Abstract: Scene parsing refers to parsing the scene image into a set of coherent semantic regions. It has been a long-standing and challenging topic in computer vision as well as artificial intelligence. In the thesis, we explore how to develop effective as well as efficient deep neural networks to perform scene parsing.

In our first chapter, we adopt Convolutional Neural Networks (CNNs) to be our parametric model to learn discriminative features and classifiers for local patch classification. Based on the occurrence frequency distribution of classes, an ensemble of CNNs (CNN-Ensemble) are learned, in which each CNN component focuses on learning different and complementary visual patterns. The local beliefs of pixels are output by CNN-Ensemble. Considering that visually similar pixels are indistinguishable under local context, we leverage the global scene semantics to alleviate the local ambiguity. The global scene constraint is mathematically achieved by adding a global energy term to the labeling energy function, and it is practically estimated in a non-parametric framework. A large margin based CNN metric learning method is also proposed for better global belief estimation. In the end, the integration of local and global beliefs gives rise to the class likelihood of pixels, based on which maximum marginal inference is performed to generate the label prediction maps. Even without any post-processing, we achieve very promising results on the challenging SiftFlow and Barcelona benchmarks.

We observe the limitation of CNNs that are trained from scratch. Then, we adapt the pre-trained CNN (e.g. VGG-16) to extract high-level features. In our second chapter, we discuss how to effectively capture the rich contextual dependencies over image regions. Specifically, we propose Directed Acyclic Graph - Recurrent Neural Networks (DAG-RNN) to perform context aggregation over locally connected feature maps. More specifically, DAG-RNN is placed on top of pre-trained CNN (feature extractor) to embed context into local features so that their representative capability can be enhanced. In comparison with plain CNN (as in Fully Convolution Networks - FCN), DAG-RNN is empirically found to be significantly more effective at aggregating context. Therefore, DAG-RNN demonstrates noticeably performance superiority over FCNs on scene segmentation. Besides, DAG-RNN entails dramatically less parameters as well as demands fewer computation operations, which makes DAG-RNN more favorable to be potentially applied on resource-constrained embedded devices. Meanwhile, the class occurrence frequencies are extremely imbalanced in scene segmentation, so we propose a novel class-weighted loss to train the segmentation network. The loss distributes reasonably higher attention weights to infrequent classes during network training, which is essential to

boost their parsing performance. We evaluate our segmentation network on three challenging public scene segmentation benchmarks: Sift Flow, Pascal Context and COCO Stuff. On top of them, we achieve very impressive segmentation performance.

Considering that scene segmentation demands multi-level visual recognition ranging from low-level (e.g. boundary detection) to high-level (e.g. general object recognition). In our third chapter, we first discuss and compare two widely used adaptation approaches of pre-trained CNN to retain lower-level features - ``dilation" and ``skip". By slightly modifying the parametrization of skip layers, we demonstrate that segmentation network with our skip layers delivers a very promising network architecture. Furthermore, we propose and place a convolutional context network (CCN) on top of pre-trained CNNs, which is used to aggregate contexts for high-level feature maps so that their representative capability can be enhanced. In order to retain as much detailed low-level information as possible from pre-trained CNN, we introduce ``dense skip" network architecture. We name our segmentation network improved fully convolutional network (IFCN) based on its significantly enhanced structure over FCN. We carry out careful ablation studies to justify each contribution individually. Without bells and whistles, IFCN achieves state-of-the-arts on ADE20K, Pascal Context and Pascal VOC 2012 segmentation datasets.