Anurag Arnab, Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Måns Larsson, Alexander Kirillov, Bogdan Savchynskyy, Carsten Rother, Fredrik Kahl, and Philip H.S. Torr

Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation

Combining probabilistic graphical models with deep learning for structured prediction

Semantic segmentation is the task of labeling every pixel in an image with a predefined object category. It has numerous applications in scenarios where the detailed understanding of an image is required, such as in autonomous vehicles and medical diagnosis. This problem has traditionally been solved with probabilistic models known as *conditional random fields* (*CRFs*) due to their ability to model the relationships between the pixels being predicted. However, deep neural networks (DNNs) recently have been shown to excel at a wide range of computer vision problems due to their ability to automatically learn rich feature representations from data, as opposed to traditional handcrafted features. The idea of combining CRFs and DNNs have achieved state-of-the-art results in a number of domains. We review the literature on combining the modeling power of CRFs with the representation-learning ability of DNNs, ranging from early work that combines these two techniques as independent stages of a common pipeline to recent approaches that embed inference of probabilistic models directly in the neural network itself. Finally, we summarize future research directions.

Introduction

Scene understanding is a long-standing problem in the field of computer vision and involves developing algorithms to interpret the contents of images with the level of comprehension of a human. Perceptual tasks, such as visual scene understanding, are performed effortlessly by humans. However, replicating the visual cortex on a computer has proven to be a challenging problem that has yet to be completely solved, almost 50 years after it was first posed by an undergraduate student at the Massachusetts Institute of Technology as a summer research project [1].

There are many ways of describing a scene, and current computer vision research addresses most of these problems independently, as illustrated in Figure 1. A high-level summary of a scene can be obtained by predicting image tags that describe the objects in the picture (such as *person*) or the scene (such as *city* or *office*). This task is known as *image classification*. The object detection task, on the other hand, aims

Digital Object Identifier 10.1109/MSP.2017.2762355
Date of publication: 9 January 2018

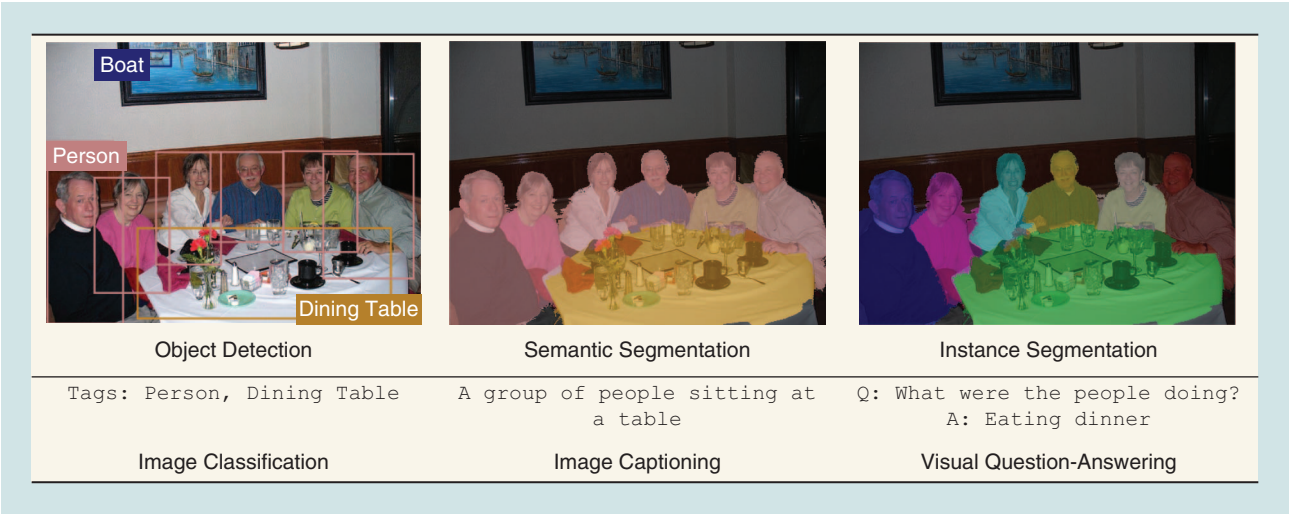

FIGURE 1. An example of various scene understanding tasks. Some tasks, such as image classification, provide a high-level description of the image by classifying whether certain tags exist. Other tasks like object detection, semantic segmentation, and instance segmentation provide more detailed and localized information about the scene. Researchers have also begun to bridge the gap between natural language processing and computer vision with tasks such as image captioning and visual question-answering.

to localize different objects in an image by placing bounding boxes around each instance of a predefined object category. Semantic segmentation, the main focus of this article, aims for a more precise understanding of the scene by assigning an object category label to each pixel within the image. Recently, researchers have also begun tackling new scene understanding problems such as instance segmentation, which aims to assign a unique identifier to each segmented object in the image, as well as bridging the gap between natural language processing and computer vision with tasks such as image captioning and visual question-answering, which aim at describing an image in words, and answering textual questions from images, respectively.

Scene understanding tasks, such as semantic segmentation, enable computers to extract information from real-world scenarios and leverage this information to accomplish given tasks. Semantic segmentation has numerous applications including the following:

- autonomous vehicles, which need a precise, pixel-level understanding of their environment
- developing robots, which can navigate and manipulate objects in their environment
- diagnosing medical conditions by segmenting cells, tissues, and organs of interest
- image- and video-editing and developing "smart glasses," which describe the scene to the blind.

Semantic segmentation has traditionally been approached using probabilistic models known as CRFs, which explicitly model the correlations among the pixels being predicted. However, in recent years, DNNs have been shown to excel at a wide range of computer vision and machine-learning problems, as they can automatically learn expressive feature representations from massive data sets. Despite the representational power of DNNs, state-of-the-art segmentation algorithms, which are benchmarked on public computer vision data sets and evaluation servers where the test set is withheld, all include a CRF within their pipelines. Some approaches include CRFs as a separate stage of the pipeline, while the leading ones incorporate it within the neural network itself.

In this article, we review CRFs and DNNs in the context of dense, pixelwise prediction tasks and explain how CRFs can be incorporated into neural networks to combine the advantages of these two models. Markov random fields, CRFs, and, more generally, probabilistic graphical models are ubiquitous tools with a long history of applications in a variety of domains spanning computer vision, computer graphics, and image processing [2]. This is due to their ability to model correlations in the variables being predicted. DNNs, on the other hand, are also fast becoming the de facto method of choice in a variety of machine-learning tasks as they can learn rich feature representations automatically from data. It is, therefore, a natural idea to combine CRFs with neural networks in a joint framework—an approach that has been successful in a number of domains. From a theoretical perspective, it is interesting to investigate the connections between CRFs and DNNs and explore how inference algorithms for probabilistic graphical models can be framed as neural networks themselves. We review CRFs and DNNs and their integration in the remainder of this article.

CRFs

A naïve way of performing dense prediction tasks like semantic segmentation is to classify each pixel independently using some features derived from the image. However, such independent pixelwise classification often produces unsatisfactory results that are inconsistent with the visual features in the image. For example, an independent pixelwise classification can predict a few spurious incorrect labels in the middle of a blob of pixels that are classified to have the same label (e.g., a

few *dog* pixels in the middle of a blob that is classified as *cat*, as shown in Figure 2). To predict the label of each pixel, a local classifier uses a small spatial context in the image [as shown by the patch in Figure 2(c)], and this often leads to noisy predictions. Better results can be obtained by acknowledging that we are predicting a structured output and explicitly modeling the problem to include our prior knowledge about a good pixelwise prediction result. For example, we know that objects are usually continuous and thus we expect nearby pixels to be assigned the same object label. CRFs are models that are widely used to achieve this. In the following, we provide a tutorial introduction to CRFs in the semantic image segmentation setting.

CRFs are a classical tool for modeling complex structures consisting of a large number of interrelated parts. In the aforementioned example of image segmentation, these parts correspond to separate pixels. Each pixel u is associated with a finite set of its possible states $L = \{l_1, l_2, \ldots, l_L\}$, modeled by a variable $X_u \in L$. In the example in Figure 2, these finite states are the labels that can be assigned to each pixel, i.e., $L = \{\text{person, cat, dog, background}\}$. Each state has an associated unary cost $\psi_u(X_u = x \mid I)$, which has to be paid to assign label x to the pixel u, given the image I. This unary cost is typically obtained from a classifier, as described in the previous paragraph, and with only the unary cost, we would be performing independent, per-pixel predictions. To model interactions between pixels, pairwise costs are introduced. The value $\psi_{u,v}(X_u = x, X_v = y \mid I)$ is the pairwise cost for assigning a pair of labels x and y to the pixels u and v, respectively. In semantic segmentation, a common pairwise cost to use is the Potts model (derived from statistical mechanics) where the cost is zero when two neighboring pixels have the same label, and λ (a real, positive scalar) if they have different labels. This cost encourages nearby pixels to take on the same label and is based on our prior knowledge that objects are generally continuous. This pairwise weight, λ, typically depends on other features in the image—for example, the cost is higher if pixels are closer to each other in terms of spatial coordinates or if the appearance of two pixels is similar (by comparing pixel intensity values in an appropriate color space). Costs can also be defined over more than two simultaneously interacting variables, in which case they are known as *higher-order potentials*. However, we ignore these in the rest of this section for simplicity.

In terms of graph theory, a CRF can be understood as a graph $(V, \mathcal{E})$, with nodes V corresponding to the image pixels, and edges $\mathcal{E}$ connecting those node pairs for which a pairwise cost is defined. The following graphs are common in segmentation literature: 1) four- or eight-grid graphs (which we will denote as the *grid CRF*), where only neighboring pixels in the image are connected by graph edges and 2) fully connected graphs, where all pairs of pixels are connected by edges [Figure 3(b) and (c)]. Intuitively, grid graphs (such as the four-grid graph in Figure 2) can only propagate information to a limited number of neighbors. By contrast, fully connected graphs enable long-range interactions between pixels, which can lead to more precise segmentations as shown in Figure 3. However, grid graphs were traditionally favored in segmentation systems [3], [9], [10] since there exist efficient inference algorithms to solve the corresponding segmentation problem.

Let X_u be the variable associated with the node $u \in V$ and let $\mathbf{X}$ be the vector formed by the X_u variables under some ordering of V. An assignment $\mathbf{x}$ to $\mathbf{X}$ is known as a *configuration* or a *labeling*, i.e., a configuration assigns a label to each node in the CRF. Inference of the CRF involves finding a configuration $\mathbf{x}$, such that the total unary and pairwise costs, also called the *energy*, are minimized. The corresponding problem is called *energy minimization for CRFs*, where the energy is given by

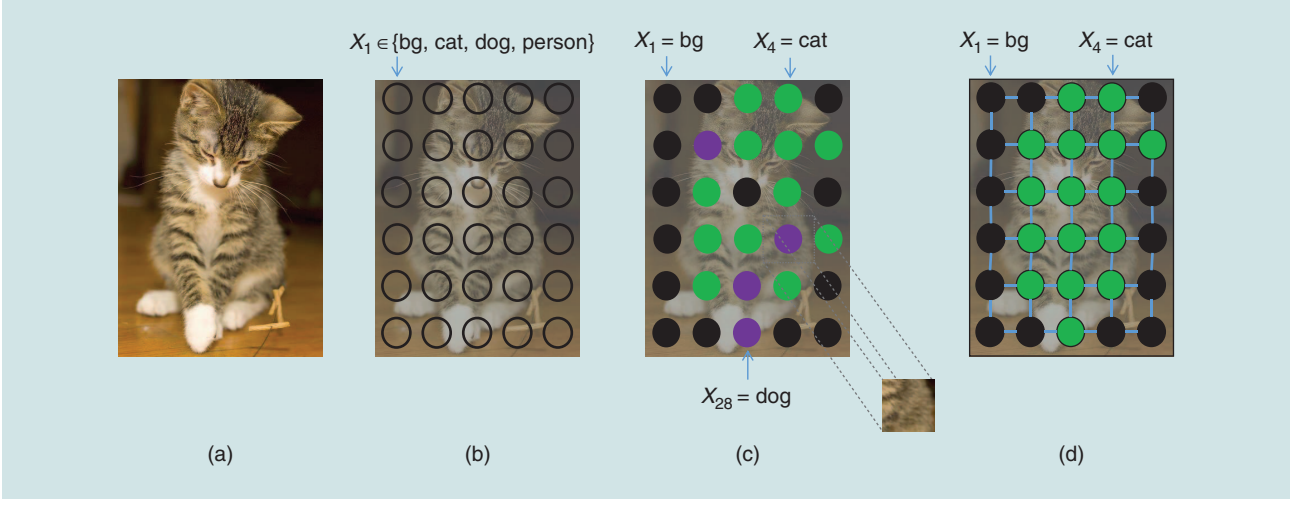

FIGURE 2. An overview of semantic segmentation. (a) The input image we wish to segment. Every pixel, u, is associated with a random variable X_u, which takes on a label from a predefined set, as shown in (b). A naive method of performing this would be to train a classifier to predict the semantic label of each pixel, using features derived from the image. However, as we can see from (c), this tends to result in very noisy segmentations. This is because, to predict a pixel, the classifier would only have access to local pixel features, which are often not discriminative enough. By taking the relationship between different pixels into account (such as the fact that if a pixel has been labeled *cat*, those nearby pixels are likely to take the same label since cats are continuous objects) with a CRF, we can improve our labeling, as shown in (d).

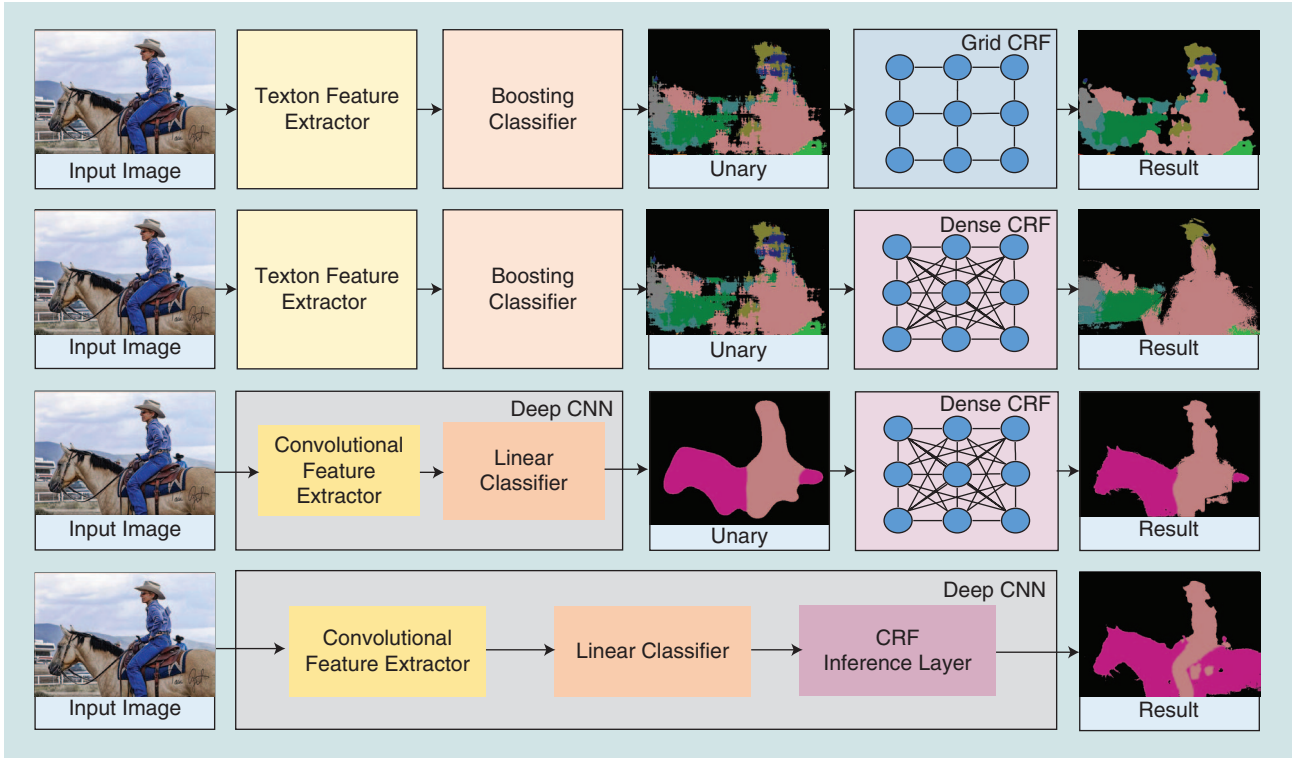

FIGURE 3. The evolution of semantic segmentation systems. (a) shows the early "TextonBoost" work [3] that computed unary potentials using Texton [3] features, and a CRF with limited eight-connectivity. (b) presents the DenseCRF work of Krahenbuhl and Koltun [4] that used densely connected pairwise potentials and approximate mean-field inference. The more expressive model achieved significantly improved performance. As shown in (c), numerous works, such as [5], have replaced the early handcrafted features with DNNs, which can learn features from large amounts of data, and used the outputs of a CNN as the unary potentials of a DenseCRF model. In fact, works such as [6] showed that unary potentials from CNNs (without any CRF) on their own achieved significantly better results than previous methods. Current state-of-the-art methods, as shown in (d), have all combined inference of a CRF within the DNN itself, allowing for end-to-end training [7], [8]. Result images for this figures were obtained using the publicly available code of [4]–[7] and [9]. CNN: convolutional neural network.

$$E(\mathbf{x},I) := \sum_{u\in V} \psi_u(X_u = x_u \mid I) + \sum_{\{u,v\}\in\mathcal{E}} \psi_{u,v}(X_u = x_u,\ X_v = x_v \mid I). \tag{1}$$

Although the energy minimization problem is NP-hard [11], a number of exact and approximate algorithms exist to obtain acceptable solutions (see [12] for an overview). Exact algorithms typically apply to only special cases of the energy, while approximate algorithms efficiently find a solution to a simplification of the original problem. In particular, the most popular methods for image segmentation were initially based on the reduction of the energy minimization problem or its parts to the st-min-cut problem [13]. However, the complexity of these algorithms grow as the graph becomes more dense.

By contrast, fully connected graphs with a specific type of pairwise cost can be efficiently (albeit approximately) addressed by mean-field algorithms, as detailed in the next section. These fully connected graphs are now the most common model in segmentation.

The learning problem involves estimating cost functions based on a training set $(I^{(k)}, \mathbf{x}^{(k)})_{k=1}^{n}$ consisting of n pairs of images and corresponding ground-truth labelings. The costs must be designed in a way that the inference performed for the image I returns a labeling that is close to the ground truth $\mathbf{x}$. Since statistical parameter estimation builds a basis for the learning theory, one defines a probability distribution $p(\mathbf{x}\mid I) = 1/Z(I)\exp\{-E(\mathbf{x},I)\}$ on the set of labelings. Here $Z(I) = \Sigma_{\mathbf{x}}\exp\{-E(\mathbf{x},I)\}$ is a normalization factor known as the *partition function*, which ensures that the distribution sums to one (to make it a probability distribution).

A popular learning problem formulation is based on the maximum likelihood principle and consists in finding costs ψ_u and $\psi_{u,v}$ that maximize the probability of the training set $(I^{(k)},\mathbf{x}^{(k)})_{k=1}^{n}$. The crucial subproblem, which determines the computational complexity of such estimation, consists of computing marginal probabilities for each label x_u in each node u and the pair (x_u, x_v) of labels in each pair of nodes (u,v), connected by an edge in the associated graph. The marginal probabilities build sufficient statistics for the distribution p and therefore need to be computed to perform learning. To compute the marginal probability $p_u(x_u)$ for the label x_u in the graph node u, one has to perform the summation $\Sigma_{\mathbf{x}':x'_u=x_u}p(\mathbf{x}')$ over all possible labelings where node u takes on label x_u. Combinatorial complexity of this problem arises because the number of such labelings is exponentially large and the explicit summation is typically computationally infeasible. Moreover, it is a well-known result [14] stating that this problem is NP-hard, or, loosely speaking, there is no hope for a reasonably fast algorithm being able to perform

such computations in general. A number of approaches have been proposed to approximate these computations. Next, we review the ones that are most popular.

Mean-field inference facilitating dense pairwise terms

Although computing the marginal probabilities is a hard problem in general, it can be easily done in certain special cases. In particular, it is an easy task if the graphical model only contains unary terms (and therefore no edges). In this case, all marginal probabilities are simply inversely proportional to the unary costs. Distributions corresponding to such graphs without edges are called *fully factorized*. The mean-field approach approximates the distribution $p(\mathbf{x} \mid I)$ for a given image I with a fully factorized distribution, $Q(\mathbf{x})$. The approximation is done by minimizing the Kullback–Leibler divergence between these two distributions. This minimization constitutes a nonconvex problem, therefore only a local minimum is typically found by optimization algorithms.

In practice, it is particularly important that there is an efficient algorithm to this end in the special case, when the graphical model is associated with a fully connected graph and pairwise costs $\psi_{uv}(x_u, x_v \mid I)$ for each pair of labels form a Gaussian distribution (up to a normalizing constant) for $x_u \neq x_v$ and otherwise equal to zero. This model was first introduced by Krähenbühl and Koltun [4] and is known as *DenseCRF*. The pairwise costs were the sum of two Gaussian kernels: a Gaussian blurring filter, $\exp\left(-\| p_u - p_v \|^2 / 2\sigma_\gamma^2\right)$, and an edge-preserving bilateral filter, $\exp\left(\left(-\| p_u - p_v \|^2 / 2\sigma_\alpha^2\right) - \left(\| I_u - I_v \|^2 / 2\sigma_\beta^2\right)\right)$, where p_u and p_v denote the positions of pixels u and v, I_u and I_v the color intensity vectors of these pixels, and σ the bandwidth of these filters. Although approximate mean-field inference was used, the model was more expressive compared to previous grid CRFs and thus achieved significantly improved results with the faster run time (facilitated by fast Gaussian filtering techniques [15]). As a result, this has become the de facto CRF model for most segmentation tasks and has achieved the best performance on multiple public benchmarks.

Stochastic sampling-based estimation of marginals

Another approach for approximating marginal probabilities is based on sampling, in particular, on the Gibbs sampling method [16]. In this case, instead of computing the sum over all labelings, one samples such labelings and computes frequencies of all configurations in each node. The advantage of this approach is that, in theory, the estimated marginals eventually converge to the true ones.

Variational approximations

The problem of computing marginals can be reformulated as a minimization problem [17]. Although this minimization problem is still as difficult as computing marginal probabilities, efficient approximations of the objective function exist—and this approximation of the objective function can be minimized quickly. In the section "Learning Arbitrary Potentials in CRFs," we return to these approximations in the context of joint training of DNN and CRF models.

In this section, we have introduced CRFs, a common probabilistic graphical model used in semantic segmentation. The performance of these models are, however, influenced heavily by the unary term produced by the classifier. Convolutional neural networks (CNNs) have proven to be excellent classifiers in a variety of tasks, and we review their application to dense prediction tasks in the next section.

CNNs for dense prediction

DNNs have recently been shown to excel in a number of machine-learning tasks, and most problems within the computer vision community are now solved by a class of these neural networks known as *CNNs*. A detailed overview of neural networks can be found in [18], with which we share our terminology. In contrast to previous classification algorithms, which require the user to manually design discriminative features, neural networks are able to automatically learn features from data when trained with stochastic gradient descent (SGD) (or one of its variants) to minimize a training objective function. While the backpropagation algorithm (a method for efficiently computing gradients of the parameters of a neural network with respect to an objective function, for use in conjunction with optimization via SGD) for training neural networks has been used since the 1980s [19], it was the emergence of large-scale data sets such as ImageNet [20] and the parallel computational power of graphics processing units (GPUs) that have enabled neural networks to become very successful in most machine-learning tasks. This became apparent to the computer vision community during the 2012 ImageNet Challenge, where the only entry using a neural network, AlexNet [21], achieved the best performance by a significant margin.

AlexNet [21], like many subsequent neural network architectures, is composed of a sequence of convolutional layers followed by ReLU nonlinearities and pooling layers. A sequence of convolutional filters allows a neural network to learn hierarchical representations of images where complex patterns are composed of simpler patterns captured in earlier stages of the network. Convolutional layers are common in computer vision tasks since they preserve spatial information, and also since they are translationally equivariant—they have the same response at different parts of the image.

The last layers of successful CNN architectures [21]–[23] are typically inner-product or *fully connected* layers (as common in traditional multilayer perceptrons [19], [24] which used these layers throughout the network). These layers consider all features in the input to make the final prediction in the case of image classification, and the final fully connected layer can be thought of as a linear classifier (operating on the nonlinear features produced by preceding layers of the neural network). The final output of the network is a C-dimensional vector, where C is the number of classes and each element in the vector represents the probability of each class appearing in the image. Hence, a typical CNN designed for image classification can be thought of as a function which maps an image of a fixed size to a C-dimensional vector of probabilities of each class appearing in the image.

Girshick et al. [25] showed that CNN architectures designed to excel in ImageNet classification could, with minimal

modifications, be adapted for other scene understanding tasks such as object detection and semantic segmentation. Furthermore, it was possible for Girschick et al. to fine-tune their network from a network already trained on ImageNet since most of the layers were the same. Fine-tuning from an existing ImageNet-pretrained model provided better parameter-initialization for training via backpropagation, and has been found to improve performance in many computer vision tasks. Therefore, the work of Girshick et al. suggested that CNNs for semantic segmentation should be based on ImageNet trained architectures as well.

A key idea to extending CNNs designed for image classification to other more complex tasks such as semantic segmentation is realizing that a fully connected layer can be considered as a convolutional layer, where the filter size is the same as the size of the input feature map [6], [26]. Long et al. [6] converted the fully connected layers of common architectures such as AlexNet [21] and VGG [22] into convolutional layers, and named them *fully convolutional networks* (*FCNs*). Since these networks consist of only convolutional-, pooling- and ReLU nonlinearity layers, they can operate on any arbitrarily sized image. However due to max-pooling in the network, the output would be a downsampled version of the input, as shown in Figure 4. Common architectures such as AlexNet [21], VGG [22] and ResNet [23] all consist of five pooling layers of size 2×2, and hence, the output is downsampled by a factor of 32 in these FCNs. Long et al. showed that even by simply bilinearly upsampling the coarse predictions up to the original size of the image, state-of-the-art performance at the time of publication could be achieved. This method is simple to implement, can be initialized with the parameters of a CNN trained on ImageNet, and then be fine-tuned on smaller data sets, which significantly improves results over initializing with random weights.

Although the fully convolutional approach of Long et al. achieved state-of-the-art performance, the predictions of the model were still quite coarse and "blobby," since the max-pooling stages in earlier parts of the network resulted in a lot of spatial information being lost. As a result, fine structures and object boundaries were usually segmented poorly. This has led to a lot of follow-up work on improving the segmentation performance of neural networks.

Chen et al. [5] used the outputs of a CNN as the unary potentials of a DenseCRF model, and showed that applying a CRF as postprocessing on these unaries could significantly improve results and provide sharper boundaries [as shown in Figure 3(c)]. In fact, the absolute performance improvement from applying DenseCRF on CNN unaries was greater than that of Textonboost [3] unaries [5]. Other works have improved the CNN architecture by addressing the loss of resolution caused by max-pooling. It is not possible to completely remove max-pooling from a CNN architecture for segmentation, since it will mean that layers deeper down will not have sufficient context or receptive field to make a good prediction. To combat this issue, Atrous [5] or Dilated [27] convolutions have been proposed (inspired by the "algorithme à trous" used in computing the undecimated wavelet transform [28]), which enables the receptive field of a convolution filter to be increased without increasing the number of parameters in

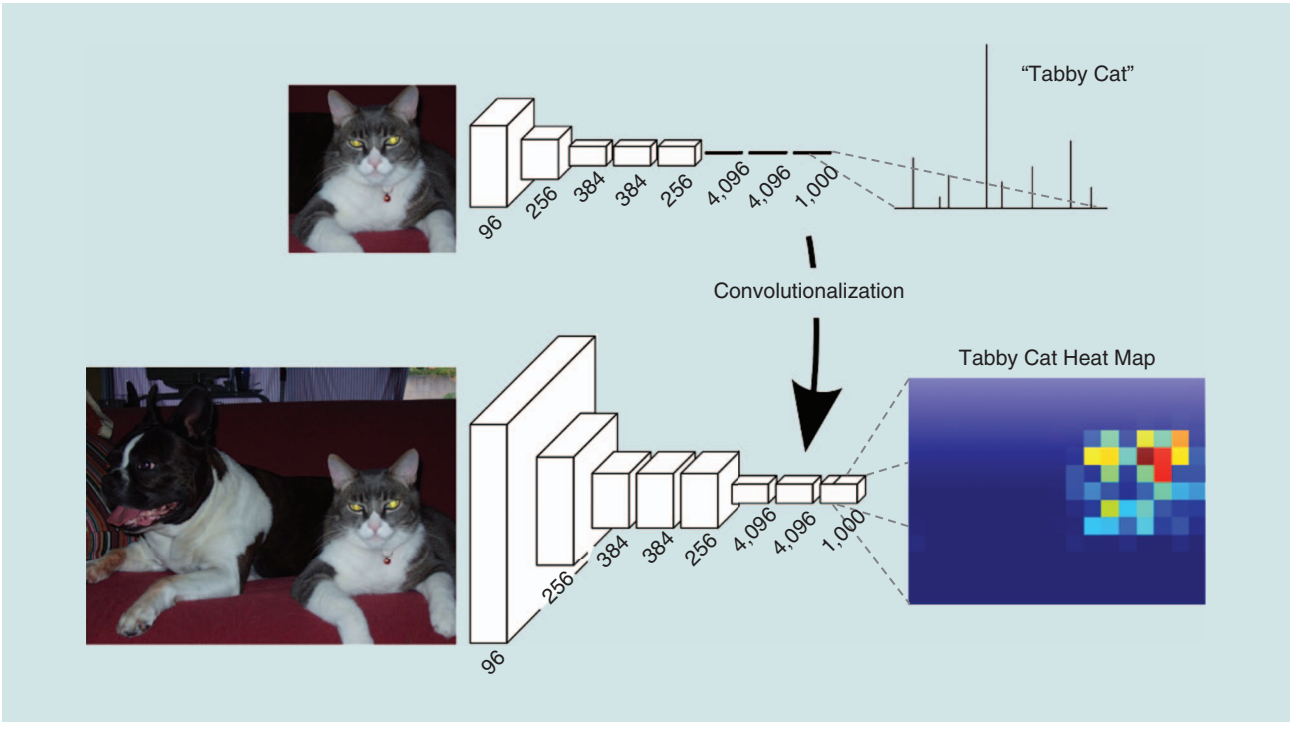

FIGURE 4. FCNs. Fully connected layers can easily be converted into convolutional layers by recognizing that a fully connected layer is simply a convolutional layer where the size of the convolutional filter and input feature map are identical. This enables CNNs trained for image classification to output a coarse segmentation when input a larger image. This simple method enables good initialization and efficient training of CNNs for pixelwise prediction. (Figure used with permission from [6].)

the filter. In these works, the last two max-pooling layers were removed, and Atrous convolutions were used thereafter to ensure a large receptive field. Note that it is not possible to remove all max-pooling layers in the network, due to the memory requirements of processing images at full resolution. Other works have learned more complex networks to upsample the low-resolution output of an FCN: in [29] an additional "decoder" network is learned, which progressively "unpools" the initial prediction to obtain the final full-resolution output. Ghiasi and Fowlkes [30] learn the basis functions with which to upsample in a coarse-to-fine architecture.

Although many architectural innovations have been proposed to improve the segmentation accuracy of neural networks, they have all benefited from additional refinement by a CRF. Furthermore, as Table 1 shows, algorithms that have achieved state-of-the-art results on public benchmarks such as Pascal VOC [31] have all incorporated CRFs as part of the neural network and trained it jointly with the unary part of the network end to end [8], [32], [33]. Similar trends are also being observed on the Cityscapes [34] and ADE20k [35] data sets, which were released in the last year. Intuitively, the improvement from these approaches stems from the fact that the parameters of the unary part of the network, and those of the CRF, may learn to optimally cooperate with each other.

The rest of this article focuses on these approaches that combine CRFs and CNNs in an end-to-end differentiable network. We elaborate on how mean-field inference of CRFs can be unrolled and interpreted as a recurrent neural network (RNN) in the section "CRFs as RNNs," and in the section "Learning Arbitrary Potentials in CRFs," we describe other approaches that enable arbitrary potentials to be learned.

Table 1. Results of recent algorithms on the Pascal VOC 2012 test set. Only the first submission, from 2012, does not use any deep learning. All of the other methods use a base CNN architecture derived from an ImageNet pretrained network. Evaluation is performed by a public server on a withheld test-set. The performance metric is the Intersection over Union (IoU) [31].

Method	IoU [%]	Base Network
Methods not using deep learning		
O2P [36]	47.8	–
Methods not using a CRF		
SDS [37]	51.6	AlexNet
FCN [6]	67.2	VGG
Zoom-out [38]	69.6	VGG
Methods using CRF for postprocessing		
DeepLab [5]	71.6	VGG
EdgeNet [39]	73.6	VGG
BoxSup [40]	75.2	VGG
Dilated Conv [27]	75.3	VGG
Centrale Boundaries [41]	75.7	VGG
DeepLab Attention [42]	76.3	VGG
LRR [30]	79.3	ResNet
DeepLab v2 [43]	79.7	ResNet
Methods with end-to-end CRFs		
CRF as RNNs [7]	74.7	VGG
Deep Gaussian CRF [8]	75.5	VGG
Deep parsing network (DPN) [44]	77.5	VGG
Context [32]	77.8	VGG
Higher-order CRF [33]	77.9	VGG
Deep Gaussian CRF [8]	80.2	ResNet

Mean-field inference as a neural network

Chen et al. [5] showed that state-of-the-art semantic segmentation results could be achieved by using the output of an FCN as the unary potentials of the DenseCRF model of [4]. However, the CRF was used as postprocessing, and FCN parameters were learned by backpropagation while CRF parameters were cross-validated (the authors tried a large number of different CRF parameters, and finally selected those which gave the highest performance on a validation set).

This section details how mean-field inference of a DenseCRF model can be incorporated into the neural network itself, as a separate "mean-field inference module," an idea that was developed concurrently by Zheng et al. [7] and Schwing and Urtasun [45]. This enables joint training of both the CNN and CRF parameters by backpropagation. Intuitively, we can expect better results from this approach as the CNN and CRF learn parameters which are compatible with each other due to the joint training. The cross-validation strategy of other works, such as [5], cannot update the parameters of the CNN such that they are optimal for the chosen CRF parameters. Zheng et al. named their approach *CRF-as-RNN*, and this achieved the best results when that paper was published.

Mean-field is an iterative algorithm, and crucially for optimization via SGD, the derivative of the output with respect to the input of each iteration can be calculated analytically. Therefore, we can unroll the inference algorithm across its time-steps, and form an RNN [18]. An RNN is a type of neural network, usually used to model sequential data, where the output of one iteration is used as the input of the next iteration and all iterations share the same parameters. In this case, the sequence is formed from the output of the iterative mean-field inference algorithm on each time step. When training the network, we can backpropagate through the RNN and into the previous CNN to optimize all parameters jointly. Furthermore, as shown in [7] and described next, for the DenseCRF model, the inference algorithm turns out to consist of standard CNN operations, making its implementation simple and efficient in standard neural network libraries. In the section "Incorporating Higher-Order Potentials," we describe how this idea can be extended beyond DenseCRF to other types of potentials, while the section "Other Examples of Unrolling Inference Algorithms in Neural Networks" mentions how the idea of unrolling inference algorithms has subsequently been employed in other domains using deep learning.

CRFs as RNNs

As mentioned in the section "Mean-Field Inference Facilitating Dense Pairwise Terms," mean field is an approximate inference method that approximates the true probability distribution, $P(\mathbf{X} \mid I)$, with a simpler distribution, $Q(\mathbf{X} \mid I)$. For notational simplicity, we omit the conditioning on the image, I, throughout the remainder of the article. In mean field, $Q(\mathbf{X})$ is assumed to be a product of independent marginals, $Q(\mathbf{X}) = \prod_u Q_u(X_u)$. The KL divergence between $P(\mathbf{X})$ and $Q(\mathbf{X})$ is then iteratively minimized. The maximum a posteriori (MAP) estimate of $P(\mathbf{X})$ is approximated as the MAP estimate of $Q(\mathbf{X})$. Since $Q(\mathbf{X})$ is fully factorized, the MAP estimate is simply the label that maximizes each independent marginal Q_u.

In the case of DenseCRF [4] (introduced in the section "Mean-Field Inference Facilitating Dense Pairwise Terms"), where the energy is of the form of (1), and the pairwise potentials are sums of Gaussian kernels,

$$\begin{aligned}\psi_{u,v}(x_u, x_v) &= \mu(x_u, x_v)\, k(\mathbf{f}_u, \mathbf{f}_v)\\ k(\mathbf{f}_u, \mathbf{f}_v) &= w^{(1)} \exp\left(-\frac{\|p_u - p_v\|^2}{2\sigma_\gamma^2}\right)\\ &\quad + w^{(2)} \exp\left(-\frac{\|p_u - p_v\|^2}{2\sigma_\alpha^2} - \frac{\|I_u - I_v\|^2}{2\sigma_\beta^2}\right),\end{aligned} \tag{2}$$

the mean-field update equations take the form

Algorithm 1. Mean-field inference for dense CRF [4], composed from common CNN operations.

$Q_u(l) \leftarrow \frac{1}{\sum_{l'} \exp(U_u(l'))} \exp(U_u(l))$ ▷ Initialization

while not converged **do**

$\tilde{Q}_u^{(m)}(l) \leftarrow \sum_{v \neq u} k^{(m)}(\mathbf{f}_u, \mathbf{f}_v)\, Q_v(l)$ for all m ▷ Message Passing

$\check{Q}_u(l) \leftarrow \sum_m w^{(m)} \tilde{Q}_u^{(m)}(l)$ ▷ Weighting Filter Outputs

$\hat{Q}_u(l) \leftarrow \sum_{l' \in L} \mu(l, l')\, \check{Q}_u(l')$ ▷ Compatibility Transform

$\check{Q}_u(l) \leftarrow U_u(l) - \hat{Q}_u(l)$ ▷ Adding Unary Potentials

$Q_u(l) \leftarrow \frac{1}{\sum_{l'} \exp(\check{Q}_u(l'))} \exp(\check{Q}_u(l))$ ▷ Normalizing

end while

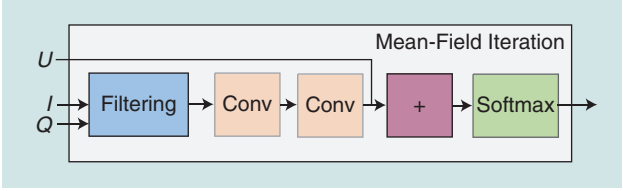

FIGURE 5. A mean-field iteration expressed as a sequence of common CNN operations. The updated equation of mean-field inference of a DenseCRF model [see (3)], can be broken down into a series of smaller steps, as shown in Algorithm 1. Note that not only are these steps all differentiable, but they are all standard neural network operations as well. This allows a single mean-field iteration to be efficiently implemented as a neural network.

$$Q_u(l) = \frac{1}{Z_u} \exp\left\{-\psi_u(l) - \sum_{l' \in L} \mu(l, l') \sum_{m=1}^{M} w^{(m)} \sum_{v \neq u} k^{(m)}(f_u, f_v)\, Q_v(l')\right\}. \tag{3}$$

The $\mu(\cdot,\cdot)$ function represents the compatibility of the labels assigned to variables X_u and X_v. In DenseCRF, the common Potts model (see the section "CRFs") was used, where $\mu(x_u, x_v) = 0$ if $x_u = x_v$ and 1 otherwise. The DenseCRF model has parallels with CNNs: in DenseCRF with Gaussian pairwise potentials [4], a Gaussian blurring filter and an edge-preserving bilateral filter are used to compute the pairwise term. The coefficients of the bilateral filter depend on the image itself (pixel intensity values), which differs from a convolution layer in a CNN, where the weights are fixed after training. Moreover, although the filter can potentially be as large as the image, it is parameterized only by its bandwidth. The pairwise potential is further parameterized by the weights of each filter, $w^{(1)}$ and $w^{(2)}$, and the label compatibility function, $\mu(\cdot,\cdot)$, which are both learned in the framework of [7].

Algorithm 1 shows how we can break this updated equation down into simpler steps [7]. Moreover, we can see that these steps all consist of common CNN operations and are all differentiable. Therefore, the authors were able to backpropagate through the mean-field inference algorithm and into the original CNN. This allowed them to jointly optimize the parameters of the CNN and the CRF and achieve the best-published results on the VOC data set at the time.

Expressing a mean-field iteration as a sequence of standard neural network operations

The message-passing step, which involves filtering the approximated marginals, Q, can be computed efficiently using fast-filtering techniques that are common in signal processing literature [15]. This was leveraged by [7] to reduce the computational complexity of this step from $O(N^2)$ (the complexity of a naïve implementation of the bilateral filter, where N is the number of pixels in the image) to $O(N)$. Computing the gradient of the output of the filtering step with respect to its input can also be performed using similar techniques. The next two steps—weighting filter outputs and the compatibility transform—can both be viewed as convolutions with a 1×1 kernel. In both cases, the parameters of these two steps were learned as the neural network was trained. The addition step is another common operation that is trivial to implement in a neural network. Finally, note that both the normalizing and initialization steps are equivalent to applying a softmax operation. This operation is ubiquitous in neural network literature as it is the activation function used in multinomial logistic regression.

The fact that Zheng et al. [7] viewed the compatibility transform as a convolutional filter whose weights were learned meant that they did not restrict themselves to the Potts model (see the section "CRFs"). This is in contrast to other methods such as [5], which cross-validated CRF parameters separately and assumed a Potts model, and is another reason for the improved performance of this method relative to the works published before it.

Mean-field inference as an RNN

One iteration of the mean-field algorithm can be formulated as sequence of common CNN layers as shown in Figure 5. By performing multiple mean-field iterations with the output of one iteration becoming the input of the next iteration, the mean-field inference algorithm can be formulated as an RNN, as shown in Figure 4(a). If we denote the unary potentials as U (the output of the initial CNN), then one mean-field iteration can be expressed as $Q^{t+1} = f_\theta(U, Q^t, I)$, where Q^t are the current estimation of the marginal probabilities and I is the image. The vector, θ denotes the parameters of the mean-field iteration, which are shared among all iterations. In the case of Zheng et al., they were the weights for the filter outputs, w, and the compatibility transform, $\mu(\cdot,\cdot)$ represented as a convolutional layer.

Q^0 is initialized as the softmax-normalized unary potentials (log probabilities) output by the initial CNN. Following the original DenseCRF work [4], Zheng et al. [7], computed a fixed number, T, of mean-field iterations. Thus, the final output of the module can be read as Q^T. In practice, $T = 5$ iterations were used by [7] as it was empirically observed that the mean-field had converged at this time. RNNs are known to be susceptible to the vanishing/exploding gradients problem [18]: computing the gradient of the output of one iteration with respect to its input requires multiplying by the parameters being learned. This repeated multiplication can cause the gradients to explode (if the parameters are greater than one) or vanish (if the parameters are less than one). However, the fact that only five iterations need to be performed in practice means that this problem is averted.

As shown in Figure 6, the final network implemented by Zheng et al. consists of an FCN [6], which predicts pixel-level labels without considering the structure of the output variables, followed by a CRF that can be trained end to end. The complete system therefore unites the strengths of CNNs—which can learn rich feature representations automatically from data—and CRFs—which can model the structure and correlations between the variables that are being predicted. Furthermore, parameters of both the CNN and CRF can be learned end to end via backpropagation.

In practice, Zheng et al. first trained an FCN (specifically the FCN8-s architecture of [6]) for semantic segmentation using the standard softmax cross-entropy loss. They then inserted the CRF-as-RNN layer and continued training their network, since it is necessary for the FCN part of the model to be initialized well before training the CRF. As shown in Table 1, the CRF-as-RNN method outperformed DeepLab [47] (which also uses an FCN, but uses the CRF as a postprocessing step) by 2%. In their paper, Zheng et al. [7] reported an improvement over just the FCN of 5.1%.

Incorporating higher-order potentials

The CRF-RNN framework considered the particular case of unrolling mean-field inference as an RNN for the DenseCRF model. Arnab et al. [33] showed that this framework could be extended to different types of higher-order potentials as well. Higher-order potentials (as mentioned in the section "CRFs"), model correlations between cliques of pixels larger than two pixels as in DenseCRF.

Arnab et al. [33] considered two different higher-order potentials: first, a detection potential was formulated, which encouraged consistency between the outputs of an object detector (i.e., Figure 1) and the final segmentation. This helped in cases where the segmentation unaries were poor and missed an object, but a complementary object detector had not. The potential was also formulated such that false detections could be ignored. A second potential was based on superpixels (a grouping of pixels into perceptually similar units, using specialized algorithms) and encouraged consistency over larger regions, helping to clean up spurious noise in the output. The mean-field updates for these more complex potentials were still differentiable, although they no longer consisted of commonly used neural network operations.

Higher-order potentials were shown to be effective in improving semantic segmentation performance before the adoption of deep learning [48]. In fact, potentials based on object detectors [49] and superpixels [50] had been proposed previously. However, by learning the parameters of these potentials jointly with the weights of an FCN, Arnab et al. [33] reduced the error of CRF-as-RNN by 12.6% on the VOC benchmark (Table 1).

Liu et al. [44] formulated another approach of incorporating higher-order relations. While [7] and [33] performed exact mean-field inference of a CRF as an RNN, Liu et al. approximated a single iteration of mean-field inference with a number of carefully designed convolution layers. Their deep parsing network (DPN) was also trained end to end, although like [7], they first pretrained the initial part of their network before finally training the entire network. As shown in Table 1, this approach achieved very competitive results similar to [33]. Figure 7 shows a comparison between FCN, CRF-as-RNN, DPN, and the higher-order CRF on a common image.

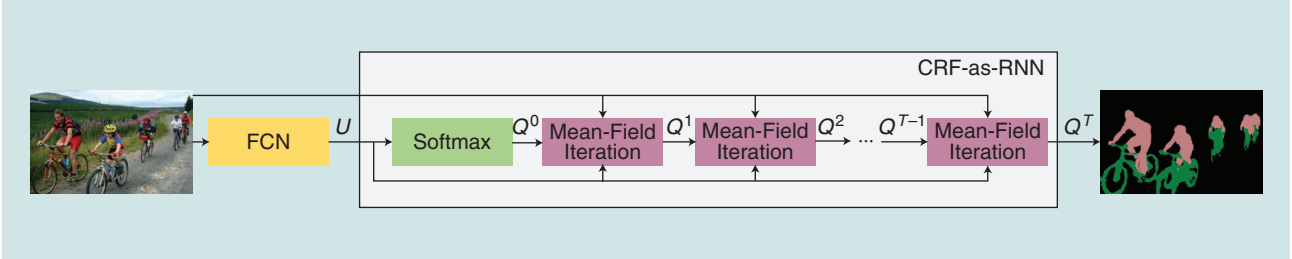

FIGURE 6. The final end-to-end trainable network of Zheng et al. [7]. The final system consists of an FCN [6] followed by a CRF. The authors showed that the iterative mean-field inference algorithm could be unrolled and seen as an RNN. This CRF inference module was named *CRF-as-RNN*.

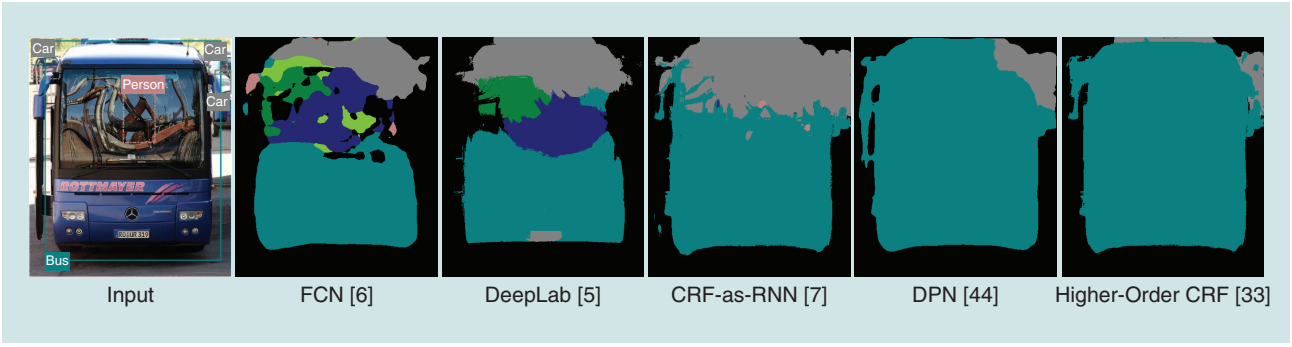

FIGURE 7. A comparison of various semantic segmentation methods. FCN tends to produce "blobby" outputs which do not respect the edges of the image (from the Pascal VOC validation set). DeepLab, which refines outputs of an FCN with DenseCRF, produces an output that is consistent with edges in the image. CRF-RNN and DPN both train a CRF jointly within a neural network; achieving better results than DeepLab. Unlike other methods, the higher-order CRF can recover from incorrect segmentation unaries since it also uses cues from an external object detector, while being robust to false-positive detections (like the incorrect "person" detection). Object detections produced by [46] have been overlaid on the input image, but only [33] uses this information.

Benefits of end-to-end training

An alternative to unrolling mean-field inference of a CRF and training the whole network end to end is to train only the "unary" part of the network and use the CRF as a postprocessing step whose parameters are determined via cross-validation. We refer to this approach as *disjoint* training of the CNN and the CRF, and show in Table 2 that end-to-end training outperforms disjoint training in the case of CRF-RNN [7], which only has pairwise potentials, and the higher-order CRF of [33].

Many recent works in the literature have used CRFs as a postprocessing step. However, as shown in Table 1, the best-performing ones have incorporated the CRF as part of the network itself. Intuitively, joint training of a CRF with a CNN allows the two modules to learn to optimally cooperate with each other.

Error analysis

Figures 3 and 7 show that the densely connected pairwise potentials of a CRF improve the segmentation quality at boundaries of objects. To quantify the improvements that CRFs make to the overall segmentation, we separately evaluate segmentation performance on the "boundary" and "interior" regions of the image, as done by [40] and [33]. As shown in Figure 8(c) and (d), we consider a narrow band (trimap [50]) around the "void" labels annotated in the VOC 2012 reduced validation set. The mean IoU of pixels lying within this band is termed the *boundary IoU* while the *interior IoU* is evaluated outside this region.

Figure 8 illustrates our results as the trimap width is varied for FCN [6], CRF-as-RNN [7], DPN [44], and higher-order CRF [33]. We can see that all of the CRF models improve the boundary IoU over FCN significantly, although CRF-as-RNN and higher-order CRF are almost identical. Moreover, all of the methods incorporating CRFs also show an improvement in the Interior IoU as well, indicating that the spatial and appearance consistency encouraged by CRFs is not limited to only improving segmentation at object boundaries. Furthermore, the higher-order potentials of [33] and [44] show a substantial increase in interior IoU over CRF-as-RNN, and an even bigger improvement over FCN. Higher-order potentials encourage consistency over larger regions of the image [33] and model contextual relationships between object classes [44]. As a result, they show larger improvements at the interior of objects being segmented.

Table 2. A comparison of mean IoU (%) obtained on the VOC 2012 reduced validation set from end to end and disjoint training (adapted from [33]).

Method	Mean IoU [%]
Unary only	68.3
Pairwise CRF trained disjointly	69.5
Pairwise CRF trained end to end	72.9
Higher-order CRF trained disjointly	73.6
Higher-order CRF trained end to end	75.8

Other examples of unrolling inference algorithms in neural networks

Unrolling inference algorithms as neural networks is a powerful idea beyond semantic image segmentation. Riegler et al. [51] presented a method for depth-map superresolution by unrolling the steps of an optimization algorithm and formulating it as an end-to-end trainable network. Recently, Wang et al. [52] extended deep structured models to continuous valued output variables, and addressed tasks such as image denoising and depth refinement.

Learning arbitrary potentials in CRFs

As the previous section shows, the joint training of a CNN and a CRF with Gaussian potentials is beneficial for the semantic segmentation problem. While it is able to output spatially and appearance-consistent results, some applications may benefit even more from the usage of more generic potentials. Indeed, general potentials are able to incorporate more sophisticated knowledge about a problem than are Gaussian potentials. As an

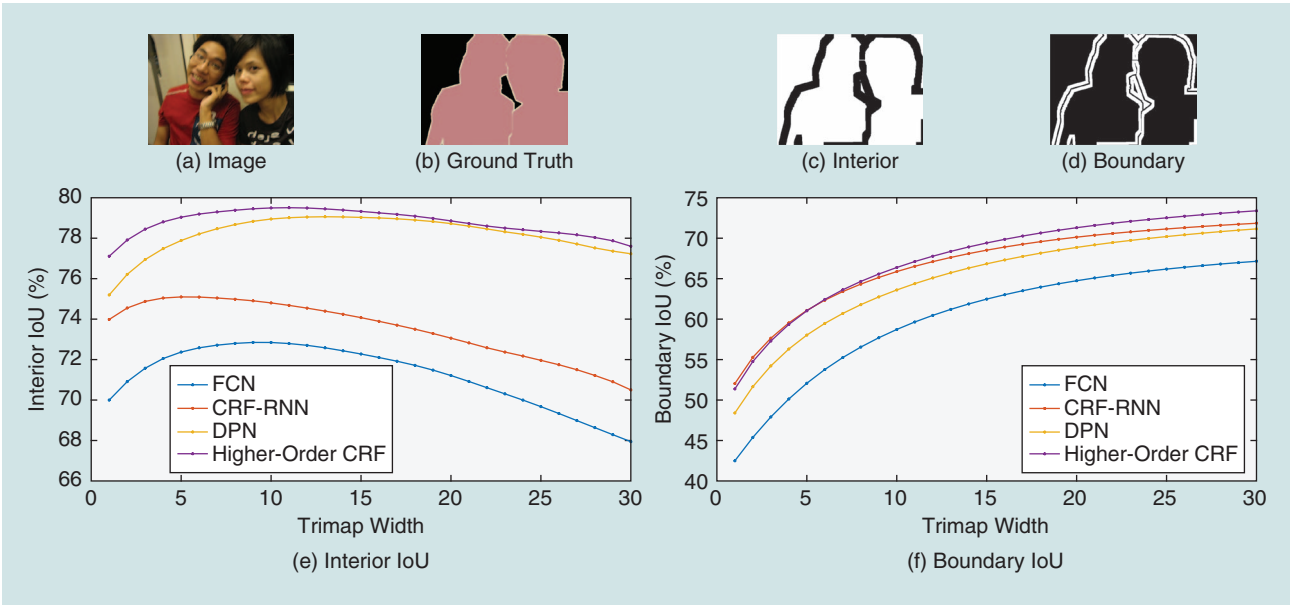

FIGURE 8. Error analysis on the VOC 2012 reduced validation set. An example image and its corresponding ground truth are shown in (a) and (b), respectively. The (c) interior and (d) boundary regions for this image are shown using a trimap width of nine pixels. Black regions in (c) and (d) are ignored in the IoU calculation. The IoU for interior and boundary regions, as a function of the trimap width, are shown in (e) and (f). (Figure is adapted from [33].)

example, for the human body part segmentation problem, parametric potentials may enforce a high-level structural constraint, i.e., the head should be located above the torso as seen in Figure 9(a). Moreover, these potentials can acquire this knowledge automatically during training. And by training a pixel-level CNN jointly with these potentials, a synergistic effect can be obtained, as observed in the section "CRFs as RNNs."

As mentioned in the section "CRFs," the main computational burden for probabilistic CNN-CRF learning consists in estimating marginal distributions for the configurations of individual variables and their pairs. Therefore, existing methods can be classified according to how they approximate these marginals. Each of the methods has different advantages and disadvantages and has a specific scope of problems where it is superior to the other. Next, we briefly review three such methods, that are applicable to learn arbitrary pairwise and unary costs. Finally, we also describe methods that are able to learn arbitrary pairwise costs without computing marginals.

Stochastic sampling-based training

This approach to learning is based on the stochastic estimation of marginal probability distributions, as described in the section "Stochastic Sampling-Based Estimation of Marginals." This method for training CNN-CRF models was proposed in [53] and applied to the problem of human body-part segmentation, where it was shown to outperform techniques based on DenseCRF as displayed in Figure 9. Advantages of this method are that it is simple and easy to parallelize. Along with a long training time, the disadvantages include the necessity to use very slow sampling-based energy minimization methods [16] for image segmentation during the prediction stage, when the neural network has already been trained. The slow prediction time is due to the fact that the best predictive accuracy is obtained when the training and prediction procedures are identical.

Piecewise training

For estimating marginal probabilities, this method replaces the marginal probabilities $p_v(x_v)$ (see the section "CRFs" for details) with values equal to the corresponding exponentiated costs $\exp\{-\psi_v(x_v)\}$ up to a normalizing constant. Although such a replacement is not grounded theoretically, it allows one to train parameters for each node and edge of the graph independently, which is computationally very efficient and easily parallelizable. Despite the lack of theoretical justification, the method was employed by Lin et al. [32] to give good practical results (it was on top of the Pascal VOC and Cityscapes leaderboards for several months) on numerous segmentation data sets, as reflected by the `Context` entry in Table 1.

Variational method-based training

As referenced in the section "Variational Approximations," the marginal probabilities were approximated with a variational technique during learning in [47]. In terms of theoretical justification the method can be positioned between stochastic training, as the most theoretically justified, and piecewise training, which lacks such a justification. Practically, the running time in the classification regime is far more acceptable than those of the stochastic method of [53]. However, the scalability of the method is questionable because of its large memory footprint: along with the training set one has to store a set of current values of the working variables. The size of this set is proportional to the number of training samples multiplied by 1) the number of nodes in a graphical model corresponding to each sample (which can be roughly estimated as a number of pixels in the corresponding image),

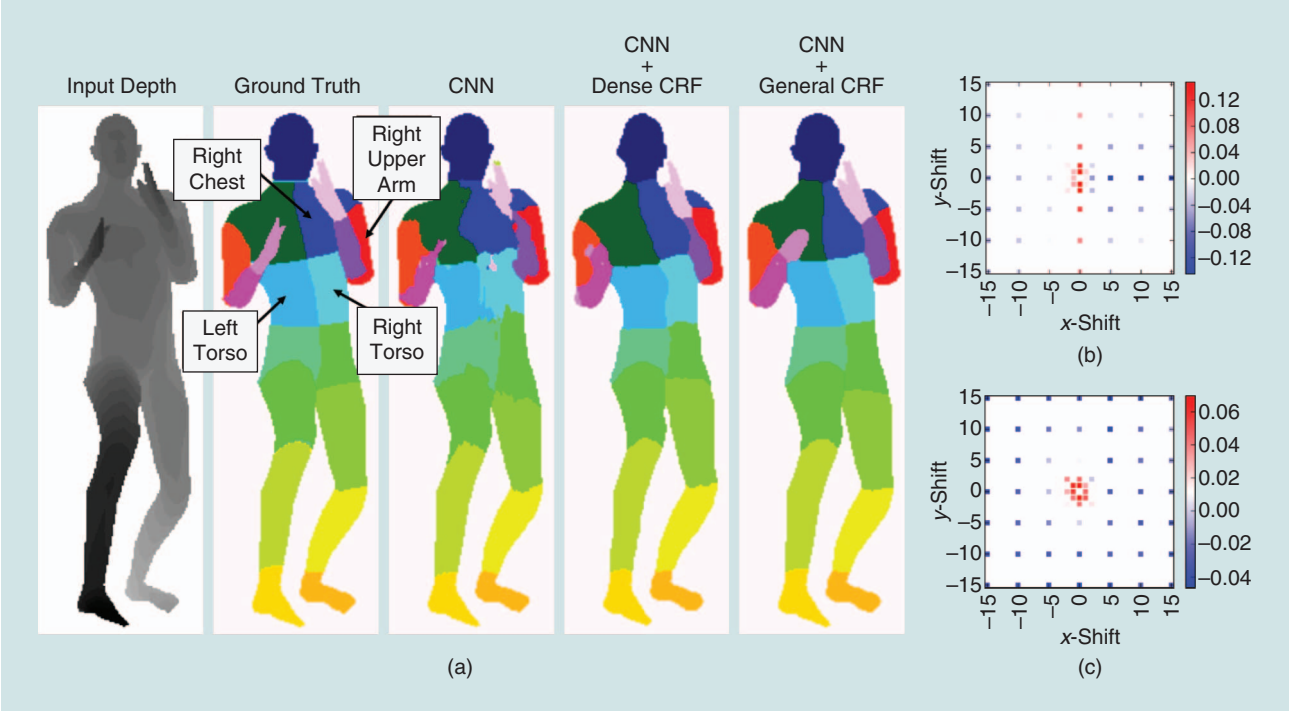

FIGURE 9. Human body parts segmentation with generic pairwise potentials. (a) From left to right: the input depth image; the corresponding ground-truth labeling for all body parts; the result of a trained CNN model; the result of CRF-as-RNN [7], where the pairwise potentials are a sum of Gaussian kernels; and the result of [53] that jointly train CNN and CRF with general parametric potentials. Note how the hands and elbows are segmented better. (b) Weights for pairwise potentials that connect the label "left torso" with the label "right torso." Red means a high energy value, i.e., a discouraged configuration, while blue means the opposite. The potentials enforce a straight, vertical border between the two labels, i.e., there is a large penalty for "left torso" on top (or below) of "right torso" (*x*-shift 0, *y*-shift arbitrary). Also, it is encouraged that "right torso" is to the right of the "left torso" (Positive *x*-shift and *y*-shift 0). (c) Weights for pairwise potentials that connect the label "right chest" with the label "right upper arm." It is discouraged that the "right upper arm" appears close to "right chest," but this configuration can occur at a certain distance. Since the training images have no preferred arm-chest configurations, all directions have similar weights. Such relations between different parts cannot by the Gaussian kernels used in CRF-as-RNN.

2) the number of edges in each graphical model, and 3) the number of possible variable configurations, which can be assigned to each graph node. In total, this typically requires one or even two orders of magnitude more storage than the training set itself. So far, this method has not been shown on large training sets.

Learning by backpropagating through inference

Formulating the steps of CRF inference as differentiable operations enable both the learning and inference of the CRF to be incorporated in a standard deep-learning framework, without having to explicitly compute marginals. An example of this was discussed in the section "CRFs as RNNs."

However, in contrast to that approach, we can look at the inference problem from a discrete optimization point of view. Here, it can be seen as an integer program where a given cost should be minimized while satisfying a set of constraints (each pixel should be assigned one label). An alternative inference approach is to do a continuous relaxation (allowing variables to be real-valued instead of integers) of this integer program and search for a feasible minimum. The solution to the original discrete problem can then be derived from the solution to the relaxed problem. Desmaison et al. [55] presented methods to solve several continuous relaxations of this problem. They showed that these methods generally achieve lower energies than mean-field-based approaches.

In the work of Larsson et al. [54], a projected gradient method based on only differential operations is presented. This inference method minimizes a continuous relaxation of the CRF energy and the weights can be learned by backpropagating the error derivative through the gradient steps. An example of the pairwise potentials learned by this method is shown in Figure 10. Chandra and Kokkinos [8] solve the energy minimization problem by formulating it as a convex quadratic program and finding the global minimum by solving a linear system. As seen in Table 1, it is the best-published approach on the PASCAL VOC 2012 data set at the time of this writing.

Methods based on discriminative learning

There is another discriminative learning technique that does not require the computation of marginal distributions. This class of methods formulate the learning as a structured support vector machine (SSVM), which is an extension to the support vector machine allowing structured output. Since solving these usually requires doing inference for several setups of weights an efficient inference method is crucial. Larsson et al. [56] utilized this for medical image segmentation doing inference with a highly efficient graph-cut method. Knöbelreiter et al. [57] proposes a very efficient GPU-parallelized energy minimization solver to efficiently perform inference for the stereo problem.

They show how learning can be made practically feasible by an SSVM formulation.

Training considerations

Optimizing DNNs in general requires careful selection of training hyperparameters, most notably the learning rate. The networks described in this article, which integrate CRFs and CNNs, are typically trained in a two-stage process [7], [8], [33], [44], [53], [54]: first, the "unary" part of the network is trained, and then the part of the network modeling inference of the CRF is appended and the network is trained end to end. It is not recommended to train from the beginning with the CRF inference layer, as the unaries produced by the first stage of the network are so poor that performing inference on them produces meaningless results while also increasing the computational time. An exception is Lin et al. [32], who train their entire network end to end. However, they have carefully chosen learning rate multipliers for different parts of their network. Liu et al. [44], on the other hand, initially train the unary part of the network and then have three separate training stages for each of the three modules in their network simulating mean-field inference of a CRF.

Advances in pixel-level prediction—along with holistic models, which address multiple scene understanding tasks—are bringing us closer to computers that understand our physical world and help enrich it.

Another important training detail for these models is the selection of the initial unary network to fine-tune end to end with the CRF inference layer. In the case of [7], [33], [54], allowing the initial unary network to converge and fine-tuning off that model tends to not produce the best results. Instead, the best performance is usually obtained by fine-tuning from a unary network that has not fully converged, in the sense that its validation error has not yet completely plateaued. Our intuition about why this happens is that once the parameters of the unary network converge to a local optimum, it is difficult for the learning algorithm to get out of this region.

Since CRF-as-RNN [7] is an iterative method, it is also necessary to specify the number of mean-field iterations to perform. The authors used five iterations of mean-field for training, and then increased this to ten iterations at test time. This choice is motivated by Figure 11, which shows empirical convergence results on the VOC 2012 reduced validation set—the majority of the improvement takes place after three iterations and the IoU begins to plateau after five iterations. Ten iterations are used at test time to obtain the best performance, while five iterations are used while training as these are enough iterations for the IoU to start plateauing and fewer iterations also decreases the training time.

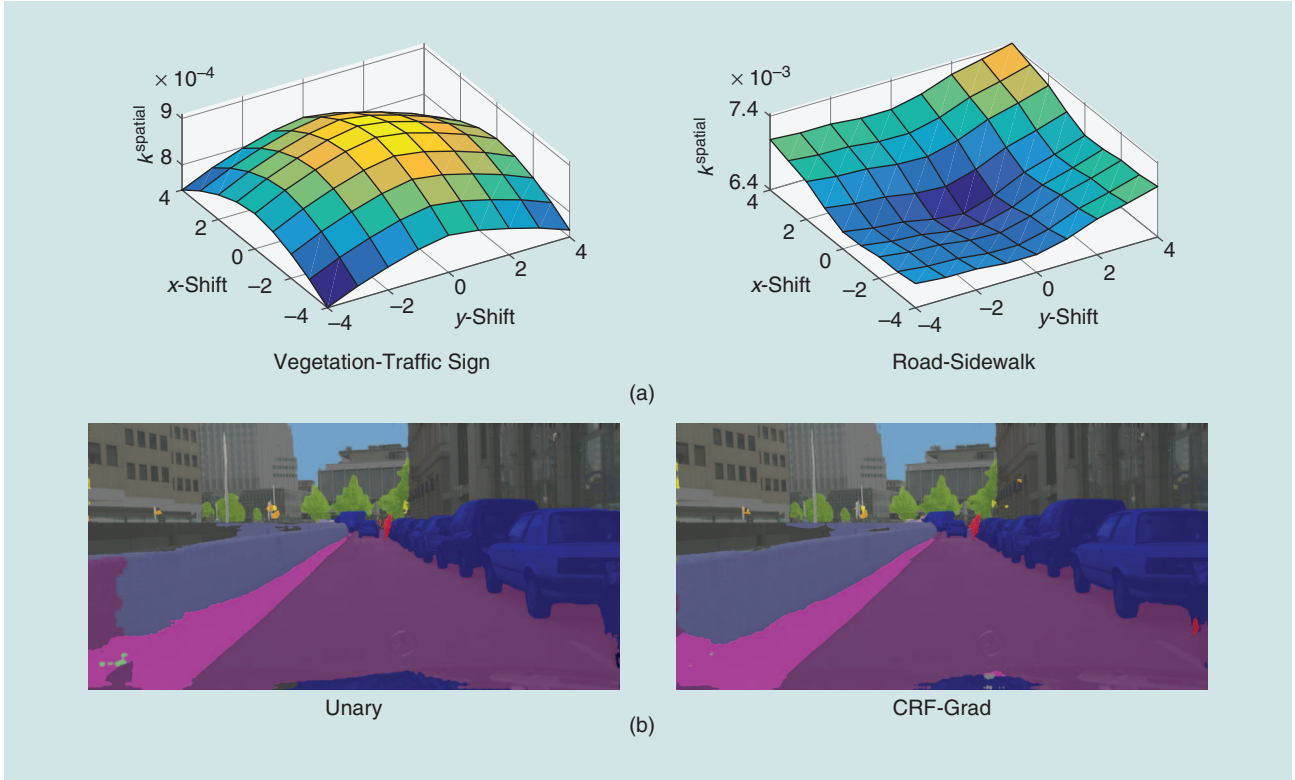

FIGURE 10. Pairwise potentials, represented as convolutional filters, learned by the method of [54]. These filters, shown in (a), model contextual relationships between classes, and their values can be understood as the energy added when setting one pixel to the first class (e.g., vegetation) and the other pixel with relative position (*x*-shift, *y*-shift) to the second class (e.g., a traffic sign). (b) shows an example result on the Cityscapes data set. The traffic lights and poles (which are challenging due to their limited spatial extent) are segmented better, and it does not label "road" being on top of the "sidewalk," due to the modeling of contextual relationships between classes.

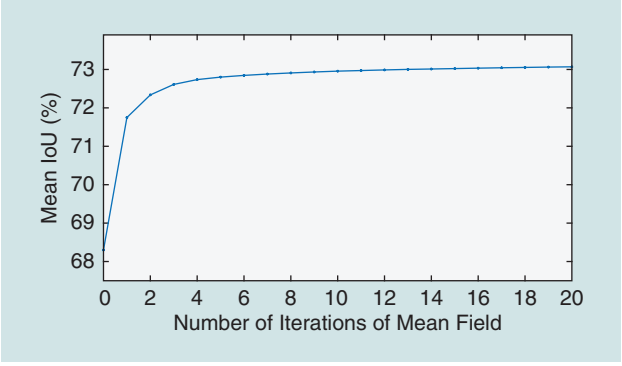

FIGURE 11. The empirical convergence of CRF-as-RNN [7]. The mean IoU on the Pascal VOC 2012 reduced validation set starts plateauing after five iterations of mean-field inference, supporting the authors' choice of training their network with five iterations of mean-field. The result at zero iterations is the mean IoU of only the unary network.

Conclusions and future directions

CRFs have a long history in structured prediction tasks in computer vision, such as semantic segmentation. DNNs, on the other hand, have recently been shown to achieve outstanding results due to their ability to automatically learn features from large data sets. In this article, we have discussed various approaches of how CRFs have been combined with DNNs to tackle pixel-labeling tasks such as semantic segmentation. The most basic method is to simply use the outputs of a DNN as the unary potentials of a well-studied CRF model (such as DenseCRF [4]) as a separate postprocessing step, e.g., [47]. We then discussed how the mean-field inference algorithm for CRFs could be formulated as an RNN, and thus be incorporated as another module or "layer" of an existing neural network. This method, however, was limited to only pairwise potentials of a specific form. Finally, we described several approaches to learning arbitrary potentials in CRFs. We also noted that the idea of unrolling inference algorithms as neural networks has since been applied in other fields as well.

As neural networks are universal function approximators, it is possible that network architectures could be designed that do not require explicit CRFs to model smoothness priors and achieve the same performance. This, however, remains an open research question as our understanding of DNNs is still very limited. Moreover, it is not clear if the smoothness priors incorporated by a CRF could be modeled by generic neural network layers as efficiently (in terms of the number of parameters). It is also an open question as to whether we need to develop more sophisticated training algorithms to achieve this.

The works described in this article have all been fully supervised learning scenarios, where large costs have been incurred in collecting data sets with per-pixel annotations. Given the substantial increase in segmentation performance on public benchmarks such as Pascal VOC [31], a future direction is to achieve similar accuracy levels using weakly supervised annotations (for example, image tags as annotation). In such scenarios with limited annotations, incorporating additional prior knowledge is of greater importance, and CRFs provide a method of doing so [58], [59]. Instance segmentation (Figure 1) is another emerging area of scene understanding research, and early works have incorporated end-to-end CRFs within their systems [60].

For some tasks (such as face detection on cameras), rough bounding boxes suffice. However, pixel-level understanding of a scene is required for tasks such as autonomous vehicles and medical diagnosis where detailed information is required. Advances in pixel-level prediction—along with holistic models, which address multiple scene understanding tasks—are bringing us closer to computers that understand our physical world and help enrich it.

Acknowledgments

This work was supported by grants ERC-2012-AdG 321162-HELIOS, ERC-647769, EPSRC Seebibyte EP/M013774/1, EPSRC/MURI EP/N019474/1, the Clarendon Fund, the Swedish Research Council (grant number 2016-04445), the Swedish Foundation for Strategic Research (Semantic Mapping and Visual Navigation for Smart Robots), and Vinnova/FFI (Perceptron, grant number 2017-01942).

Authors

Anurag Arnab (anurag.arnab@eng.ox.ac.uk) obtained his undergraduate degree in electrical and computer engineering from the University of Cape Town, South Africa, in 2014. He is currently a D.Phil. (Ph.D.) degree student at the University of Oxford, United Kingdom, under the supervision of Prof. Philip H.S. Torr. His research interests are in deep learning and probabilistic graphical models and their applications in scene understanding tasks such as semantic segmentation and instance segmentation.

Shuai Zheng (szheng@robots.ox.ac.uk) received his B.Eng. degree in information engineering from the Beijing Institute of Technology, China, in 2008 and his M.Eng. degree in pattern recognition and intelligent systems from the Institute of Automation, Graduate University of the Chinese Academy of Sciences, Beijing, in 2011, where he was supervised by Prof. Kaiqi Huang and Prof. Tieniu Tan. He received his Ph.D. degree in computer vision and machine learning from Oxford University, United Kingdom, in 2016 under the guidance of Prof. Philip H.S. Torr. His research interests involve deep learning and its applications in computer vision such as semantic segmentation, instance segmentation, image editing, biometrics, and visual search.

Sadeep Jayasumana (sadeep.jay@gmail.com) received his undergraduate degree in electronics and telecommunication engineering in 2010 from the University of Moratuwa, Sri Lanka. He received his Ph.D. degree in computer vision from the Australian National University, Canberra, in 2014, under the supervision of Prof. Richard Hartley. He was a postdoctoral research fellow in the Torr Vision Group at the University of Oxford, United Kingdom, from 2014 to 2016. He is a research scientist at Five AI, Cambridge, United Kingdom. His research interests include deep learning, semantic segmentation, and machine learning on manifold-valued data.

Bernardino Romera-Paredes (bernard24@gmail.com) received his B.Sc. degree in computer engineering from the

Universidad de Murcia, Spain, in 2009 and his M.Sc. degree in computational statistics and machine learning from University College London in 2010. He received his Ph.D. degree from University College London in 2014, supervised by Prof. Massimiliano Pontil and Prof. Nadia Berthouze. He was a postdoctoral research fellow in the Torr Vision Group at the University of Oxford, United Kingdom, from 2014 to 2016. He has published in top-tier machine-learning conferences such as Advances in Neural Information Processing Systems, the International Conference on Machine Learning, and the International Conference on Computer Vision. His research focuses on multitask and transfer learning methods applied to computer vision tasks. He is a research scientist at Google DeepMind.

Måns Larsson (mans.larsson@chalmers.se) received his B.Sc. degree in engineering physics and his M.Sc. degree in complex adaptive systems from Chalmers University of Technology, Sweden, in 2012 and 2015, respectively, where he is currently a Ph.D. student under the supervision of Prof. Fredrik Kahl. His research interests include deep learning, probabilistic graphical models, and semantic segmentation tasks in both medical image analysis and computer vision.

Alexander Kirillov (alexander.kirillov@iwr.uni-heidelberg.de) received his specialist degree in applied mathematics and computer science from Lomonosov Moscow State University, Russia, in 2013. He is a Ph.D. student at the Universität Heidelberg, Germany, jointly supervised by Prof. Carsten Rother and Prof. Dmitry Vetrov. He has published in top computer vision and machine-learning venues such as the Conference on Computer Vision and Pattern Recognition, the International Conference on Computer Vision, and Advances in Neural Information Processing Systems. His main research interests are deep-learning models for structured output and energy minimization problems for computer vision applications.

Bogdan Savchynskyy (bogdan.savchynskyy@tu-dresden.de) received his bachelor's and master's degrees in applied mathematics from the National Technical University of Ukraine "Kyiv Polytechnic Institute" in 2000 and 2002, respectively. He received his Ph.D. degree in computer science (systems and tools of artificial intelligence) in 2007 from the National Academy of Sciences of Ukraine. Currently, he is a senior researcher at the Universität Heidelberg, Germany. He was awarded the 2003 President of Ukraine Scholarship for Young Scientists and the 2014 Best Student Paper Award from the Conference on Computer Vision and Pattern Recognition. He received funding on the project Exact Relaxation-Based Inference in Graphical Models from the German Research Foundation. His research interests include mathematical optimization problems in machine learning and computer vision.

Carsten Rother (carsten.rother@iwr.uni-heidelberg.de) received the diploma degree with distinction in 1999 from the University of Karlsruhe, Germany, and the Ph.D. degree in 2003 from the Royal Institute of Technology Stockholm, Sweden, both in computer science. He is a full professor at the Universität Heidelberg, Germany. His research interests are the field of computer vision and machine learning, such as segmentation, matching, object recognition, and pose estimation, with applications to natural science, in particular biology and medicine. He won the Best Paper (honorable mention) Award at the International Conference on Computer Vision and Pattern Recognition 2005, the Asian Conference on Computer Vision, and the Association for Computing Machinery Conference on Human Factors in Computing Systems 2007. He was awarded the DAGM Olympus Prize in 2009.

Fredrik Kahl (fredrik.kahl@chalmers.se) received his M.Sc. degree in computer science and technology in 1995 and his Ph.D. degree in mathematics in 2001, both from Lund University, Sweden. His thesis was awarded the Best Nordic Thesis Award in pattern recognition and image analysis 2001–2002 at the Scandinavian Conference on Image Analysis 2003. He received the Marr Prize at the 2005 IEEE International Conference on Computer Vision. In 2007, he was awarded an ERC starting grant from the European Research Council, which supported his team during 2008–2013. He is currently a joint professor at Chalmers University of Technology and Lund University, both in Sweden.

Philip H.S. Torr (philip.torr@eng.ox.ac.uk) completed his undergraduate degree in pure mathematics from the University of Southampton, United Kingdom, in 1990. He received his Ph.D. degree in computer vision in 1995 from Oxford University, United Kingdom. He remained at Oxford for three years as a research fellow before becoming a research scientist at Microsoft Research for six years, first in Redmond, Washington, then in Cambridge, United Kingdom, founding the vision side of the Machine Learning and Perception Group. He has won awards from several top vision conferences, including the International Conference on Computer Vision, the Conference on Computer Vision and Pattern Recognition, the European Conference on Computer Vision, Advances in Neural Information Processing Systems, and the British Machine Vision Conference. He is currently a professor at Oxford University.

References

[1] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press, 1969.

[2] A. Blake, P. Kohli, and C. Rother, *Markov Random Fields for Vision and Image Processing*. Cambridge, MA: MIT Press, 2011.

[3] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multiclass object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vision*, vol. 81, no. 1, pp. 2–23, 2009.

[4] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *Proc. Neural Information Processing Systems Conf.*, 2011, pp. 109–117.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learning Representations*, 2015.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2015, pp. 3431–3440.

[7] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1529–1537.

[8] S. Chandra and I. Kokkinos, "Fast, exact and multi-scale inference for semantic image segmentation with deep Gaussian CRFs," in *Proc. IEEE European Conf. Computer Vision*, 2016, pp. 402–418.

[9] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr, "Associative hierarchical CRFs for object class image segmentation," in *Proc. IEEE Int. Conf. Computer Vision*, 2009, pp. 739–746.

[10] X. He, R. Zemel, and M. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2004, pp. 695–702.

[11] M. Li, A. Shekhovtsov, and D. Huber, "Complexity of discrete energy minimization problems," in *Proc. IEEE European Conf. Computer Vision*, 2016, pp. 834–852.

[12] J. H. Kappes, B. Andres, F. A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. X. Kausler, T. Kröger, J. Lellmann, N. Komodakis, B. Savchynskyy, and C. Rother, "A comparative study of modern inference techniques for structured discrete energy minimization problems," *Int. J. Comput. Vision*, vol. 115, no. 2, pp. 1–30, 2015.

[13] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 2, pp. 147–159, 2004.

[14] A. Bulatov and M. Grohe, "The complexity of partition functions," *Theoretical Comput. Sci.*, vol. 348, no. 2–3, pp. 148–186, 2005.

[15] A. Adams, J. Baek, and M. A. Davis, "Fast high-dimensional filtering using the permutohedral lattice," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 753–762, 2010.

[16] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, no. 6, pp. 721–741, 1984.

[17] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Machine Learn.*, vol. 1, no. 1–2, pp. 1–305, 2008.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error-propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1. Cambridge, MA: MIT Press, 1986, pp. 318–362.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Information Processing Systems Conf.*, 2012, pp. 1106–1114.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations*, 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016, pp. 770–778.

[24] F. Rosenblatt, "Principles of neurodynamics. Perceptrons and the theory of brain mechanisms," DTIC Document, Cornell Aeronautical Lab, Inc., Buffalo, NY, Tech. Rep. AD0256582, 1961.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2014, pp. 580–587.

[26] A. Giusti, D. C. Ciresan, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks," in *Proc. IEEE Int. Conf. Image Processing*, 2013, pp. 4034–4038.

[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learning Representations*, 2015.

[28] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*, J.-M. Combes, A. Grossmann, and P. Tchamitchian, Eds. New York: Springer, 1990, pp. 286–297.

[29] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1520–1528.

[30] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *Proc. IEEE European Conf. Computer Vision*, 2016, pp. 519–534.

[31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[32] G. Lin, C. Shen, A. van den Hengel, and I. D. Reid, "Exploring context with deep structured models for semantic segmentation," arXiv Preprint, arXiv:1603.03183, 2016.

[33] A. Arnab, S. Jayasumana, S. Zheng, and P. Torr, "Higher order conditional random fields in deep neural networks," in *Proc. IEEE European Conf. Computer Vision*, 2016, pp. 524–540.

[34] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016, pp. 3213–3223.

[35] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20k dataset," arXiv Preprint, arXiv:1608.05442, 2016.

[36] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Free-form region description with second-order pooling," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 37, no. 6, pp. 1177–1189, 2015.

[37] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. IEEE European Conf. Computer Vision*, 2014, pp. 297–312.

[38] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2015, pp. 3376–3385.

[39] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016, pp. 4545–4554.

[40] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1635–1643.

[41] I. Kokkinos, "Pushing the boundaries of boundary detection using deep learning," in *Proc. Int. Conf. Learning Representations*, 2016.

[42] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016, pp. 3640–3649.

[43] L.-C. 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," arXiv Preprint, arXiv:1606.00915, 2016.

[44] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Deep learning Markov random field for semantic segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, to be published.

[45] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," arXiv Preprint, arXiv:1503.02351, 2015.

[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Neural Information Processing Systems Conf.*, 2015, pp. 91–99.

[47] L.-C. Chen, A. G. Schwing, A. L. Yuille, and R. Urtasun, "Learning deep structured models," in *Proc. IEEE Int. Conf. Machine Learning*, 2015, pp. 1785–1794.

[48] V. Vineet, J. Warrell, and P. H. Torr, "Filter-based mean-field inference for random fields with higher-order terms and product label-spaces," in *Proc. IEEE European Conf. Computer Vision*, 2012, pp. 31–44.

[49] C. Wojek and B. Schiele, "A dynamic conditional random field model for joint labeling of object and scene classes," in *Proc. IEEE European Conf. Computer Vision*, 2008, pp. 733–747.

[50] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vision*, vol. 82, no. 3, pp. 302–324, 2009.

[51] G. Riegler, M. Rüther, and H. Bischof, "ATGV-Net: Accurate depth super-resolution," in *Proc. IEEE European Conf. Computer Vision*, 2016, pp. 268–284.

[52] S. Wang, S. Fidler, and R. Urtasun, "Proximal deep structured models," in *Proc. Neural Information Processing Systems Conf.*, 2016, pp. 865–873.

[53] A. Kirillov, D. Schlesinger, S. Zheng, B. Savchynskyy, P. Torr, and C. Rother, "Joint training of generic CNN-CRF models with stochastic optimization," in *Proc. 13th Asian Conf. Computer Vision,* 2016, pp. 221–236.

[54] M. Larsson, A. Arnab, F. Kahl, S. Zheng, and P. H. S. Torr, "A projected gradient descent method for CRF inference allowing end-to-end training of arbitrary pairwise potentials," presented at the 11th Int. Conf. Energy Minimization Methods Computer Vision Pattern Recognition, 2017.

[55] A. Desmaison, R. Bunel, P. Kohli, P. H. Torr, and M. P. Kumar, "Efficient continuous relaxations for dense CRF," in *Proc. IEEE European Conf. Computer Vision*, 2016, pp. 818–833.

[56] M. Larsson, J. Alvén, and F. Kahl, "Max-margin learning of deep structured models for semantic segmentation," in *Proc. 20th Scandinavian Conf. Image Analysis*, 2017, pp. 28–40.

[57] P. Knöbelreiter, C. Reinbacher, A. Shekhovtsov, and T. Pock, "End-to-end training of hybrid CNN-CRF models for stereo," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2017, pp 2339–2348.

[58] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. IEEE European Conf. Computer Vision,* 2016, pp. 695–711.

[59] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *Proc. IEEE European Conf. Computer Vision,* 2016, pp. 413–432.

[60] A. Arnab and P. H. S. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2017, pp. 441–450.

SP