

DAG-Recurrent Neural Networks For Scene Labeling

Bing Shuai*, Zhen Zuo*, Bing Wang, Gang Wang †

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

{bshuai001, zzuol, wanggang, wang0775}@ntu.edu.sg

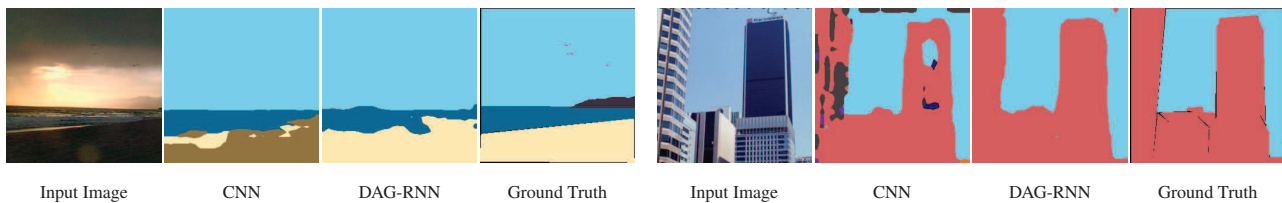


Figure 1: (Best viewed in color) With the local representations extracted from Convolutional Neural Networks (CNNs), the ‘sand’ pixels (in the first image) are likely to be misclassified as ‘road’, and the ‘building’ pixels (in the second image) are easy to get confused with ‘streetlight’. Our DAG-RNN is able to significantly boost the discriminative power of local representations by modeling their contextual dependencies. As a result, it can produce smoother and more semantically meaningful labeling maps.

Abstract

In image labeling, local representations for image units are usually generated from their surrounding image patches, thus long-range contextual information is not effectively encoded. In this paper, we introduce recurrent neural networks (RNNs) to address this issue. Specifically, directed acyclic graph RNNs (DAG-RNNs) are proposed to process DAG-structured images, which enables the network to model long-range semantic dependencies among image units. Our DAG-RNNs are capable of tremendously enhancing the discriminative power of local representations, which significantly benefits the local classification. Meanwhile, we propose a novel class weighting function that attends to rare classes, which phenomenally boosts the recognition accuracy for non-frequent classes. Integrating with convolution and deconvolution layers, our DAG-RNNs achieve new state-of-the-art results on the challenging SiftFlow, CamVid and Barcelona benchmarks.

1. Introduction

Scene labeling refers to associating one of the semantic classes to each pixel in a scene image. It is usually defined as a multi-class classification problem based on their surrounding image patches. However, some classes may be in-

distinguishable in a close-up view. As an example in Figure 1, the ‘sand’ and ‘road’ pixels are hard to be distinguished even for humans with limited context. In contrast, their differentiation becomes conspicuous when they are considered in the global scene. Thus, how to equip local features with a broader view of contextual awareness is a pivotal issue in image labeling.

In this paper, recurrent neural networks (RNNs) [10][18] are introduced to address this issue by modeling the contextual dependencies of local features. Specifically, we adopt undirected cyclic graphs (UCGs)¹ to represent the connectivity of pixels in images. Due to the loopy property of UCGs, RNNs are not directly applicable to UCG-structured images. Thus, we decompose the UCG to several directed acyclic graphs (DAGs). In other words, an UCG-structured image is represented by the combination of several DAG-structured images. Then, we develop the DAG-RNNs, a generalization of RNNs [9][10], to process DAG-structured images. Each hidden layer is generated independently through applying DAG-RNNs to the corresponding DAG-structured image, and they are aggregated to produce the context-aware feature maps. In this case, the local representations are able to embed the abstract gist of the image, so their discriminative power are enhanced remarkably.

We integrate the DAG-RNNs with the convolution and deconvolution layers, thus giving rise to an end-to-end trainable full labeling network. Functionally, the convolution

*Equal contribution

†Corresponding author

¹UCG refers to the undirected graph in which there exists a closed walk with no repetition of edges.

layer transforms RGB raw pixels to compact and discriminative representations. Based on them, the proposed DAG-RNNs model the contextual dependencies of local features, and output the improved context-aware representation. The deconvolution layer upsamples the feature maps to match the dimensionality of the desired outputs. Overall, the full labeling network accepts variable-size images and generates the corresponding dense label prediction maps in a single feed-forward network pass. Furthermore, considering that the class frequency distribution is highly imbalanced in natural scene images, we propose a novel class weighting function that attends to rare classes.

We test the proposed labeling network on three popular and challenging scene labeling benchmarks (SiftFlow [15], CamVid [2] and Barcelona [28]). On these datasets, we show that our DAG-RNNs are capable of greatly enhancing the discriminative power of local representations, which leads to dramatic performance improvements over baselines (CNNs). Meanwhile, the proposed class weighting function is able to boost the recognition accuracy for rare classes. Most importantly, our full labeling network significantly outperforms current state-of-the-art methods.

Next, related works are firstly reviewed, compared and discussed in Section 2. Section 3 elaborates the details of the DAG-RNNs and how they are applied to image labeling. Besides, it presents the details of the full labeling network and the class weighting function. The detailed experimental results and analysis are presented in Section 4. In the end, section 5 concludes the paper.

2. Related Work

Scene labeling (also termed as scene parsing, semantic segmentation) is one of the most challenging problems in computer vision. It has attracted more and more attention in recent years. Here we would like to highlight and discuss three lines of works that are most relevant to ours.

The first line of work is to explore the contextual modeling. One attempt is to encode context into local representation. For example, Farabet et al. [8] stacks surrounding contextual windows from different scales; Pinheiro et al. [19] increases the size of input windows. Sharma et al. [20] adopts recursive neural networks to propagate global context to local regions. However, they do not consider any connectivity structure for image units, thus their correlations are not effectively captured. In contrast, we interpret the structure of images as UCGs, within which the connections of image units allow the DAG-RNNs to explicitly model their dependencies. Another attempt is to pass context to local classifiers by building probabilistic graphical models (PGM). For example, Shotton et al. [21] formulates the unary and pairwise features in a 2nd-order Conditional Random Field (CRF). Krähenbühl et al. [12] build a fully connected graph to enforce higher order la-

beling coherence. Zheng et al.[36] transforms the efficient Conditional Random Fields (CRF) [12] to a neural network, so the inference of CRF equals to applying the same neural network recurrently until some fixed point (convergence) is reached. Shuai et al.[22] transfers the global-order dependencies in a non-parametric framework to disambiguate the local confusions. Our work also differs from them. First, the label dependencies are defined in terms of compatibility functions over cliques in PGMs, while such dependencies are modeled through a recurrent weight matrix in RNNs.

Some of the previous work exploit ‘recurrent’ ideas in a different way. They generally refer to applying the identical model recurrently at different iterations (layers). For example, Pinheiro et al.[19] attaches the RGB raw data with the output of the Convolutional Neural Network (CNN) to produce the input for the same CNN in the next layer. Tu et al.[30] augments the patch feature with the output of the classifier to be the input for the next iteration, and the classifier parameters are shared across different iterations. Our work differs from them significantly. They model the context in the form of intermediate outputs (usually local beliefs), which implicitly encodes the neighborhood information. In contrast, the contextual dependencies are modeled explicitly in DAG-RNNs by propagating information via the recurrent connections.

Recurrent neural networks (RNNs) have achieved great success in temporal dependency modeling for chain-structured data, such as natural language and speeches. Zuo et al. [38] applies 1D-RNN to model weak contextual dependencies in image classification. Graves et al. [9] generalizes 1D-RNN to multi-dimensional RNN (MDRNN) and applies it to offline arabic handwriting recognition. Shuai et al.[23] also adopts 2D-RNN to real-world image labeling. Recently, Tai et al.[27] and Zhu et al.[37] demonstrate that considering tree structure (constituent / parsing trees for sentences) is beneficial for modeling the global representation of sentences. Our proposed DAG-RNN is a generalization of chain-RNNs [1][11], tree-RNNs [27][37] and 2D-RNNs [9][23], and it enables the network to model long-range semantic dependencies for graphical structured images. The most relevant work to ours is [23]. Comparing with it, (1), we generalize 2D-RNN to DAG-RNN and show benefits in quantitative labeling performance; (2), we integrate the convolution layer, deconvolution layer with our DAG-RNNs to a full labeling network; and (3), we adopt a novel class weighting function to address the extremely imbalanced class distribution issue in natural scene images. Moreover, the proposed network achieves state-of-the-art performance on a variety of scene labeling benchmarks.

3. Approach

To densely label an image I , our full labeling network process the image sequentially: (1), Convolution layer pro-

duces the corresponding feature map \mathbf{x} . Each feature vector in \mathbf{x} summarizes the information from a local region in I . (2), DAG-RNNs model the contextual dependency among elements in \mathbf{x} , and generates the intermediate feature map \mathbf{h} , whose element is a feature vector that implicitly embeds the abstract gist of the image. (3), Deconvolution layer [16] upsamples the feature maps. From which, the dense label prediction maps are derived. We start by introducing the proposed DAG-RNNs, and the details of the full network are elaborated in the following sections.

3.1. RNNs Revisited

A recurrent neural network (RNN) is a class of artificial neural network that has recurrent connections, which equip the network with memory. In this paper, we focus on the Elman-type network [7]. Specifically, the hidden layer $\mathbf{h}^{(t)}$ in RNNs at time step t is expressed as a non-linear function over current input $\mathbf{x}^{(t)}$ and hidden layer at previous time step $\mathbf{h}^{(t-1)}$. The output layer $\mathbf{y}^{(t)}$ is connected to the hidden layer $\mathbf{h}^{(t)}$. Mathematically, given a sequence of inputs $\{\mathbf{x}^{(t)}\}_{t=1:T}$, an Elman-type RNN operates by computing the following hidden and output sequences:

$$\begin{aligned} \mathbf{h}^{(t)} &= f(U\mathbf{x}^{(t)} + W\mathbf{h}^{(t-1)} + b) \\ \mathbf{y}^{(t)} &= g(V\mathbf{h}^{(t)} + c) \end{aligned} \quad (1)$$

where U, W are weight matrices between the input and hidden layers, and among the hidden units themselves, while V is the output matrix connecting the hidden and output layers; b, c are corresponding bias vectors and $f(\cdot), g(\cdot)$ are element-wise nonlinear activation functions. The local information $\mathbf{x}^{(t)}$ is progressively stored in the hidden layers by applying Equation 1. In other words, the contextual information (the summarization of past sequence information) is explicitly encoded into local representation $\mathbf{h}^{(t)}$.

Training a RNN can be achieved by optimizing a discriminative objective with a gradient-based method. Back Propagation through time (BPTT) [32] is usually used to calculate the gradients. This method is equivalent to unfolding the network in time and using back propagation in a very deep feed-forward network except that the weights across different time steps (layers) are shared.

3.2. DAG-RNNs

The aforementioned RNN is designed for chain-structured sequential data (e.g. sentences or speeches), where temporal dependency is modeled. However, the connectivity structure of image units are beyond chain. In other words, traditional chain-structured RNNs are not suitable for images. For instance, we can reshape the feature tensor $\mathbf{x} \in \mathbb{R}^{h \times w \times d}$ to $\hat{\mathbf{x}} \in \mathbb{R}^{(h \cdot w) \times d}$, and generate the chain representation by connecting contiguous elements in $\hat{\mathbf{x}}$. Such a structure loses spatial relationship of image units, as two

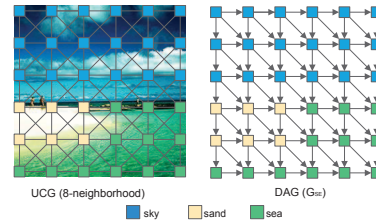


Figure 2: An 8-neighborhood UCG and one of its induced DAG in the southeastern (SE) direction.

adjacent units in image plane may not necessarily be neighbors in the chain. The graphical representations that respect the 2-D neighborhood system are more plausible solutions, and they are pervasively adopted in probabilistic graphical models (PGM). Therefore in this work, undirected cyclic graphs (UCG, an example is shown in Figure 2) are used to represent the connectivity structure of image units.

Due to the loopy structure of UCGs, they are unable to be unrolled to an acyclic processing sequence. Therefore, RNNs are not directly applicable to UCG-structured images. To address this issue, we approximate the topology of UCG by a combination of several directed acyclic graphs (DAGs), each of which is applicable for our proposed DAG-RNNs (one of the induced DAGs is depicted in Figure 2). Namely, an UCG-structured image is represented as the combination of a set of DAG-structured images. We now start introducing the detailed mechanism of our DAG-RNNs here, and later elaborate how they are applied to UCG-structured images in the next section.

We first assume that the topology of an image I is represented as a DAG $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_i\}_{i=1:N}$ is the vertex set and $\mathcal{E} = \{e_{ij}\}$ is the arc set (e_{ij} denotes an arc from v_i to v_j). The structure of the hidden layer \mathbf{h} follows the same topology as \mathcal{G} . Therefore, a forward propagation sequence can be generated by traversing \mathcal{G} , on the condition that one node should not be processed until all its predecessors are processed. The hidden layer $\mathbf{h}^{(v_i)}$ is represented as a nonlinear function over its local input $\mathbf{x}^{(v_i)}$ and the summarization of hidden representation of its predecessors. The local input $\mathbf{x}^{(v_i)}$ is obtained by aggregating (e.g. average pooling) from constituent elements in the feature tensor \mathbf{x} . In detail, the forward operation of DAG-RNNs is calculated by the following equations:

$$\begin{aligned} \hat{\mathbf{h}}^{(v_i)} &= \sum_{v_j \in \mathcal{P}_{\mathcal{G}}(v_i)} \mathbf{h}^{(v_j)} \\ \mathbf{h}^{(v_i)} &= f(U\mathbf{x}^{(v_i)} + W\hat{\mathbf{h}}^{(v_i)} + b) \\ \mathbf{o}^{(v_i)} &= g(V\mathbf{h}^{(v_i)} + c) \end{aligned} \quad (2)$$

where $\mathbf{x}^{(v_i)}, \mathbf{h}^{(v_i)}, \mathbf{o}^{(v_i)}$ are the representations of input, hidden and output layers located at v_i respectively, $\mathcal{P}_{\mathcal{G}}(v_i)$

is the direct predecessor set of vertex v_i in the graph \mathcal{G} , $\hat{\mathbf{h}}^{(v_i)}$ summarizes the information of all the predecessors of v_i . Note that the recurrent weight W in Equation 2 is shared across all predecessor vertexes in $\mathcal{P}_{\mathcal{G}}(v_i)$. We may learn a specific recurrent matrix W for each predecessor when vertexes (except source and sink vertex) if the DAG \mathcal{G} have a fixed number of predecessors. In this case, a finer-grained dependency may be captured.

The derivatives are computed in the backward pass, and each vertex is processed in the reverse order of forward propagation sequence. Specifically, to derive the gradients at v_i , we look at equations (besides Equation 2) that involve $\mathbf{h}^{(v_i)}$ in the forward pass:

$$\begin{aligned} \forall v_k \in \mathcal{S}_{\mathcal{G}}(v_i) \\ \mathbf{h}^{(v_k)} &= f(U\mathbf{x}^{(v_k)} + W\mathbf{h}^{(v_i)} + W\tilde{\mathbf{h}}^{(v_k)} + b) \\ \tilde{\mathbf{h}}^{(v_k)} &= \sum_{v_j \in \mathcal{P}_{\mathcal{G}}(v_k) - \{v_i\}} \mathbf{h}^{(v_j)} \end{aligned} \quad (3)$$

where $\mathcal{S}_{\mathcal{G}}(v_i)$ is the direct successor set for vertex v_i in the graph \mathcal{G} . It can be inferred from Equation 2, 3 that the errors backpropagated to the hidden layer ($d\mathbf{h}^{(v_i)}$) at v_i have two sources: direct errors from v_i ($\frac{\partial \mathbf{o}^{(v_i)}}{\partial \mathbf{h}^{(v_i)}}$), and summation over indirect errors propagated from its successors ($\sum_{v_k \in \mathcal{S}_{\mathcal{G}}(v_i)} \frac{\partial \mathbf{o}^{(v_k)}}{\partial \mathbf{h}^{(v_k)}} \frac{\partial \mathbf{h}^{(v_k)}}{\partial \mathbf{h}^{(v_i)}}$). The derivatives at v_i can then be computed by the following equations:

$$\begin{aligned} \Delta V^{(v_i)} &= g'(\mathbf{o}^{(v_i)})(\mathbf{h}^{(v_i)})^T \\ d\mathbf{h}^{(v_i)} &= V^T g'(\mathbf{o}^{(v_i)}) + \sum_{v_k \in \mathcal{S}_{\mathcal{G}}(v_i)} W^T d\mathbf{h}^{(v_k)} \circ f'(\mathbf{h}^{(v_k)}) \\ \Delta W^{(v_i)} &= \sum_{v_k \in \mathcal{S}_{\mathcal{G}}(v_i)} d\mathbf{h}^{(v_k)} \circ f'(\mathbf{h}^{(v_k)})(\mathbf{h}^{(v_i)})^T \\ \Delta U^{(v_i)} &= d\mathbf{h}^{(v_i)} \circ f'(\mathbf{h}^{(v_i)})(\mathbf{x}^{(v_i)})^T \end{aligned} \quad (4)$$

where \circ denotes the Hadamard product. By abuse of notation, $g'(\cdot) = \frac{\partial L}{\partial \mathbf{o}} \frac{\partial \mathbf{o}}{\partial g}$ is the derivative of loss function L with respect to the output function g , and $f'(\cdot) = \frac{\partial \mathbf{h}}{\partial f}$.² It is the second term of $d\mathbf{h}^{(v_i)}$ in Equation 4 that enables DAG-RNNs to propagate local information, which behaves similarly to the message passing [34] in probabilistic graphic models.

3.3. Decomposition

We decompose the UCG \mathcal{U} to a set of DAGs $\mathcal{G}^{\mathcal{U}} = \{\mathcal{G}_1, \dots, \mathcal{G}_d, \dots\}$. Hence, the UCG-structured image is represented as the combination of a set of DAG-structured images. Next, DAG-RNNs are applied independently to each DAG-structured image, and the corresponding hidden layer

²Here, g (and f) are composition functions. For instance, $\mathbf{o} = g(p(\mathbf{h}))$ where $p(\mathbf{h}) = V\mathbf{h} + c$. $\frac{\partial \mathbf{o}}{\partial g}$ actually refers to $\frac{\partial \mathbf{o}}{\partial p(\mathbf{h})}$.

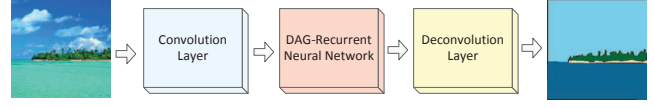


Figure 3: The architecture of the full labeling network, which consists of three functional layers: (1), convolution layer: it produces discriminative feature maps; (2), DAG-RNN: it models the contextual dependency among elements in the feature maps; (3), deconvolution layer: it upsamples the feature maps to output the desired sizes of label prediction maps.

\mathbf{h}_d is generated. The aggregation of the independent hidden layers yields the output layer \mathbf{o} . These operations can be mathematically expressed as follows:

$$\begin{aligned} \mathbf{h}_d^{(v_i)} &= f(U_d \mathbf{x}^{(v_i)} + \sum_{v_j \in \mathcal{P}_{\mathcal{G}_d}(v_i)} W_d \mathbf{h}_d^{(v_j)} + b_d) \\ \hat{\mathbf{o}}^{(v_i)} &= g(\sum_{\mathcal{G}_d \in \mathcal{G}^{\mathcal{U}}} V_d \mathbf{h}_d^{(v_i)} + c) \end{aligned} \quad (5)$$

where U_d, W_d, V_d and b_d are weight matrices and bias vector for the DAG \mathcal{G}_d , $\mathcal{P}_{\mathcal{G}_d}(v_i)$ is the direct predecessor set of vertex v_i in \mathcal{G}_d . This decomposition strategy is reminiscent of the tree-reweighted max-product algorithm (TRW) [31], which represents the problem on the loopy graphs as a convex combination of tree-structured problems.

We consider the following criterions for the decomposition. Topologically, the combination of DAGs should be equivalent to the UCG \mathcal{U} , so any two vertexes can be reachable. Besides, the combination of DAGs should allow the local information to be routed to anywhere in the image. In our experiment, we use the four context propagation directions (southeast, southwest, northwest and northeast) suggested by [9][23] to decompose the UCG. One example of the induced DAG of the 8-neighborhood UCG in the southeast direction is shown in Figure 2.

3.4. Full Labeling Network

The skeleton architecture of the full labeling network is illustrated in Figure 3. The network is end-to-end trainable, and it takes input as raw RGB images with any size. It outputs the label prediction maps with the same size of inputs. In more detail, the convolution layer is used to produce compact yet highly discriminative features for local regions. Next, the proposed DAG-RNN is used to model the semantic contextual dependencies of local representations. Finally, the deconvolution layer [16] is introduced to upsample the feature maps by learning a set of deconvolution filters.

To train the network, we adopt the average weighted cross entropy loss. It is formally written as:

$$\begin{aligned} L &= -\frac{1}{N} \sum_{v_i \in \mathcal{I}} \sum_{j=1}^c \mathbf{w}_j \mathbf{y}_j^{(v_i)} \log(\tilde{\mathbf{o}}_j^{(v_i)}) \\ \tilde{\mathbf{o}} &= \text{deconv}(\hat{\mathbf{o}}) \end{aligned} \quad (6)$$

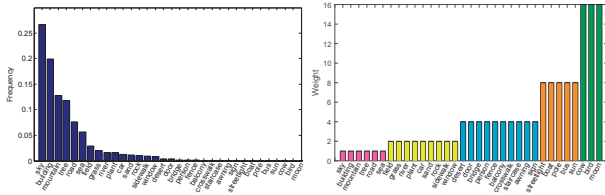


Figure 4: Graphical visualization of the class frequencies (left) and weights (right) on the siftFlow datasets [15]. The classes are sorted in the descending order based on their occurrence frequencies in training images.

where N is the number of image units in image I ; \mathbf{w} is the class weight vector, in which w_j stands for the weight for class j ; $\mathbf{y}^{(v_i)}$ is the binary label indicator vector for the image unit located in v_i ; $\tilde{\mathbf{o}}^{(v_i)}$ is the corresponding class likelihood vector, which is upsampled from $\hat{\mathbf{o}}^{(v_i)}$ by the deconvolution layer. The errors propagated from DAG-RNNs to the convolution layer for image unit v_i are calculated based on the following equations:

$$\Delta x^{(v_i)} = \sum_{\mathcal{G}_d \in \mathcal{G}^d} U_d^T d\mathbf{h}_d^{(v_i)} \circ f'(\mathbf{h}_d^{(v_i)}). \quad (7)$$

3.5. Attention to Rare Classes

In scene images, the class distribution is extremely imbalanced. Namely, very few classes account for large percentage of pixels in images. An example is demonstrated in Figure 4. It's therefore common to put more attention to rare classes, in order to boost their recognition precisions.

In the patch-based CNN training, Farabet et al. [8] and Shuai et al. [24] oversample the rare-class pixels to address this issue. (Different sampling methods are introduced and discussed in [24].) It's however inapplicable to adopt this strategy in our network training, which is a complex structure learning problem. Meanwhile, as the classes are distributed severely unequally in scene images, it's also problematic to weight classes according to their inverse frequencies. As an example, the frequency ratio between the most frequent (sky) and the most rare class (moon) on the SiftFlow dataset is 3.5×10^4 . If the above class weighting criterion is adopted like in [5][17], the frequent classes will be under-attended. Hence, we define the weighting function \mathbf{w} as follows:

$$\mathbf{w}_j = k^{\lceil \log_{10}(\eta/f_j) \rceil} \quad (8)$$

where $\lceil \cdot \rceil$ is the integer ceiling operator, f_j is the occurrence frequency of the class j , η denotes the threshold that discriminates the rare classes. Specifically, a class is identified as rare if its frequency is smaller than η , otherwise, it is a frequent class. k is a constant that controls the importance of rare classes ($k = 2$ in our experiments). The proposed weighting function has the following properties: (1), it attends to rare classes by assigning them higher weights; (2), the degree of attention for rare classes grows exponentially

based on their ratio magnitudes w.r.t the threshold η ; We use the following criterion to determine the value of η : the accumulated frequency of all the non-rare classes is 85%. We call it 85%-15% rule, and [33] uses a similar rule.

4. Experiments

We justify our method on three popular and challenging real-world scene image labeling benchmarks: SiftFlow [15], CamVid [2] and Barcelona [28]. Two types of scores are reported: the percentage of all correctly classified pixels (**Global**), and average per-class accuracy (**Class**).

4.1. Baselines

The convolution neural network (CNN), which jointly learns features and classifiers is used as our first baseline. In this case, the parameters are optimized to maximize the independent prediction accuracy for local patches. Another baseline is the network that shares the same architecture with our DAG-RNNs, while removes the recurrent connections. Mathematically, the W_d and b_d in Equation 5 are fixed to $\mathbf{0}$. In this case, the DAG-recurrent neural network degenerates to an ensemble of four plain two-layer neural networks (CNN-ENN). The performance disparity between the baselines and DAG-RNNs clearly illuminates the efficacy of our dependency modeling method.

4.2. Implementation Details

We use the following two networks to be the convolution layers in our experiments:

- **CNN-65**: The network consists of five convolutional layers, the kernel sizes of which are $8 \times 8 \times 3 \times 64$, $6 \times 6 \times 64 \times 128$, $5 \times 5 \times 128 \times 256$, $4 \times 4 \times 256 \times 256$ and $1 \times 1 \times 256 \times 64$ respectively. Each of the first three convolutional layers are followed by a ReLU and non-overlapping 2×2 max pooling layer. The parameters of this network is learned from image patches (65×65) of the target dataset only (**Setting 1**).
- **VGG-conv5**: The network borrows its architecture and parameters from VGG-verydeep-16 net [25]. In detail, we discard all the layers after the 5th pooling layer to yield the desired convolution layer. The network is pre-trained on ImageNet dataset and fine-tuned on the target dataset. [6] (**Setting 2**).

In DAG-RNNs, the adopted non-linear functions (refer to Equation 2) are ReLU [13] for hidden neurons: $f(x) = \max(0, x)$ and *softmax* for output layer g . The dimensionality of hidden layer \mathbf{h} is empirically set to 64 for CNN-65 and 128 for VGG-conv5 respectively.³ In our experi-

³Based on our preliminary results, we didn't observe too much performance improvement by using larger \mathbf{h} (e.g. 128 in CNN-65, and 256 in VGG-conv5) on the siftFlow dataset. In addition, the networks with larger capacity incur much heavier computation burdens.

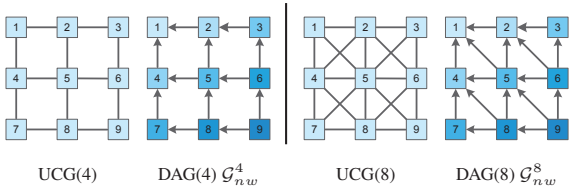


Figure 5: Two UCGs (with 4, 8 neighborhood system) and their induced DAGs in the northwestern (NW) direction.

ments, we consider two UCGs with 4 and 8 neighborhood systems. Their induced DAGs in the northwestern direction are shown in Figure 5. In comparison with DAG(4), DAG(8) enables information to be propagated in shorter paths, which is critical to prevent the long-range information from vanishing. As exemplified in Figure 5, the length of propagation path from v_9 to v_1 in \mathcal{G}_{nw}^8 is halved to that in \mathcal{G}_{nw}^4 ($4 \rightarrow 2$ steps). In practice, we apply the function g after the deconvolution layer. The upsampling factor for the deconvolution layer is set to be 2 and 4 for CNN-65 and VGG-conv5 respectively. Namely, the ground truth maps are subsampled during training, while the label prediction maps are further upsampled by bilinear operation in the testing (validation) phase.

The full labeling network is trained by stochastic gradient descent with momentum. The parameters are updated after one image finishes its forward and backward passes. The learning rate is initialized to be 10^{-3} , and decays exponentially with the rate of 0.9 after 10 epoch. The reported results are based on the model trained in 35 epoches. We tune the parameters and diagnoses the network performance based on CNN-65. We also include the results of VGG-conv5 to see whether our proposed DAG-RNNs are beneficial for the highly discriminative representation from the state-of-the-art VGG-verydeep-16 net [25].

4.3. SiftFlow Dataset

The SiftFlow dataset has 2688 images generally captured from 8 typical outdoor scenes. Every image has 256×256 pixels, which belong to one of the 33 semantic classes. We adopt the training/testing split protocol (2488/200 images) provided by [15] to perform our experiments. Following the 85%-15% criterion, the class frequency threshold $\eta = 0.05$. Statistically, out of 33 classes, 27 of them are regarded as infrequent class. The graphical visualization of the weights for different classes are depicted in Figure 4.

The quantitative results are listed in Table 1, where the upper part presents the performance of methods under setting 1. Our baseline CNN-65 achieves very promising results, which proves the effectiveness of the convolution layer. We also notice that results of CNN-65 fall behind CNN-65-ENN on the average class accuracy. This phenomenon is also observed on the CamVid and Barcelona benchmarks, as shown by Table 2 and 3 respectively. This

Methods	Global	Class
Byeon <i>et al.</i> [4]	70.1%	22.6%
Liu <i>et al.</i> [15]	74.8%	N/A
Farabet <i>et al.</i> [8]	78.5%	29.4%
Pinheiro <i>et al.</i> [19]	77.7%	29.8%
Tighe <i>et al.</i> [29]	79.2%	39.2%
Sharma <i>et al.</i> [20]	79.6%	33.6%
Shuai <i>et al.</i> [22]	80.1%	39.7%
Yang <i>et al.</i> [33]	79.8%	48.7%
CNN-65	76.1%	32.5%
CNN-65-ENN	76.1%	37.0%
CNN-65-DAG-RNN(4)	80.5%	42.6%
CNN-65-DAG-RNN(8)	81.1%	48.2%
Long <i>et al.</i> [16]	85.2%	51.7%
Caesar <i>et al.</i> [5]	N/A	55.6%
VGG-conv5-ENN	84.0%	48.8%
VGG-conv5-DAG-RNN(8)	85.3%	55.7%

Table 1: Quantitative performance of our method on the siftFlow dataset. The numbers (in brackets) following the DAG-RNN denote the neighborhood system of the UCG.

result indicates that the proposed class weighting function significantly boosts the recognition accuracy for rare classes. By adding DAG-RNN(8), our network reaches 81.1% (48.2%) on global (class) accuracy⁴, which outperforms the baseline (CNN-65-ENN) by 5% (11.2%). Meanwhile, we observe promising accuracy gain (global: 0.6% / class: 5.6%) by switching DAG-RNN(4) to DAG-RNN(8), in which we believe that long-range dependencies are better captured as information propagation paths in DAG(8) are shorter than those in DAG(4). Such performance benefits can be observed consistently on the CamVid (0.5% / 2.0%) and Barcelona (1.1% / 1.6%) datasets, as evidenced in Table 2 and 3 respectively. Moreover, in comparison with other representation learning nets, which are fed with much richer contextual input (133×133 patch in [19], 3-scale 46×46 patches in [8]), our DAG-RNNs outperform theirs by a large margin. Importantly, our results match the state-of-the-art under this setting.

Furthermore, we initialize our convolution layers with VGG-verydeep-16 [25], which has been proven to be the state-of-the-art feature extractor. The quantitative results under setting 2 are listed in the lower body of Table 1. Our baseline VGG-conv5-ENN (pre-trained on ImageNet [6]) surpasses the best performance of methods under setting 1. This result indicates the significance of large-scale data in deep neural network training. Interestingly, our DAG-RNN(8) is still able to further improve the discriminative power of local features by modeling their dependencies, thereby leading to a phenomenal (6.9%) average class accuracy boost. Note that Fully Convolution Networks (FCNs) [16] uses activations (feature maps) from multiple convolu-

⁴If we disassemble the full labeling network to two disjoint parts - CNN-65 and DAG-RNN(8), and they are optimized independently, the corresponding accuracies are 80.1% and 42.7%. The performance discrepancy indicates the importance of the joint optimization for the full network.

Methods	Global	Class
Tighe <i>et al.</i> [28]	78.6%	43.8%
Sturgess <i>et al.</i> [26]	83.8%	59.2%
Zhang <i>et al.</i> [35]	82.1%	55.4%
Bulò <i>et al.</i> [3]	82.1%	56.1%
Ladický <i>et al.</i> [14]	83.8%	62.5%
Tighe <i>et al.</i> [29]	83.9%	62.5%
CNN-65	84.3%	53.2%
CNN-65-ENN	84.1%	58.1%
CNN-65-DAG-RNN(4)	88.2%	66.3%
CNN-65-DAG-RNN(8)	88.7%	68.3%
VGG-conv5-ENN	91.0%	76.5%
VGG-conv5-DAG-RNN (8)	91.6%	78.1%

Table 2: Quantitative performance of our method on the CamVid dataset.

tion layers, whereas our VGG-conv5-ENN only use feature maps from conv5 layer. Hence, there is a slight performance gap between our VGG-conv5-ENN and FCNs. Nonetheless, our VGG-conv5-DAG-RNN(8) still performs comparably with FCNs on global accuracy, and significantly outperforms it on the class accuracy. Importantly, our full labeling network also achieves new state-of-the-art performance under this setting. The detailed per-class accuracy is listed in Table 4.

4.4. CamVid Dataset

The CamVid dataset [2] contains 701 high-resolution images (960×720 pixels) from 4 driving videos at daytime and dusk (3 daytime and 1 dusk video sequence). Images are densely labelled with 32 semantic classes. We follow the usual split protocol [26][29] (468/233) to obtain training/testing images. Similar to other works [2][3][26][29], we only report results on the most common 11 categories. According to the 85%-15% rule, 4 classes are identified as rare, and η is 0.1.

The quantitative results are given in Table 2. Our baseline networks (CNN-65, CNN-65-ENN) achieve very competitive results. By explicitly modeling contextual dependencies among image units, our CNN-65-DAG-RNN(8) brings phenomenal performance benefit (4.6% and 10.2% for the global and class accuracy respectively). Moreover, in comparison with state-of-the-art methods [3][14][26][29], our CNN-65-DAG-RNN(8) outperforms theirs by a large margin (4.8% / 5.8%), demonstrating the profitability of adopting high-level features learned from CNN and context modeling with our DAG-RNNs. Furthermore, the VGG-conv5-ENN alone performs excellently. Even though the performance starts saturating, our DAG-RNN(8) is able to consistently improve the labeling results.

4.5. Barcelona Dataset

The barcelona dataset [28] consists of 14871 training and 279 testing images. The size of the images varies across different instances, and each pixel is labelled as one of the 170

Methods	Global	Class
Tighe <i>et al.</i> [28]	66.9%	7.6%
Farabet <i>et al.</i> [8]	46.4%	12.5%
Farabet <i>et al.</i> [8]	67.8%	9.5%
CNN-65	69.0%	10.5%
CNN-65-ENN	69.0%	11.0%
CNN-65-DAG-RNN(4)	71.3%	12.9%
CNN-65-DAG-RNN(8)	72.4%	14.5%
VGG-conv5-ENN	73.3%	21.1%
VGG-conv5-DAG-RNN(8)	74.6%	24.6%

Table 3: Quantitative performance of our method on the Barcelona dataset.

semantic classes. The training images range from indoor to outdoor scenes, whereas the testing images are only captured from the barcelona street scene. These issues pose Barcelona as a very challenging dataset. Based on the 85%-15% rule, 147 classes are identified as rare classes, and the class frequency threshold η is 0.005.

Table 3 presents the quantitative results. From which, we clearly observe that our baseline networks (CNN-65 and CNN-65-ENN) achieve very competitive results, which has already matched the state-of-the-art results. The introduction of DAG-RNN(8) leads to promising performance improvement, therefore the full labeling network clinches the new state-of-the-art under setting 1. More importantly, under setting 2, even though the VGG-conv5-ENN is extraordinarily competitive, the DAG-RNN(8) is still able to enhance its labeling performance significantly.

4.6. Effects of DAG-RNNs to Per-class Accuracy

In this section, we investigate the effects of our DAG-RNNs for each class. The detailed per-class accuracy for the SiftFlow dataset is listed in Table 4. Under setting 1, we find that the contextual information encoded through our DAG-RNN(8) is beneficial for almost all classes. In this case, the local representations from CNN-65 are not strong, so their discriminative power can be greatly enhanced by modeling their dependencies. In line with it, we observe remarkable performance benefit (+11.2%) for almost all classes. Under setting 2, the VGG-conv5 net is pre-trained on the ImageNet dataset [6], and it recognizes most classes excellently. Notwithstanding the local representations are highly discriminative in this situation, our DAG-RNN(8) further tremendously improves their representative power for rare classes. Statistically, we observe a phenomenal 8.6% accuracy gain for rare classes. Under both settings, modeling the dependencies among local features enables the classification to be contextual aware. Therefore, the local ambiguities are mitigated to a large extent. However, we fail to observe commensurate accuracy improvements for extremely-small-size and rare 'object' classes (e.g. bird and bus), we conjecture that the weak local information may have been overwhelmed by context (e.g. a small bird is swallowed by the broad sky in Figure 1).

	sky	building	tree	mountain	road	sea	field	car	sand	river	plant	grass	window	sidewalk	rock	bridge	door	fence	person	staircase	awning	sign	boat	crosswalk	pole	bus	balcony	streetlight	sun	bird	global	class
Frequency	27.1	20.2	12.6	12.4	6.93	5.61	3.64	1.64	1.41	1.37	1.33	1.22	1.07	0.89	0.85	0.36	0.26	0.24	0.23	0.18	0.11	0.11	0.06	0.05	0.04	0.03	0.03	0.02	0.01	0.004	-	-
CNN-65-ENN	94.1	84.9	77.6	74.2	80.9	61.1	30.7	71.0	27.7	34.7	21.8	63.5	27.8	47.0	21.0	8.8	35.7	33.7	29.3	6.8	0.37	16.3	1.4	48.1	0	0.43	34.5	6.0	71.4	0	76.1	37.0
CNN-65-DAG-RNN(8)	95.9	87.3	82.5	77.9	85.8	70.2	43.5	80.1	52.9	65.4	37.2	57.8	40.4	59.1	27.6	31.4	51.8	38.0	35.6	50.9	7.8	31.0	5.14	82.8	0	0	54.9	8.59	85.5	0	81.1	48.2
VGG-conv5-ENN	96.0	91.1	84.4	82.9	90.4	83.5	48.8	77.5	63.5	57.6	32.6	60.2	34.9	66.0	25.8	20.0	51.9	44.0	38.6	45.9	26.5	33.7	14.9	50.2	1.1	0	32.7	9.1	99.9	0	84.0	48.8
VGG-conv5-DAG-RNN(8)	96.3	90.8	82.1	85.1	89.2	84.8	55.4	84.2	67.9	75.3	51.5	64.8	45.2	63.5	45.7	37.3	56.8	44.7	36.2	58.7	18.3	40.0	63.3	65.2	18.4	1.4	45.8	5.4	97.9	0	85.3	55.7

Table 4: Per-class accuracy comparison on the SiftFlow dataset. All the numbers are displayed in the percentage scale. The statistics for class frequency is obtained in test images. For reading convenience, the frequent and rare classes are placed in the same block.

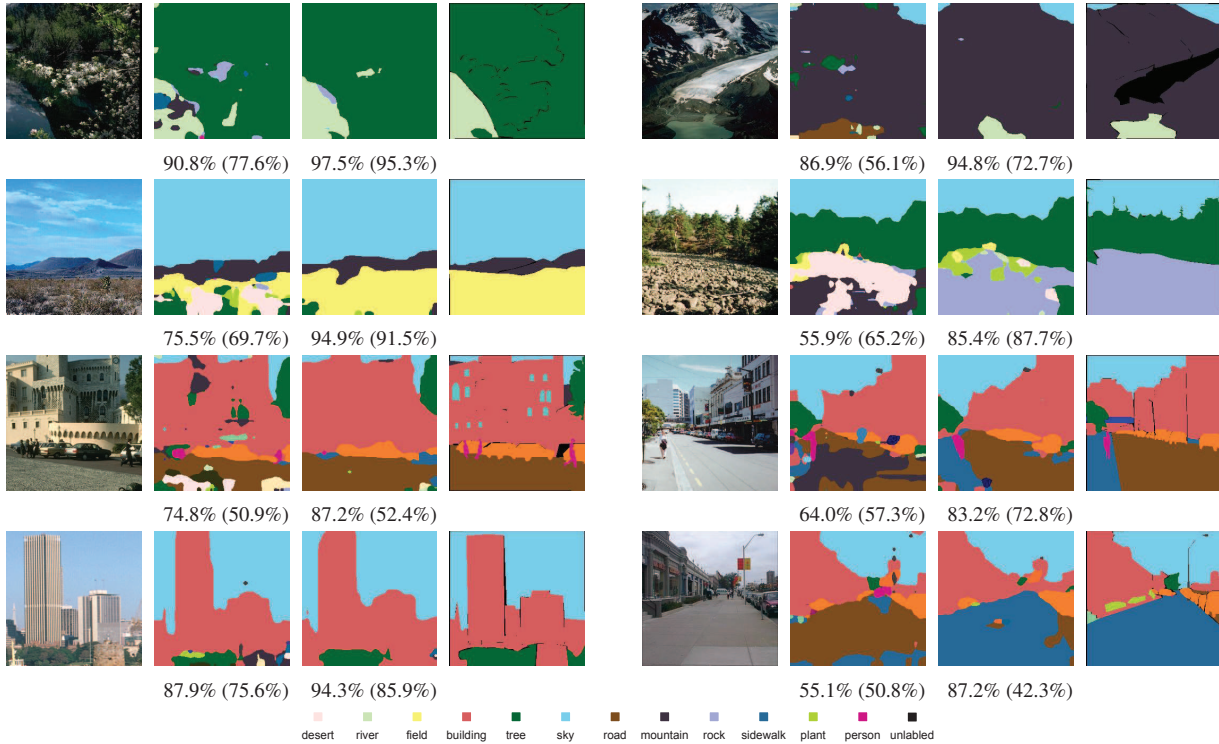


Figure 6: (Best viewed in color) Qualitative labeling results. We show input images, local prediction maps (CNN-65-ENN), contextual labeling maps (CNN-65-DAG-RNN(8)) and their ground truth respectively. The numbers outside and inside the parentheses are global and class accuracy respectively.

4.7. Discussion of Modeled Dependency

We show a number of qualitative labeling results in Figure 6. By looking into them, we can have some interesting observations. The DAG-RNNs are capable of (1), enforcing local consistency: neighborhood pixels are likely to be assigned to the same labels. In Figure 6, the left-panel examples show that confusing regions are smoothed by using our DAG-RNNs. (2), ensuring semantic coherence: the pixels that are spatially far away are usually given labels that could co-occur in a meaningful scene. For example, the ‘desert’ and ‘mountain’ classes are usually not seen together with ‘trees’ in a ‘open country’ scene, so they are corrected to ‘stone’ in the second example of the right panel. More examples of this kind are shown in the right panel. These results illuminate that useful contextual dependencies may have been captured by our DAG-RNNs.

5. Conclusion

In this paper, we propose DAG-RNNs to model the contextual dependencies of local features. DAG-RNNs are able to encode the global gist of images to local features, so their discriminative power can be enhanced. Furthermore, the proposed class weighting function is experimentally proved to be effective towards the recognition enhancement for rare classes. Integrating with the convolution and deconvolution layers, our DAG-RNNs achieve state-of-the-art results on three challenging scene labeling benchmarks. We also demonstrate that useful contexts are captured by our DAG-RNNs, which is helpful for generating smooth and semantically sensible labeling maps in practice.

Acknowledgements: The research is supported by Singapore Ministry of Education (MOE) Tier 2 ARC28/14, and Singapore A*STAR Science and Engineering Research Council PSF1321202099. The authors would also like to thank NVIDIA for their generous donation of GPU.

References

- [1] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu. Advances in optimizing recurrent networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8624–8628. IEEE, 2013.
- [2] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV (1)*, pages 44–57, 2008.
- [3] S. R. Bulò and P. Kotschieder. Neural decision forests for semantic image labelling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 81–88. IEEE, 2014.
- [4] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with lstm recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3547–3555, 2015.
- [5] H. Caesar, J. Uijlings, and V. Ferrari. Joint calibration for semantic segmentation. In *BMVC*, 2015.
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [7] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.
- [8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1915–1929, 2013.
- [9] A. Graves. Offline arabic handwriting recognition with multidimensional recurrent neural networks. In *Guide to OCR for Arabic Scripts*, pages 297–313. Springer, 2012.
- [10] A. Graves, A.-R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE, 2013.
- [11] O. Irsoy and C. Cardie. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 720–728, 2014.
- [12] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [14] L. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and crfs. In *Computer Vision–ECCV 2010*, pages 424–437. Springer, 2010.
- [15] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1972–1979. IEEE, 2009.
- [16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.
- [17] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feed-forward semantic segmentation with zoom-out features. In *Computer Vision and Pattern Recognition, 2015. CVPR 2015. IEEE Conference on*. IEEE, 2015.
- [18] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*, 2012.
- [19] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *Proceedings of The 31st International Conference on Machine Learning*, pages 82–90, 2014.
- [20] A. Sharma, O. Tuzel, and M.-Y. Liu. Recursive context propagation network for semantic scene labeling. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2014.
- [21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Computer Vision–ECCV 2006*, pages 1–15. Springer, 2006.
- [22] B. Shuai, G. Wang, Z. Zuo, B. Wang, and L. Zhao. Integrating parametric and non-parametric models for scene labeling. In *Computer Vision and Pattern Recognition, 2015. CVPR 2015. IEEE Conference on*. IEEE, 2015.
- [23] B. Shuai, Z. Zuo, and W. Gang. Quaddirectional 2d-recurrent neural networks for image labeling. *Signal Processing Letters, IEEE*, 2015.
- [24] B. Shuai, Z. Zuo, G. Wang, and B. Wang. Scene parsing with integration of parametric and non-parametric models. *Image Processing, IEEE Transactions on*, 2016.
- [25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [26] P. Sturgess, K. Alahari, L. Ladický, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *Computer Vision–BMVC 2009*. 2009.
- [27] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.
- [28] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *Computer Vision–ECCV 2010*, pages 352–365. Springer, 2010.
- [29] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3001–3008. IEEE, 2013.
- [30] Z. Tu. Auto-context and its application to high-level vision tasks. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
- [31] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. *Information Theory, IEEE Transactions on*, 51(7):2313–2335, 2005.
- [32] P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.
- [33] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *Computer*

Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 3294–3301. IEEE, 2014.

- [34] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.
- [35] C. Zhang, L. Wang, and R. Yang. Semantic segmentation of urban scenes using dense depth maps. In *Computer Vision–ECCV 2010*, pages 708–721. Springer, 2010.
- [36] 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. *arXiv preprint arXiv:1502.03240*, 2015.
- [37] X. Zhu, P. Sobhani, and H. Guo. Long short-term memory over tree structures. *arXiv preprint arXiv:1503.04881*, 2015.
- [38] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen. Convolutional recurrent neural networks: Learning spatial dependencies for image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–26, 2015.