Bipartite Conditional Random Fields for Panoptic Segmentation

Sadeep Jayasumana[1]
sadeep@apache.org
Kanchana Ranasinghe[2]
kahnchana@gmail.com
Sahan Liyanarachchi[2]
150360a@uom.lk
Mayuka Jayawardhana[2]
150273j@uom.lk
Harsha Ranasinghe[2]
150504v@uom.lk
Sina Samangooei[3]
sina@five.ai

[1] Work done at Five AI
[2] University of Moratuwa, Moratuwa, Sri Lanka
[3] Five AI, Cambridge, UK

Abstract

We tackle the panoptic segmentation problem with a conditional random field (CRF) model. Panoptic segmentation involves assigning a semantic and an instance label to each pixel of a given image. At each pixel, the semantic label and the instance label should be compatible. Furthermore, a good panoptic segmentation should have a number of other desirable properties such as the spatial and color consistency of the labeling. To tackle this problem, we propose a CRF model, named Bipartite CRF or BCRF, with two types of random variables for semantic and instance labels. In this formulation, various energies are defined within and across the two types of random variables to encourage a consistent panoptic segmentation. We propose a mean-field-based efficient inference algorithm for solving the CRF and empirically show its convergence properties. This algorithm is fully differentiable, and therefore, BCRF inference can be included as a trainable module in any deep network. In the experimental evaluation, we quantitatively and qualitatively show that the BCRF yields superior panoptic segmentation results in practice. Our code is publicly available at: https://github.com/sahan-liyanaarachchi/bcrf-detectron.

1 Introduction

Panoptic segmentation of images is a problem that has received considerable attention in computer vision recently. It combines two well-known computer vision tasks: semantic segmentation and instance segmentation. The goal of panoptic segmentation is to assign a semantic label and an instance label for each pixel in the image as presented formally in [10].

Although semantic segmentation and instance segmentation are apparently very related problems, current state of the art methods in computer vision solve these in substantially different ways. The semantic segmentation problem is usually solved with a fully convolutional network architecture such as FCN [20] or DeepLab [4], whereas the instance segmentation problem is solved using an object detector based method such as Mask-RCNN [7].

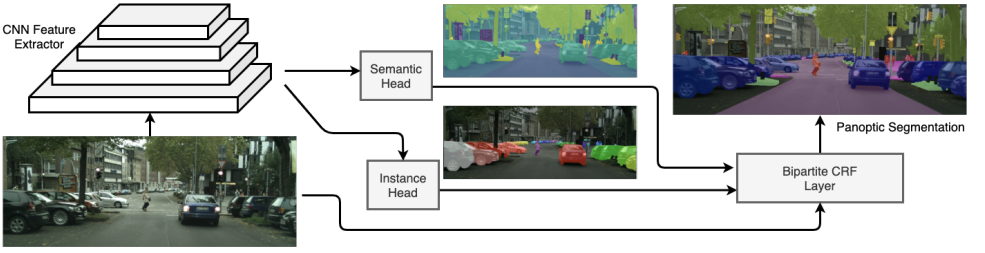

Figure 1: **BCRF in an end-to-end trainable deep net.** The Bipartite CRF proposed in this paper can be used to combine the predictions of a semantic segmentation model and an instance segmentation model to obtain a consistent panoptic segmentation.

Each of these architectures have their own strengths and weaknesses. For example, fully-convolutional network based semantic segmentation methods have a wide field of view, specially when used with dilated convolutions [28], and therefore can make semantic segmentation predictions with global information about the image. In contrast, region proposal based networks, such as Mask-RCNN, focus on specific regions of interest during the later stages of the network and make predictions using strong local features available within a given region of interest [8]. It is natural to think of a systematic way of combining the complementary strengths of these two different approaches.

We propose a Conditional Random Field (CRF) based framework for panoptic segmentation: Bipartite Conditional Random Fields (BCRF). This module performs probabilistic inference on a graphical model to obtain the best panoptic label assignment given the semantic segmentation classifier, the instance segmentation classifier, and the image itself. Our framework provides a heuristic-free, probabilistic method to combine semantic segmentation results and instance segmentation results - yielding a panoptic segmentation with consistent labeling across the entire image. We formulate our bipartite CRF using different energy functions to encourage the spatial, appearance and semantic consistency of the final panoptic segmentation. The optimal labeling is then obtained by performing mean field inference on the bipartite CRF - solving for both the semantic segmentation and the instance segmentation in a jointly optimal way.

Importantly, we show that our proposed BCRF inference is fully differentiable with respect to the parameters used within the CRF and also the semantic segmentation and instance segmentation classifier inputs. Therefore, the BCRF module can be used as a first-class citizen of a deep neural network to perform panoptic segmentation. A deep network equipped with the BCRF module is capable of structured prediction of consistent panoptic labels and is end-to-end trainable. We show an example application of this framework and demonstrate that superior results can be gained by probabilistic combination of a semantic segmentation classifier and an instance segmentation classifier in the BCRF framework.

2 Related Work

2.1 Panoptic Segmentation

Since its formal introduction by Kirillov *et al.* [10], the task of panoptic segmentation has gained popularity, with multiple works attempting to transform existing network architectures to tackle this task [6, 16, 17, 19, 26, 27]. The work in [26] presents a parameter-free panoptic head that logically combines instance and semantic logits. Our work achieves a dif-

ferent goal of individually optimising the two sets of logits learning arbitrary complex mappings between them. Also, the panoptic combination head in [26] could be used on top of our module instead of the generic combination in [10] that we use for further improvement. The spatial ranking methods described in [14, 17] optimize ranking between overlapping instance masks. The work in [16] uses two attention modules to optimize the background segmentation. Our framework performs both these tasks together using our cross-potential terms while enforcing the two branches to have a consensus in their outputs. The BCRF module is thus more robust in terms of information integration.

Another similar recent work by Arnab *et al.* [1] moves in a slightly new direction by using a CRF to obtain instance segmentation outputs from a semantic segmentation using bounding box (from an object detection network) and instance shape cues. Our work differs from this in three significant ways: presence of pixel-wise cross potentials, using instance mask cues from a region-based network, and the ability to explicitly learn and model relationships between classes.

2.2 Conditional Random Fields

Conditional Random Fields (CRFs) are a class of statistical modeling models excellent at structured prediction tasks such as semantic segmentation. While early methods of CRFs for semantic segmentation [8, 22] used 4-connected or 8-connected locally connected graphs, the development of an efficient mean field based inference algorithm [12] to solve fully connected CRFs with Gaussian edge potentials resulted in a resurgence of its use in deep networks. The authors of [29] showed that this CRF inference algorithm can be formulated as a Recurrent Neural Network (RNN), which plugged into a fully convolutional network could obtain the state-of-the-art in semantic image segmentation. Similar trainable CRF models have been used in works such as [2, 21] for semantic segmentation and [1] for instance segmentation. In [15], where the problem of panoptic segmentation with weak and semi supervision was addressed, the authors used a CRF for refining instance segmentation labels. However, it worked on homogeneous instance labels only and therefore was similar in spirit to previous fully connected CRFs.

In our work, we propose a bipartite CRF operating on the semantic segmentation task and the instance segmentation task *simultaneously*. This CRF has energies within semantic segmentation labels, energies within instance segmentation labels, and also energies *across* semantic and instance segmentation labels. To the best of our knowledge, this is the first time a bipartite CRF with cross connections between semantic and instance labels has been proposed in the context of pixel-wise labeling.

3 Background: Conditional Random Fields

A CRF, used in the context of pixel-wise label prediction, models pixel labels as random variables that form a Markov Random Field (MRF) when conditioned upon the image. CRFs have primarily been used in computer vision for semantic image segmentation. In this setting, CRFs encourage the desirable properties of a good segmentation, such as the spatial consistency (e.g. spatially neighboring pixels should have the same label) and color consistency (e.g. a semantic segmentation boundary should correspond to an edge in the image) through various energy functions used in the formulation. A CRF formulation usually has energy terms arising from an imperfect classifier (sometimes known as the unary energy) and energy terms encouraging the consistency properties of the segmentation (sometimes known

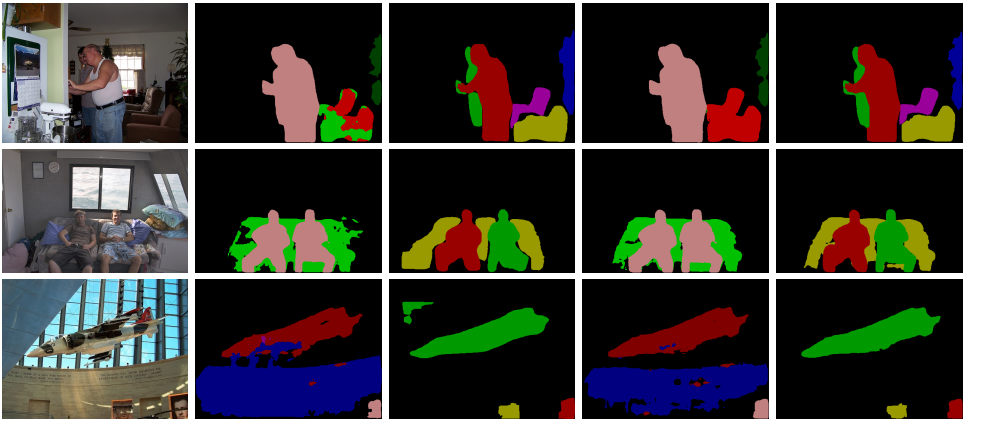

Table 1: **Visualizations on Pascal VOC Dataset.** Columns left to right: original image, semantic output and instance output before BCRF, semantic output and instance output after BCRF.

as the pairwise energy). Some semantic CRF models also include higher order energy terms to encourage higher order consistency properties such as consistency of the labeling within super-pixels [12].

Once an appropriate energy function is formed, the optimal labeling is found as the labeling that minimizes the CRF energy (or equivalently, maximizes the probability). This is known as the inference of the CRF. The exact inference of a CRF with dense pairwise connections is intractable and hence approximate inference methods such as mean field variational inference has to be utilized to solve the CRF in reasonable time [17]. For a detailed treatment of CRFs, the reader is referred to [11].

4 Bipartite CRFs

We propose a CRF formulation with bipartite random variables to capture interactions between semantic labels and instance labels. Inference of this CRF gives the jointly most probable semantic and instance segmentation (and therefore, the panoptic segmentation) for a given image.

For each pixel i, define a pair of discrete random variables (X_i, Z_i) to denote its semantic label and the instance label, respectively. For each i, X_i can take values in $\mathcal{L} = \{l_1, l_2, \ldots, l_L\}$, where each l_j is a semantic label and L is the number of semantic labels (includes both stuff and thing classes). Therefore, $\mathcal{L} = \mathcal{L}_{\text{stuff}} \cup \mathcal{L}_{\text{things}}$, where $\mathcal{L}_{\text{stuff}}$ is the set of stuff class labels and $\mathcal{L}_{\text{things}}$ the set of thing class labels. Similarly, for each i, Z_i can take values in $\mathcal{T} = \{\texttt{inst}_0, \texttt{inst}_1, \ldots, \texttt{inst}_{N_{inst}}\}$, where N_{inst} is the number of instances detected in the image, and the label $\texttt{inst}_0$ is reserved to represent the "no instance" case (the pixel belongs to a stuff class).

Let $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ and $\mathbf{Z} = [Z_1, Z_2, \ldots, Z_N]$, where N is the number of the pixels in the image. A joint assignment $(\mathbf{x}, \mathbf{z})$ to these two random vectors $(\mathbf{X}, \mathbf{Z})$ gives a unique semantic label and an instance label to each pixel i, and therefore represents a panoptic segmentation of the image. Note that, $\mathbf{x} \in \mathcal{L}^N$ and $\mathbf{z} \in \mathcal{T}^N$. In this work, we discuss the probability of such assignments and formulate the probability distribution function so that the "good" panoptic segmentation will have a high probability. We then perform inference on this formulation to find the assignment that maximizes the probability to obtain the best

panoptic segmentation.

The probability of a panoptic segmentation $(\mathbf{x},\mathbf{z})$, given the image I, can be modeled as a Gibbs distribution of the following form:

$$\Pr(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}|I)=\frac{1}{\mathcal{Z}(I)}\exp(-E(\mathbf{x},\mathbf{z}|I)), \tag{1}$$

where $\mathcal{Z}(I)=\sum_{(\mathbf{x},\mathbf{z})}\exp(-E(\mathbf{x},\mathbf{z}|I))$, is a normalization constant, sometimes known as the partition function. The term $E(\mathbf{x},\mathbf{z}|I)$ is known as the energy of the configuration $(\mathbf{x},\mathbf{z})$. Hereafter, we drop the conditioning on I in the notation for brevity. The energy of our bipartite CRF is defined as follows:

$$E(\mathbf{x},\mathbf{z})=\sum_i \phi(x_i)+\sum_{i<j}\Phi(x_i,x_j)+\sum_i \psi(z_i)+\sum_{i<j}\Psi(z_i,z_j)+\sum_i \omega(x_i,z_i)+\sum_{i<j}\Omega(x_i,z_j) \tag{2}$$

where x_i and z_i are the elements of the vectors $\mathbf{x}$ and $\mathbf{z}$, respectively. The meaning of each term will be described in detail below. Note that, since a "good" panoptic segmentation should have a high probability, it should have a low energy. Various terms in Eq. (2) should therefore encourage a good panoptic segmentation by penalizing disagreements with our prior knowledge about a consistent panoptic segmentation.

4.1 Semantic & Instance Components of the CRF

In the following, we discuss the first two terms of the energy function in Eq. (2). The first term encourages the semantic segmentation result to be consistent with the initial classifier,

$$\phi(X_i=x_i)=-\log(\Pr_0(X_i=x_i)), \tag{3}$$

where $\Pr_0(.)$ is the classifier probability score for the semantic segmentation. The second term in Eq. (2) encourages the smoothness of the semantic labeling,

$$\Phi(X_i=x_i,X_j=x_j)=\mu(x_i,x_j)\,\mathrm{Sim}_\Phi(i,j), \tag{4}$$

where $\mu:\mathcal{L}\times\mathcal{L}\rightarrow\mathbb{R}$ is the label compatibility function, and $\mathrm{Sim}_\Phi(i,j)$ is a similarity measure between the pixels i and j. This term penalizes assigning different labels to a pair of pixels that are "similar". Following [12], we use a mixture of Gaussians as the similarity measure and define a general similarity function,

$$\mathrm{Sim}_\chi(i,j)=\sum_m w_{\chi,m}\exp\left(-\frac{\|\mathbf{f}_i^{(m)}-\mathbf{f}_j^{(m)}\|^2}{2\sigma_{\chi,m}^2}\right) \tag{5}$$

where $\mathbf{f}_i$ is a feature vector for pixel i containing information such as its spatial location and bilateral features (RGB + spatial coordinates). We use the same spatial and bilateral features used in [12]. The similarity measure Sim_Φ is derived accordingly.

The next two terms of the energy function in Eq. (2) perform the same for instance classification. Similarly we assume the existence of an initial classifier, such as Mask R-CNN. Despite methods like Mask R-CNN providing fixed-size predictions with respect to the bounding boxes of the detections, these predictions can be easily mapped to the full image by using bilinear interpolation and trivial coordinate transforms similar to [26] as follows.

If there are N detections in the MaskRCNN output, instance segmentation unary potentials is a tensor of shape $[im_{height}, im_{width}, N+1]$. There are (N + 1) channels to accommodate no instance at channel 0, i.e. at each pixel, the unary potential is a vector of length (N + 1) that contains negative logits (see Eq. (3)) corresponding to the detection confidence of each

particular instance. Each pixel in the image may belong to none, one or multiple instances, since overlaps between bounding boxes are possible. These cases are handled as follows: 1) None: "no-instance" channel (channel 0) will have a high negative score. 2) One: The corresponding channel will have a negative score proportional to the detection confidence of the MaskRCNN. 3) Multiple (say u and v): channels u and v will have negative scores proportional to the confidence scores for the two detections. In all cases, other channels will be set to zero.

Similar to the semantic segmentation case, the third term in Eq. (2) encourages the panoptic segmentation to be consistent with the instance classifier probabilities. The fourth term in Eq. (2) encourages instance label consistency across the whole image by penalizing assigning different instance labels to similar pixels:

$$\Psi(Z_i = z_i, Z_j = z_j) = [z_i \neq z_j]\,\mathrm{Sim}_\Psi(i,j). \tag{6}$$

The compatibility transform in this case is fixed to be $[z_i \neq z_j]$, where $[.]$ is the Iverson bracket. The similarity measure Sim_Ψ is derived from Eq. (5).

4.2 Cross Potentials in the CRF

An important contribution of this paper is the introduction of cross potentials between the semantic segmentation and instance segmentation. The semantic segmentation and the instance segmentation are highly related problems and therefore the solutions should agree: the semantic label at any pixel has to be compatible with the instance label at that pixel. For example, if the instance labeling says that the pixel i belongs to an instance of a person class, the semantic label at pixel i should also have the person label. If the initial classifier results for the instance segmentation and the semantic segmentation do not agree, one of them should correct itself depending on the interactions of other terms in the CRF.

The first cross potential term (the fifth term in Eq. (2)), encourages instance label and the semantic label at a given pixel to agree:

$$\omega(X_i = x_i, Z_i = z_i) = f(x_i, \mathrm{class}(z_i)). \tag{7}$$

Here, $\mathrm{class}(z_i)$ is the class label of the instance z_i with $\texttt{inst}_0$ mapped to a special class null. Note that, for all valid instances, the class label can be obtained from the instance classifier (e.g. Mask R-CNN). The function $f(.,.) : (\mathcal{L}, \mathcal{L}_{\text{things}} \cup \{\text{null}\}) \rightarrow \mathbb{R}_0^+$, captures the cost of incompatibility and is defined as follows:

$$f(x_i, \mathrm{class}(z_i)) = \begin{cases} 0, & \text{if } x_i = \mathrm{class}(z_i) \\ 0, & \text{if } x_i \in \mathcal{L}_{\text{stuff}} \text{ and } \mathrm{class}(z_i) = \text{null} \\ \eta(x_i, \mathrm{class}(z_i)), & \text{otherwise.} \end{cases} \tag{8}$$

The above function covers three cases: 1) If the semantic label and the class label of the instance label match, there will be no penalty for such assignment since there is no incompatibility in this case. 2) If the semantic segmentation assigns a stuff label and the instance segmentation assigns $\texttt{inst}_0$ label, there will be no penalty in that case either. 3) If the semantic label and the instance label mismatch, there will be a penalty with the magnitude decided by the function $\eta(.,.) : \mathcal{L}_{\text{things}} \cup \{\text{null}\} \times \mathcal{L}_{\text{things}} \cup \{\text{null}\} \rightarrow \mathbb{R}^+$. This function is learned from data as described in Section 5.

The last term in Eq. (2), encourages the consistency of semantic label and the instance label among similar looking pixels and has the form:

$$\Omega(X_i = x_i, Z_j = z_j) = f(x_i, \mathrm{class}(z_j))\ \mathrm{Sim}_\Omega(i,j), \tag{9}$$

where each symbol has the meaning described above and Sim_Ω is derived from Eq. (5).

5 Inference and Parameter Optimization

The best panoptic segmentation given the model described in Section 4 is the assignment $(\mathbf{x},\mathbf{z})$ that maximizes the probability in Eq. (1). However, since the graphical model used in BCRF has dense connections between the pixels, the exact inference is infeasible. We therefore use an approximate parallel mean field inference algorithm following [12].

In this setting, the joint probability distribution is approximated by the product of marginal distributions:

$$\Pr(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) \approx \prod_i Q_i(x_i)\, R_i(z_i), \tag{10}$$

where $Q_i(x_i) = \Pr(X_i = x_i)$ and $R_i(z_i) = \Pr(Z_i = z_i)$ are the marginal distributions. Out of all the distributions that can be written down in this factorized form, the closest distribution to the original joint distribution is found by minimizing the KL divergence [11, 12]. For our BCRF formulation, this results in the iterative algorithm detailed in Algorithm 1.

Algorithm 1 Inference on Bipartite CRF

$Q_i(l) := \text{softmax}_i(-\phi_i(l))$ and $R_i(t) := \text{softmax}_i(-\psi_i(t))$ ▷ Initialization
while not converged **do**
  $Q'_i(l) \mathrel{-}= \phi_i(l)$ ▷ Update due to the first term
  $Q'_i(l) \mathrel{-}= \sum_{l' \in \mathcal{L}} \left(\mu(l,l') \sum_{j \neq i} \text{Sim}_\Phi(i,j)\, Q_j(l')\right)$ ▷ Update due to the second term
  $R'_i(t) \mathrel{-}= \psi_i(t)$ ▷ Update due to the third term
  $R'_i(t) \mathrel{-}= \sum_{t' \in \mathcal{T}} \left([t \neq t'] \sum_{j \neq i} \text{Sim}_\Psi(i,j)\, R_j(t')\right)$ ▷ Update due to the fourth term
  $Q'_i(l) \mathrel{-}= \sum_{t \in \mathcal{T}} \left(f(l,\text{class}(t))\, R_i(t)\right)$
  $R'_i(t) \mathrel{-}= \sum_{l \in \mathcal{L}} \left(f(l,\text{class}(t))\, Q_i(l)\right)$ ▷ Updates due to the fifth term
  $Q'_i(l) \mathrel{-}= \sum_{t \in \mathcal{T}} \left(f(l,\text{class}(t)) \sum_{j \neq i} \text{Sim}_\Omega(i,j)\, R_j(t')\right)$
  $R'_i(t) \mathrel{-}= \sum_{l \in \mathcal{L}} \left(f(l,\text{class}(t)) \sum_{j \neq i} \text{Sim}_\Psi(i,j)\, Q_j(l')\right)$ ▷ Updates due to the sixth term
  $Q_i(l) := \text{softmax}_i\left(Q'_i(l)\right)$ and $R_i(t) := \text{softmax}_i\left(R'_i(t)\right)$ ▷ Normalization
end while

To make our model flexible, we deliberately include a number of parameters in the BCRF model, which we automatically learn from the training data. More specifically, the BCRF model has the following parameters:

1. Weight multipliers for different energy terms: each term in Eq. (2) is multiplied with a weight parameter, which decides the relative strength of the term. This parameterization helps learn the optimal combination of different energies in the CRF. For example, if the initial semantic segmentation model has better accuracy than the instance segmentation model, the ϕ unary energy might be weighted more than the ψ unary energy.
2. Parameters for similarity functions: Each similarity function $\text{Sim}_X(i,j)$ of the form shown in Eq. (5) has its own parameters. These learn the relative strength of spatial and appearance consistency of the panoptic segmentation.
3. Label compatibility matrices: The two functions $\mu(.,.)$ and $\eta(.,.)$ are initialized to have a zero cost for a pair of identical labels and a fixed cost for any combination of two different labels. They are then given the freedom to automatically learn the relative penalty strengths for different label combinations.

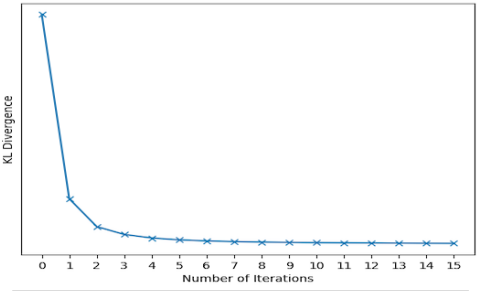

Figure 2: **Convergence of BCRF Inference** The KL divergence is plotted against the number of iterations. We pick 20 random images from the Pascal VOC validation set and average the KL divergence for each iteration across these images. It can be seen that the KL divergence measure, and therefore the inference algorithm, converges within a few iterations.

Table 2: **Results on Pascal VOC Dataset** Our baseline uses DeepLab-v3 and Mask-RCNN followed by combination using the generic method outlined in [10]. CRF only corresponds to setting the BCRF cross-potential terms to zero while BCRF is our complete network.

Method	PQ	SQ	RQ
DeeperLab [27]	67.35	-	-
Ours (baseline)	70.50	88.65	78.83
Ours (CRF only)	67.72	87.62	76.48
Ours (BCRF)	**71.76**	**89.63**	**79.33**

6 Experiments

In this section, we first show the convergence of inference for BCRF followed by how end-to-end training is performed for a deep network with BCRF. The usefulness of BCRF module is then established through experiments on public datasets. The PQ, RQ, and SQ metrics as defined in [10] are used for all experiments.

6.1 Convergence of Inference

It is difficult to provide a theoretical convergence guarantee for mean field algorithms with parallel updates [11, 23]. We therefore provide empirical evidence by estimating the KL divergence between the original joint distribution and the factorized distribution (see Eq. (10)), at the end of each iteration in Algorithm 1. Note that this KL divergence can be estimated up to a constant using the method described in [13]. Our experimental results are shown in Fig. 2. We also note that visual results do not change after about 5 iterations.

6.2 BCRF in a Deep Network

In [29], authors show that, in the semantic segmentation setting, mean field inference of a CRF with Gaussian pairwise potentials can be formulated as a Recurrent Neural Network (RNN). Since our BCRF uses an iterative mean field algorithm of similar nature, it is readily adaptable into this RNN based inference described in [29]. This formulation allows automatic optimization of the BCRF parameters described in Section 5, using backpropagation and a gradient descent algorithm. Accordingly, we build a PyTorch implementation of BCRF which is used in our experiments. Further, given a suitable loss function for panoptic segmentation, the differentials with respect to this loss can be passed on to both the semantic branch and the instance branch to optimize their parameters, and subsequently, the feature extractor CNN's parameters, thus jointly training the entire network end-to-end.

6.3 Results on Pascal VOC Dataset

In this experiment we use the architecture shown in Figure 1 with generic instance and semantic segmentation networks, initialize the BCRF parameters with ones obtained through

	PQ		SQ		RQ	
Category	Baseline	BCRF	Baseline	BCRF	Baseline	BCRF
All	41.4	**41.7**	78.3	**79.1**	50.8	**51.1**
Things	47.4	**47.9**	80.4	**82.1**	57.3	**57.7**
Stuff	32.5	**33.2**	75.1	**77.1**	40.9	**41.6**

Table 3: **Results on COCO dataset.** Comparison of mAP values against a baseline using [9].

Method	**Backbone**	**Params**	**PQ**	**PQst**	**PQst**
OCFusion [14]	ResNetXt-50	-	41.9	49.9	29.9
Panoptic FPN [9]	ResNet - 50	-	39.0	45.9	28.7
Panoptic-DeepLab [6]	Xception-71	46.7M	41.2	44.9	35.7
Axial-DeepLab [24]	Axial-ResNet-L	44.9M	43.9	48.6	36.8
Panoptic FPN with BCRF	ResNet - 50	46.0M	41.7	47.9	33.2

Table 4: **Comparison with the state-of-the-art for COCO dataset.** We compare against other similar sized networks. Panoptic FPN with BCRF (last row) is our work.

a coarse grid search, and initialize the compatibility matrices as described in Section 5. During both training and inference we used 5 mean-field iterations for BCRF. At the output, we calculate the loss function as a summation of two components: the usual pixel-wise categorical cross entropy loss for the semantic component [18] and the cross entropy loss with "matched" ground truth [1] for the instance component. We used full-image training with batch size 1, SGD with learning rate 0.0007, momentum 0.99, and run just 10 epochs to obtain the following results. In Table 2, we report the summary of the quantitative results for the PASCAL VOC validation dataset. Setting cross-potential terms to zero results in a degradation, which highlights the contribution of the BCRF module in fusing two information sources. Qualitative results are shown in Table 1.

6.4 Results on the COCO Dataset

We experiment on the COCO validation set by adding BCRF on top of Panoptic FPN [9] and training using default parameters and panoptic loss functions in its Detectron2 [25] implementation. The quantitative results are reported in Table 4.

6.5 Results on the Cityscapes Dataset

To evaluate the usefulness of BCRF without efforts for thorough end-to-end training, we simply plug in BCRF on an existing pre-trained model, followed by fine-tuning on a small subset of train images. We use a COCO pre-trained Panoptic FPN [9], and run two experiments (with and without BCRF) training on 200 randomly selected images from the Cityscapes train split. Quantitative results for the entire validation set obtained from this experiment are reported in Table 5.

	PQ		SQ		RQ	
Category	Baseline	BCRF	Baseline	BCRF	Baseline	BCRF
All	49.810	**50.299**	77.271	**77.726**	62.088	**62.412**
Things	46.247	**46.547**	77.819	**78.555**	**59.205**	59.002
Stuff	52.402	**53.028**	76.872	**77.122**	64.186	**64.892**

Table 5: **Results on Cityscapes dataset.** Panoptic segmentation results on the validation set.

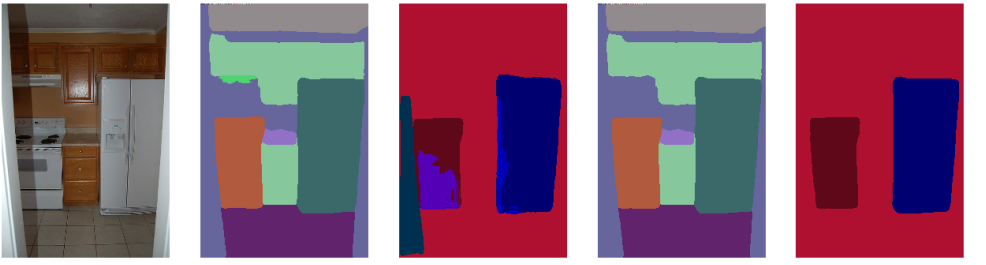

Figure 3: **Visualisation on COCO Dataset.** Columns left to right: original image, semantic output before BCRF, instance output before BCRF, semantic output after BCRF, instance output after BCRF.

6.6 Cross Potentials

Our BCRF module allows the network to learn complex class-aware relationships between the semantic and instance features belonging to each class. While there is room for it to learn a simple logical relationship, the variation of parameters learnt in Figure 4 verifies that a complex class-specific mapping has been learned by the network. For example, the high value for dining table to sofa and low value for boat to dining table (Figure 4 left) corresponds to the likelihood of finding such objects together in the natural image distribution in Pascal dataset, acting as an attention mechanism on regions of the image.

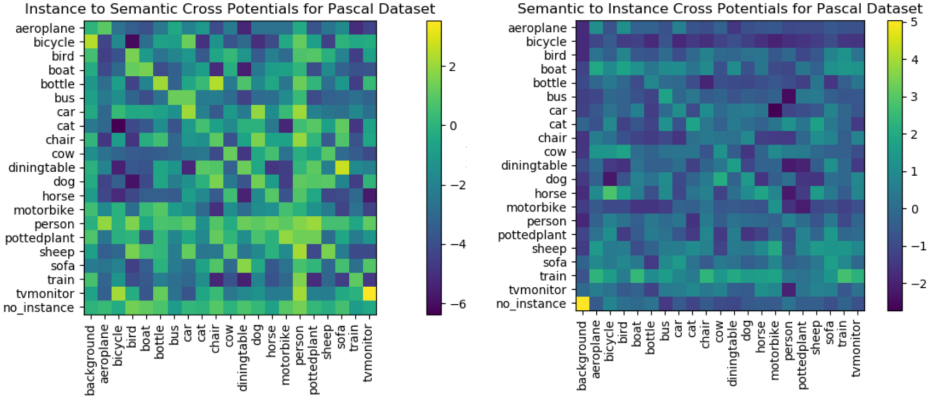

Figure 4: **Heatmap illustrating inter-class dependencies learned by BCRF.** It shows how important logits belonging to each class in one branch are for predicting each class in the other branch. Instance classes on x-axis and semantic on y-axis. Note that a logarithmic scale has been used for the legend.

7 Conclusion

We proposed a probabilistic graphical model based framework for panoptic segmentation. Our BCRF model, containing two different kinds of random variables, is capable of optimally combining the predictions from a semantic segmentation model and an instance segmentation model to obtain a good panoptic segmentation. We use different energy functions in our BCRF to encourage the spatial, appearance, and instance-semantic consistency of the panoptic segmentation. An iterative mean field algorithm is then used to find the panoptic labeling that approximately maximizes the conditional probability of the labeling given the image. We further show that the proposed BCRF framework can be used as an embedded module within a deep neural network to obtain superior results in panoptic segmentation.

References

[1] Anurag Arnab and Philip H.S Torr. Pixelwise Instance Segmentation with a Dynamically Instantiated Network. In *CVPR*, 2017.

[2] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip H. S. Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision (ECCV)*, 2016.

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. 03 2019.

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *PAMI*, 2018.

[5] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab. *ArXiv*, abs/1910.04751, 2019.

[6] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. *ArXiv*, 09 2019.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.

[8] Xuming He and Stephen Gould. An exemplar-based crf for multi-instance object segmentation. In *CVPR*, pages 296–303, 2014.

[9] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019.

[11] Koller, Daphne and Friedman, Nir. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[12] P. Krähenbühl and V. Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *NIPS*, 2011.

[13] P. Krähenbühl and V. Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials - Supplementary Material. In *NIPS*, 2011.

[14] J Lazarow et al. Learning instance occlusion for panoptic segmentation.

[15] Qizhu Li, Anurag Arnab, and Philip H. S. Torr. Weakly- and semi-supervised panoptic segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

[16] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *CVPR*, 2015.

[19] Lorenzo Porzi, Samuel Rota Bulò, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE TPAMI*, 2017.

[21] Marvin Teichmann and Roberto Cipolla. Convolutional CRFs for semantic segmentation. *BMVC*, 2019.

[22] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene parsing with object instances and occlusion ordering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3748–3755, 09 2014. doi: 10.1109/CVPR.2014.479.

[23] Philip H.S. Torr Vibhav Vineet, Jonathan Warrell. Filter-based Mean-Field Inference for Random Fields with Higher-Order Terms and Product Label-Spaces. In *ECCV*, 2012.

[24] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. 2020.

[25] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[26] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] Tien-Ju Yang, Maxwell D. Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *ArXiv*, abs/1902.05093, 2019.

[28] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *ICLR*, 2016.

[29] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P.H.S. Torr. Conditional Random Fields as Recurrent Neural Networks. In *ICCV*, 2015.