive iterative inference [19] and obtains better performance.

[6] Parallel updates are guaranteed to converge for concave CRF models, e.g. Potts [21].

[7] In its basic form, *alternating direction method* transforms problem $\min_x f(x) + g(x)$ into $\min_{x,y} f(x) + g(y)$ s.t $x = y$ and alternates optimization over x and y. This may work if optimizing f and g seperately is easier than the original problem.

[8] Cross-entropy loss $H(X(S),S)$ in [19] uses CRF layer proposal $X(S)$ generated from network output S. Dependence of X on S motivates back-propagation for this layer.

4 Experiments

Sec. 4.1 is the main experimental result of this paper. For weakly-supervised segmentation with scribbles [22], we train using different regularized losses including our proposed CRF loss, high-order normalized cut loss in [30] and kernel cut loss, as discussed in Sec. 2. We show that combining CRF (4) with normalized cut (5) *a la* KernelCut [31] yields the best performance.

In Sec. 4.2, training using regularized loss or using generated proposals are compared in terms of segmentation quality and optimization. Besides for scribbles, we also utilize our regularized loss framework for image-level labels based supervision and compare to SEC [19], a recent method based on proposal generation. We compare schemes of CRF regularization e.g. as loss, post-processing or trainable layers in Sec. 4.3 for weakly-supervised segmentation. In Sec. 4.4, we train with shortened scribbles to see how much each method degrades.We also investigate regularization loss for fully or semi-supervised segmentation (with labeled and unlabeled images), see preliminary results in Sec. 4.5.

Dataset: Most experiments are on the PASCAL VOC12 segmentation dataset. For all method, we train with the augmented dataset of 10,582 images. The scribble annotations are from [22]. Following standard protocol, mean intersection-over-union (mIoU) is evaluated on the *val* set that contains 1,449 images. For image-level label supervision, our experiment setup is the same as in [19].

Implementation details: Our implementation is based on DeepLab v2 [10]. We follow the learning rate strategy in DeepLab v2 [9] for the baseline with full supervision. For our method of regularized loss, we first train with partial cross entropy loss only. Then we fine-tune with extra regularized losses of different types. Our CRF and normalized cut regularization losses are defined at full image resolution. If the network outputs shrinked labeling, which is typical, the labeling is interpolated to original size before feeding into the loss layer.

We choose dense Gaussian kernel over RGBXY for $R_{CRF}(S)$, $R_{NC}(S)$ and $R_{KC}(S)$. As hyper-parameter, the Gaussian bandwidth is optimized via validation for DenseCRF, normalized cut and kernel cut. As is also mentioned in [30], naive forward and backward pass of such fully-connected pairwise or high-order loss layer would be prohibitively slow ($O(|\Omega|^2)$ for $|\Omega|$ pixels). For example, to implement $R_{CRF}(S)$ (4) as a loss, we need to compute its gradient w.r.t. S^k,

$$\frac{\partial R_{CRF}(S)}{\partial S^k} = -2WS^k. \tag{11}$$

For DenseCRF where W is fully connected Gaussian, computing the gradient (11) becomes a standard Bilateral filtering problem, for which many fast methods were proposed [1, 24]. We implement our loss layers using fast Gaussian filtering [1], which is also utilized in the inference of DenseCRF [20, 35, 28]. Using the same fast filtering component, we can also computer the following gradient (12) of our Kernel Cut loss (6) in linear time. Note that our CRF and KC loss layer is much faster than CRF inference layer [19, 35] since no iterations is needed.

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

	Weak				Full
	pCE only	w/ NC [30]	w/ CRF	w/ KernelCut	
DeepLab-MSc-largeFOV	56.0 (8.1)	60.5 (3.6)	63.1 (1.0)	**63.5 (0.6)**	64.1
DeepLab-MSc-largeFOV+CRF	62.0 (6.7)	65.1 (3.6)	65.9 (2.8)	**66.7 (2.0)**	68.7
DeepLab-VGG16	60.4 (8.4)	62.4 (6.4)	64.4 (4.4)	**64.8 (4.0)**	68.8
DeepLab-VGG16+CRF	64.3 (7.2)	65.2 (6.3)	66.4 (5.1)	**66.7 (4.8)**	71.5
DeepLab-ResNet101	69.5 (6.1)	72.8 (2.8)	72.9 (2.7)	**73.0 (2.6)**	75.6
DeepLab-ResNet101+CRF	72.8 (4.0)	74.5 (2.3)	**75.0 (1.8)**	**75.0 (1.8)**	76.8

Table 1: mIOU on PASCAL VOC2012 *v*al set. Our flexible framework allows various types of regularization losses, e.g. normalized cut, CRF or their combinations (KernelCut [31]) as joint loss. We achieved the state-of-the-art with scribbles. In () shows the offset to the result with full masks.

$$\frac{\partial R_{KC}(S)}{\partial S^k} = -2WS^k + \gamma\frac{S^{k'}\hat{W}S^k d}{(d'S^k)^2} - \gamma\frac{2\hat{W}S^k}{d'S^k}. \tag{12}$$

4.1 Comparison of Regularized Losses

Tab. 1 summaries the results with different regularized losses. Here we report both result with or without standard CRF post-processing on various networks. The baselines are with cross entropy losses of fully labeled masks or partial cross entropy (pCE) on seeds. We choose the weight of the regularization term to achieve the best validation accuracy. The state-of-the-art of scribble-based segmentation is from prior work [30] with extra normalized cut loss. Consistently over different networks, using the proposed CRF loss outperforms that with normalized cut loss. Our best result is obtained when combining both normalized cut loss and DenseCRF loss. Clearly, utilization of CRF loss and KernelCut loss reduce the gap toward the full supervision baseline. With DeepLab-MSc-largeFOV followed by CRF post processing, using KernelCut regularized loss achieved mIOU of 66.7%, while previous best is 65.1% with normalized cut loss [30]. Our result with scribbles approaches **97.6%** (75.0%/76.8%) of the quality of that with full supervision, yet only **3%** of all pixels are scribbled. This paper pushes the limit of weakly supervised segmentation.

To get some intuition about these losses and their regularization effect, we visualize their gradient w.r.t. segmentation $\frac{\partial R(S)}{\partial S}$ in Fig. 1. Note that the *sign* of gradients indicates whether to encourage or discourage certain labeling. The color coded gradients clearly show evidence toward better color clustering and edge alignment for normalized cut and CRF. The gradients of different losses are slightly different and complement each other. Since kernel cut is the combination of normalized cut with CRF, then its gradient is the sum of that of each.

Fig. 2 shows some qualitative examples with different losses. Results with regularized loss is better than that without. Besides, the segmentation with kernel cut loss have better edge alignment compared to that with normalized cut loss.

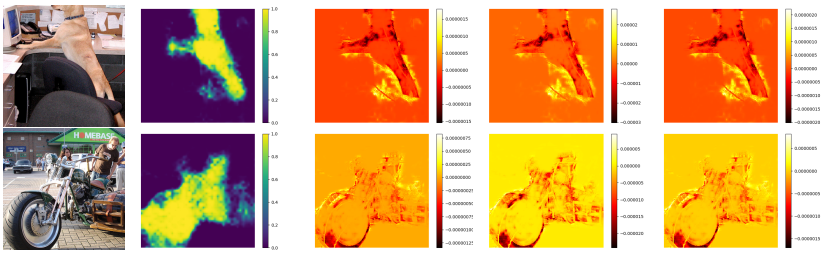


Fig. 1: Visualization of the gradient for different losses. The negative (positive) gradients are coded in red (yellow). For example, negative gradients on the dog drives the network to predict "dog" for these pixels. Also note how the dog pops out in the gradient map.

This is because of the extra pairwise CRF loss. The effect of CRF loss and normalized cut loss is different. Our Kernel Cut loss combines the benefit of regional color clustering (normalized cut) and pairwise regularization (DenseCRF).

Fig. 2: Examples on PASCAL VOC *val* set. Kernel cut as regularization loss gives qualitatively better result than that with normalized cut loss. Kernel cut based results have better edge alignment. See suppl. material for more examples.

DeepLab networks	Weak				Full
	proposal generation			regularized loss	
	GrabCut (one time)	ScribbleSup (iterative)	DenseCRF-ADM (online)	DenseCRF-GD	
MSc-largeFOV	55.5	n/a	61.3	**63.1**	64.1
MSc-largeFOV+CRF	59.7	63.1	65.4	**65.9**	68.7
VGG16	59.0	n/a	63.4	**64.4**	68.8
ResNet101	63.9	n/a	72.5	**72.9**	75.6

Table 2: Results using weak supervision (scribbles). The baseline is training with GrabCut output. ScribbleSup [22] alternates between GrabCut and CNN training, but the proposals are generated offline. It helps to have frequent online proposal updates at each iteration of training as in DenseCRF-ADM. The best (quality and speed) training is based on simple regularized loss with gradient descent (DenseCRF-GD) avoiding proposal generations.

4.2 Regularized Loss vs Proposal Generation

Here we compare our regularized loss and proposal generation methods (Sec. 3) in weakly supervised setting mainly focusing on scribbles. Proposals can be generated *offline* or *online*. One straightforward proposal method is to treat GrabCut output as "fake" ground truth for training. ScribbleSup [22] refines GrabCut output using network predicted segmentation as unary potentials. The proposals are updated but are generated offline. By online proposal generation, we let network output go through a CRF inference layer during training at each iteration. The loss for proposal generation is the cross entropy between the input and output of the CRF inference layer, see Sec.3. A recent work that generates proposals online for tag-based weakly-supervised segmentation is SEC [19].

Table 2 compares regularized loss method to variants of proposal generation. We used the public implementation of SEC's *constrain-to-boundary loss* that have mean-field inference layer and cross entropy loss layer between the proposal and network output. We didn't backpropagate through the inference layer and refer to this version as DenseCRF-ADM in Table 2. It essentially minimizes DenseCRF loss with ADM, rather than gradient descent in our method denoted as DenseCRF-GD.Compared to our regularized loss method, proposal generation gives inferior segmentation for different networks, see Table 2.

We further compare regularized loss (DenseCRF-GD) and proposal generation based approach (DenseCRF-ADM) from optimization perspective. Figure 3 compares the two methods in terms of obtained loss values besides segmentation accuracies . For DenseCRF, ADM style optimization via proposals gives higher loss than that with gradient descent. As discussed in Sec. 3, this may be due to the limitation of first-order mean-field inference for DenseCRF. Exploring ADM for other regularizer with stronger "shallow" optimizers is left as future work.

As mentioned earlier, SEC [19] was originally focused on tag-based supervision. Table 3 reports some tests for that form of weak supervision. We compare SEC with its simplification replacing their constrain-to-boundary loss by our

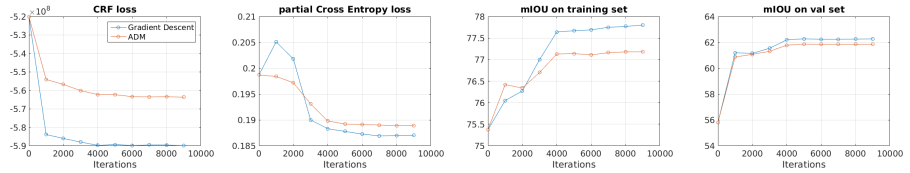


Fig. 3: We compare gradient descent and ADM (via iterative proposal generation) for dense CRF. Gradient descent gives better losses than that of ADM at convergence. Also, using gradient descent for dense CRF achieves higher mIOU on training and val set.

regularized loss. We train using different combinations of losses for supervision based on image-level labels/tags. Our CRF loss helps to improve training to 43.9% compared to 38.4% without it. There is only small improvement in segmentation mIOU when replacing constrain-to-boundary loss by CRF loss.

However, our CRF loss layer is several times faster than constrain-to-boundary layer integrating explicit iterative inferences. The segmentation accuracy and overall training speed are also reported in Tab. 3. (The results are for the DeepLab-largeFOV network.) We also tested a variant of SEC without back-propagation of mean-field layer, which we show is not helping in practice. Fig. 4 shows testing examples for our method and SEC with image tags as supervision.

		include this loss?			
Losses	Seeding loss [19]	✓	✓	✓	✓
	Expansion loss [19]	✓	✓	✓	✓
	Constrain-to-boundary loss [19]		✓	⋆	
	Our CRF loss				✓
mIOU (%)		38.4	43.7	43.8	43.9
Overall training time in s/batch		0.86	1.19 (0.33)	1.19 (0.33)	0.98 (0.12)

Table 3: Tag-based weak supervision. We train with different combinations of the losses in SEC [19] and our CRF loss. Replacing the constrain-to-boundary loss in SEC [19] by CRF loss gives minor improvement in accuracy, but training with regularized loss with gradient descent is faster since no iterative CRF inference is needed. We also compare to a variant ($\star$) of SEC without back-propagation of the CRF inference layer. Parenthesis $(\cdot)$ show the computational times for the constrain-to-boundary loss layer or our direct loss layer.

4.3 CRF as Loss, Post-processing or Trainable Layer

We are the fist to propose CRF loss though it's popular to have CRF as post-processing [10] or jointly trained with the network [35, 28]. For example, CRF-as-RNN [35] is proposed for fully supervised segmentation. Here for weakly-

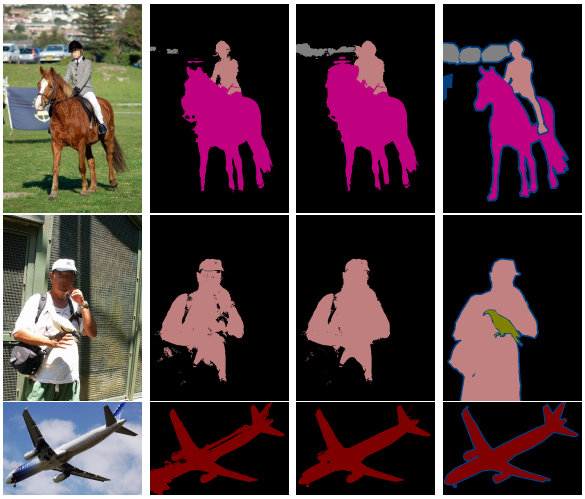


image SEC [19] w/ CRF loss ground truth

Fig. 4: Examples for supervision with image-level labels (tags). We train using the seeding loss, expansion loss in SEC [19] and our CRF loss. Similar segmentation is obtained yet we avoid any iterative mean-field inference for dense CRF.

supervised segmentation with scribbles, we train CRF-as-RNN but only minimizing partial cross entropy loss on scribbles. Table 4 compares the effects of CRF as loss, post-processing or trainable layers. End-to-end training of CRF helps a little bit (64.8% vs 64.3%), but the best is achieved with our CRF loss, which is also much more efficient without any recurrent inference. Note that the plain network trained with extra CRF loss is even better than a network trained without such loss but followed by CRF post-processing, see the fourth and second row in Table 4 (64.4% vs 64.3%). This shows the effectiveness of our CRF loss for training CNN segmentation.

training	testing	mIOU (%)
partial cross entropy loss	plain network	60.4
partial cross entropy loss	disjoint network and CRF	64.3
partial cross entropy loss end-to-end CRF	jointly trained network and CRF	64.8
partial cross entropy loss and our CRF loss	plain network	**64.4**
partial cross entropy loss and our CRF loss	disjoint network and CRF	**66.4**

Table 4: Ablation study of CRF as loss, post-processing [10] or trainable layers [35] for weakly supervised segmentation with scribbles for DeepLab-VGG16.

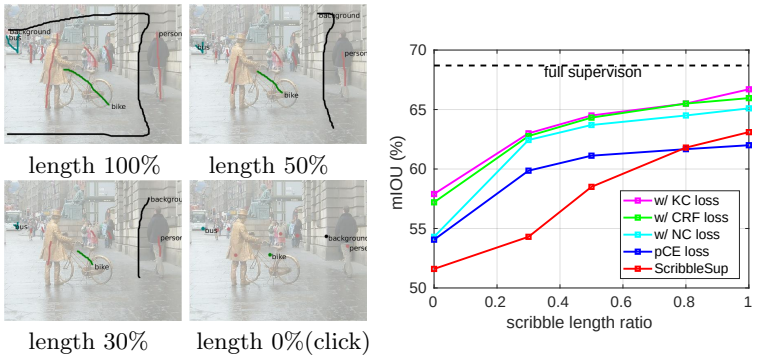


Fig. 5: Similar to [22], we shorten the scribbles. With length zero (clicks) is the most challenging case. Right plot shows mIOUs when train with shorter scribbles.

4.4 Train with Shorter Scribbles

To see the limit of our algorithm with scribble supervision, we train with shortened scribbles visualized in Fig. 5. Note that with length zero, there is only one click for each object. For different length ratios from zero to 100%, our regularized loss method achieved much better segmentation than ScribbleSup [22]. The improvement over ScribbleSup [22] is more significant for shorter scribbles.

4.5 Fully and semi supervised segmentation

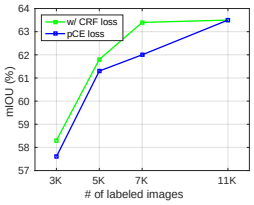
We've demonstrated the usefulness of regularized loss for weakly supervised segmentation. Here we test if it also helps full supervision or semi-supervision.

Full supervision: We add NC loss on fully labeled images besides the cross entropy loss. This experiment is on a simple saliency dataset [11] where color clustering is obvious and likely to help. As shown in Tab. 5, when we increase the weight of $R_{NC}(S)$, we indeed obtained segmentation that is more regularized. However, with extra regularization loss during training, the cross entropy loss got worse and mIOU decreased. The conclusion is that imposing regularized loss naively on labeled images doesn't help. Empirical risk minimization is in some sense optimal for fully labeled data. Extra regularization loss steers the network in the wrong direction if the regularization doesn't totally agree with the ground truth.

NC loss weight	mIOU	cross entropy loss	NC loss
0	89.85%	0.106	0.536
0.1	89.38%	0.110	0.517
0.2	89.39%	0.112	0.509
0.5	88.75%	0.125	0.485

Table 5: Negative effect of regularization loss for full supervision.

Semi supervision: For training with both labeled images and unlabeled images, our joint losses include cross entropy on labeled images and regularization on unlabeled ones. We drop the labeling for some of the the 11K images in PASCAL VOC 2012. We train DeepLab-LargeFOV with different amount of labeled & unlabeled images, see right plot. For the baseline that can only utilize labeled images, the performance degrades with less masks, as expected. For our framework, the labeled and unlabeled images are mixed and randomly sampled in each batch. We observed 0.7% 1.5% improvement with our regularized loss. Note that this result is highly preliminary. We also train with the 11K labeled images in VOC 2012 and the 10K unlabeled in VOC 2017 and achieved boosted performance from 63.5% to 64.3%. In the future, we plan to look into generalization in semi-supervised segmentation and compare to recent work [17].



5 Conclusion and Future Work

Regularized semi-supervised loss is a principled approach to semi-supervised deep learning [34, 15], in general. We utilize such principle for weakly supervised CNN segmentation. In particular, this paper is continuation of the study of losses motivated by standard shallow segmentation [30]. While [30] is entirely on normalized cut loss, here we propose and evaluate several other regularized loss based on Potts/CRF [6, 20], normalized cut [29] and KernelCut [31] regularizer. DenseCRF [20] is very popular as post-processing [10] or trainable layer [2]. We are the first to use a relaxed version of DenseCRF as loss.

In contrast to our regularized loss approach, the main stream in weakly supervised segmentation rely on generating "fake" full masks from partial input and train a network to match the proposals [22, 33, 18, 19, 23, 13]. We show that proposal methods can be viewed as approximate alternating direction method (ADM) for optimization of regularized loss. Gradient descent for DenseCRF loss gives better optimization while being more efficient than proposal generation scheme since no mean-field inference is needed.

This paper pushes the limit of weakly-supervised segmentation. Comprehensive experiments (Sec.4) with our regularized losses show (1) state-of-the-art performance for weakly supervised CNN segmentation reaching near full-supervision accuracy and (2) better quality and efficiency than proposal generating methods or normalized cut loss [30]. Besides for weak supervision, we also report preliminary results for full and semi-supervision.

In principle, our regularized loss framework allows any differentiable loss function. In the future, we plan to explore other relaxations of CRF [7, 8, 25, 12, 14, 32] as losses and corresponding efficient gradient computation. Also it would be interesting to apply our CRF regularized loss framework for weakly-supervised computer vision problems other than segmentation.

References

1. Adams, A., Baek, J., Davis, M.A.: Fast high-dimensional filtering using the permutohedral lattice. Computer Graphics Forum **29**(2), 753–762 (2010)
2. Arnab, A., Zheng, S., Jayasumana, S., Romera-Paredes, B., Larsson, M., Kirillov, A., Savchynskyy, B., Rother, C., Kahl, F., Torr, P.: Conditional random fields meet deep neural networks for semantic segmentation. IEEE Signal Processing Magazine (2017)
3. Baque, P., Bagautdinov, T.M., Fleuret, F., Fua, P.: Principled parallel mean-field inference for discrete random fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3**(1), 1–122 (2011)
5. Boykov, Y., Jolly, M.P.: *Interactive graph cuts* for optimal boundary & region segmentation of objects in N-D images. In: ICCV. vol. I, pp. 105–112 (July 2001)
6. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE transactions on Pattern Analysis and Machine Intelligence **23**(11), 1222–1239 (November 2001)
7. Chambolle, A., Darbon, J.: On total variation minimization and surface evolution using parametric maximum flows. International Journal of Computer Vision **84**(3), 288 (April 2009)
8. Chan, T.F., Esedoglu, S., Nikolova, M.: Algorithms for finding global minimizers of image segmentation and denoising models. SIAM journal on applied mathematics **66**(5), 1632–1648 (2006)
9. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge, MA (2006), http://www.kyb.tuebingen.mpg.de/ssl-book
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv:1606.00915 (2016)
11. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. IEEE TPAMI **37**(3), 569–582 (2015). https://doi.org/10.1109/TPAMI.2014.2345401
12. Couprie, C., Grady, L., Najman, L., Talbot, H.: A unifying graph-based optimization framework. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(7), 1384–1399 (July 2011)
13. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1635–1643 (2015)
14. Desmaison, A., Bunel, R., Kohli, P., Torr, P.H., Kumar, M.P.: Efficient continuous relaxations for dense crf. In: European Conference on Computer Vision. pp. 818–833. Springer (2016)
15. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
16. Grady, L.: Random walks for image segmentation. IEEE transactions on pattern analysis and machine intelligence **28**(11), 1768–1783 (2006)
17. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv:1802.07934 (2018)
18. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, HI, USA (2017)

19. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: European Conference on Computer Vision (ECCV). Springer (2016)
20. Krahenbuhl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: NIPS (2011)
21. Krähenbühl, P., Koltun, V.: Parameter learning and convergent inference for dense random fields. In: International Conference on Machine Learning (ICML) (2013)
22. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3159–3167 (2016)
23. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1742–1750 (2015)
24. Paris, S., Durand, F.: A fast approximation of the bilateral filter using a signal processing approach. International journal of computer vision **81**(1), 24–52 (2009)
25. Pock, T., Chambolle, A., Cremers, D., Bischof, H.: A convex relaxation approach for computing minimal partitions. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR) (2009)
26. Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al.: Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE transactions on medical imaging **36**(2), 674–683 (2017)
27. Rother, C., Kolmogorov, V., Blake, A.: Grabcut - interactive foreground extraction using iterated graph cuts. In: ACM trans. on Graphics (SIGGRAPH) (2004)
28. Schwing, A.G., Urtasun, R.: Fully connected deep structured networks. arXiv preprint arXiv:1503.02351 (2015)
29. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 888–905 (2000)
30. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized Cut Loss for Weakly-supervised CNN Segmentation. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City (June 2018)
31. Tang, M., Marin, D., Ayed, I.B., Boykov, Y.: Normalized Cut meets MRF. In: European Conference on Computer Vision (ECCV). Amsterdam, Netherlands (October 2016)
32. Thalaiyasingam, A., Desmaison, A., Bunel, R., Salzmann, M., Torr, P.H., Kumar, M.P.: Efficient linear programming for dense crfs. In: Conference on Computer Vision and Pattern Recognition (2017)
33. Vernaza, P., Chandraker, M.: Learning random-walk label propagation for weakly-supervised semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 3 (2017)
34. Weston, J., Ratle, F., Mobahi, H., Collobert, R.: Deep learning via semi-supervised embedding. In: Neural Networks: Tricks of the Trade, pp. 639–655. Springer (2012)
35. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1529–1537 (2015)