Ph.D. DISSERTATION

Learning Context-Aware Representations for Semantic Segmentation

의미론적 영상 분할을 위한 맥락 인식 기반 표현 학습

BY

MYEONG HEESOO

FEBRUARY 2017

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Learning Context-Aware Representations for Semantic Segmentation

의미론적 영상 분할을 위한 맥락 인식 기반 표현 학습

BY

MYEONG HEESOO

FEBRUARY 2017

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Learning Context-Aware Representations for Semantic Segmentation

의미론적 영상 분할을 위한 맥락 인식 기반 표현 학습

지도교수 이 경 무
이 논문을 공학박사 학위논문으로 제출함

2017년 2월

서울대학교 대학원

전기 컴퓨터 공학부

명 희 수

명희수의 공학박사 학위 논문을 인준함

2017년 2월

위 원 장: _____
부위원장: _____
위 원: _____
위 원: _____
위 원: _____

Abstract

Semantic segmentation, segmenting all the objects and identifying their categories, is a fundamental and important problem in computer vision. Traditional approaches to semantic segmentation are based on two main elements: visual appearance features and semantic context. Visual appearance features such as color, edge, shape and so on, are a primary source of information for reasoning the objects in an image. However, image data are sometimes unable to fully capture diversity in the object classes, since the appearance of the objects presented in real world scenes is affected by imaging conditions such as illumination, texture, occlusion, and viewpoint. Therefore, semantic context, obtained from not only the presence but also the location of other objects, can help to disambiguate the visual appearance in semantic segmentation tasks. The modern contextualized semantic segmentation systems have successfully improved segmentation performance by refining inconsistently labeled pixels via modeling of contextual interactions. However, they considered semantic context and visual appearance features independently due to the absence of the suitable representation model. Motivated by this issue, this dissertation proposes a novel framework for learning semantic context-aware representations in which appearance features is enhanced and enriched by semantic context and vice versa.

The first part of the dissertation will be devoted to semantic context-aware appearance modeling for semantic segmentation. Adaptive context aggregation network is studied to capture semantic context adequately while multiple steps of reasoning. Secondly, semantic context will be reinforced by utilizing visual appearance. Graph and example-based context model is presented for estimating contextual relationships

according to the visual appearance of objects. Finally, we propose a Multiscale Conditional Random Fields (CRFs), for integrating context-aware appearance and appearance-aware semantic context to produce accurate segmentations. Experimental evaluations show the effectiveness of the proposed context-aware representations on various challenging datasets.

keywords: Computer Vision, Object Recognition, Semantic Segmentation, Context

student number: 2009-20799

Contents

List of Tables

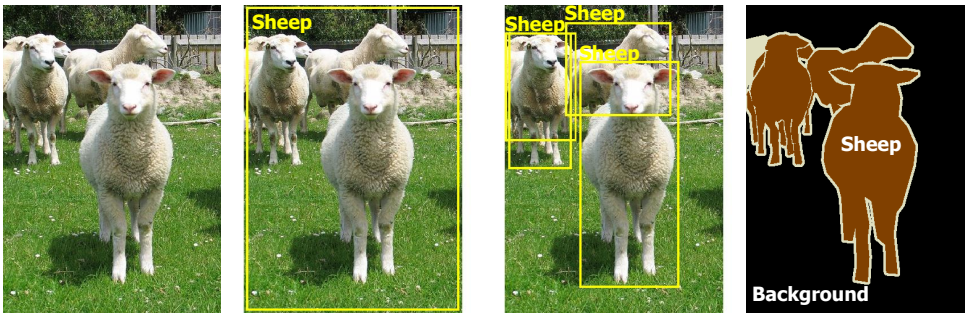
List of Figures

Chapter 1

Introduction

A human can recognize and localize the objects instantaneously with their eyes and brains. Computer vision is a research field that deals with how to give a similar ability to a machine or computer. Therefore, object recognition problem, identifying all object instances with their categories, have been a major research topic in computer vision. Robustly finding various objects from images or videos will have many potential applications, for instance, automatic inspection, surveillance system, driverless car, and medical diagnosis.

Many different types of object recognition tasks according to the level of understanding has received much attention and studied extensively over recent decades. Most notably, in image classification task, the goal is to classify a picture, for example, *sheep* as illustrated in Figure 1.1, while object detection additionally finds the location of the objects. Semantic segmentation labels every pixel in the image with the corresponding object categories. With the help of recent deep neural networks [5] and internet-based, large-scale dataset such as ImageNet [6], He *et al*. [7] reported that its image classification system had surpassed the human-level performance. How-



(a) Input image (b) Image classification (c) Object detection (d) Semantic segmentation

Figure 1.1: The three most popular object recognition tasks. (a) The given input image, (b) image classification classify the image as *sheep*, (c) object detection specify the bounding boxes of *sheep*, and (d) semantic segmentation extracts the spatial extent of *sheep*.

ever, the accuracy of many object recognition systems is still behind the human-level intelligence despite several decades of research in this area.

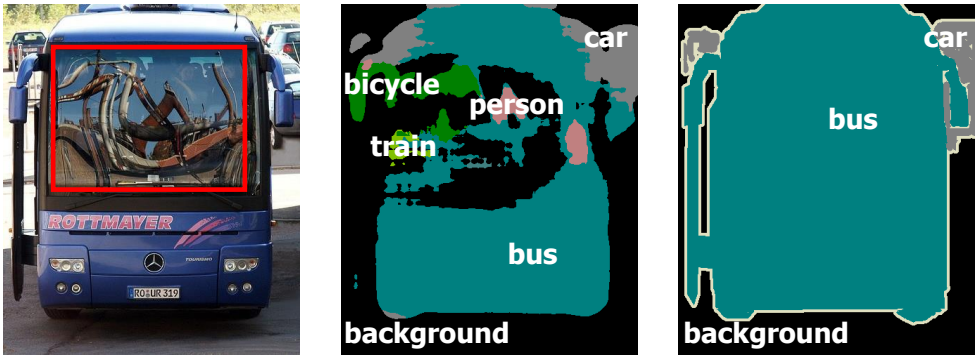
In this dissertation, we explore semantic segmentation problem in greater detail. Semantic segmentation has many advantages over the other object recognition tasks such as object detection and image classification. Image classification cannot deal with scenes composed of multiple objects, and, in object detection task, bounding box representations of an object typically contain many useless background pixels. The output of semantic segmentation provides better spatial support of objects, thereby synthesizing output of image classification and object detection in a straightforward manner. One drawback of semantic segmentation is relative higher computational cost than other object recognition tasks due to pixel-level prediction. Nonetheless, with the recent development of hardware and algorithms, there is a vast improvement in computational performance of semantic segmentation.

1.1 Backgrounds

The individual tasks in object recognition problem can be formulated as a labeling problem in which the object class labels has to be assigned to regions or images. However, an object class can have many different visual representations depends on imaging conditions such as illumination, texture, occlusion, and viewpoint. What makes this problem more difficult is that a few class may rely on functional features, for example, *chair* class. The way to overcome this challenges and make general object recognition systems is to capture robust and discriminative representations of each category.

The geometric models of objects played a crucial role in early efforts on object recognition [8]. The advantages of geometric object descriptions are invariance to viewpoint achieved by projecting 3D objects onto 2D planes under perspective projection and invariance to illumination using edge information. Nevertheless, reliable extraction of geometric information (*e.g*. lines, circles, *etc*.) is not straightforward in an unrestricted environment.

Recently, most attempts on object recognition have been focused on appearance-based approaches using advanced feature descriptors and machine learning algorithm. To tell a few representative feature descriptors, Scale-invariant feature transform (SIFT) for keypoint detection and object matching [9], the eigenface for face recognition [10], and Histogram of oriented gradients (HoG) for human detection [11] have been proposed and attracted considerable attention. Based on several feature descriptors, discriminative classifiers such as k-nearest neighbors, support vector machine, decision tree, and neural networks have been applied to various object recognition problem. More recently, deep neural networks [5] shows efficient extractions of discriminative features for object recognition problems.



(a) Input image (b) FCN ([12]) (c) Ground truth

Figure 1.2: The local appearance ambiguity marked with the red box in (a) causes difficulties in recognizing objects in (b).

1.2 Context Modeling for Semantic Segmentation Systems

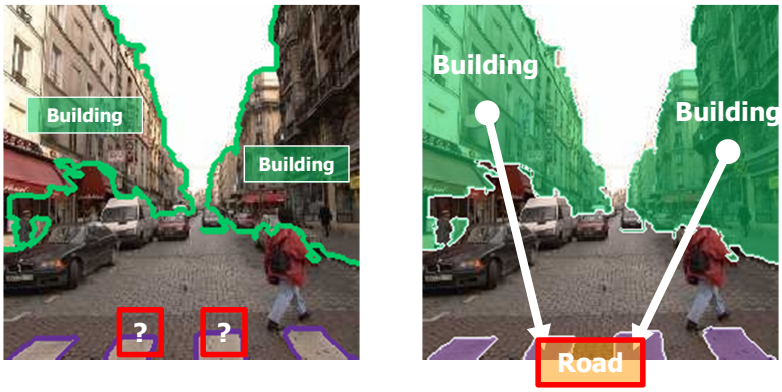
Although appearance-based approaches for semantic segmentation have shown its great performance for challenging datasets such as PASCAL VOC [12] and Microsoft COCO [13], still there are many unresolved issues. One of the most important topics is inherent ambiguity in visual information which makes it difficult to identify objects, as depicted in Figure 1.2. To this end, contextual interactions have been employed for particular recognition tasks in which simultaneously recognize multiple objects such as object detection and semantic segmentation. Many recent approaches firstly obtain object class score maps for each pixel with visual appearance. Then they employ a probabilistic graphical model with the retrieved score maps where semantic context information is represented as clique potentials to infer the final segmentation.

However, the learning process of the conventional appearance-based models does not consider the semantic context information. At the stage of calculating object scores with the appearance information, visual ambiguity may produce totally unreasonable

object likelihoods. However, it is difficult to fix in the probabilistic graphical models because of the incorrect initial object scores. This problem comes from the disentangled learning process of appearance and semantic context and limits the overall systems performance. In Figure 1.2, Long *et al*. [14] performs poorly on the windows of the bus. Semantic context information cannot resolve it because the large regions of the image are wrongly labeled.

Researchers in computer vision has widely used the term "context" in object recognition literature, but there is little consensus about what constitutes context. Very basic types of context representing local windows around pixels or regions are widely used definition. However, since most of the feature extraction methods use some degree of context windows, it is difficult to say that the all the appearance-based approaches is the contextual approaches. In this dissertation, we limit the definition of context to only the "semantic context", for example, object class co-occurrence and relative positions of objects, object-scene interaction, and things-stuff relations. Recently, context-aware semantic segmentation systems have gained wide attention and use a graphical model where semantic context are modeled as high-order potentials. In this formulation, the semantic context is regarded as *a priori* in understanding an image.

However, the problem of semantic context such as object co-occurrence is that they do not take into account the visual information. For instance, object co-occurrence matrix only considers the occurrence of the two objects in the scenes. Therefore, as in Figure 1.3, semantic context information enforces frequently appeared object class instead of the correct *crosswalk*. However, while growing the number of labels and different kinds of semantic contextual information such as *building* nearby *car*, *pedestrian* on *crosswalk* and *sky* over *sea*, extracting all different context scenarios against all object classes considering the visual appearance cannot be addressed due to complexity.



Input image Conventional context model

Figure 1.3: One of the problems of semantic context modeling. The ground truth object of the unknown regions is *crosswalk*. Based on the co-occurrence statistics, the *building* regions enforce the unknown regions to be *road* due to the strong correlation of *building* and *road*.

1.3 Dissertation Goal and Contribution

The goal of the dissertation is to learn context-aware representations for proper semantic segmentations which apply to scene understanding and to study the combination of the context-aware appearances and the appearance-aware context. The contribution of the dissertations is summarized as follows:

1. Learning context-aware representations for robustness in the semantic context, as well as the visual appearances, is proposed. The semantic context can help to disambiguate the appearance features, and the visual information can contribute to utilizing the semantic context properly.

2. Adaptive context aggregation network where the semantic context are adaptively captured while multistep reasoning is presented. In this framework, appearance

modeling while considering semantic context information is achieved.

3. Graph and example-based context model in which appearance enhances semantic context is presented. With this model, semantic context modeling while considering appearance information is produced.

1.4 Organization of Dissertation

The main part of the dissertation consists of four chapters. The first chapter deals with the methods for the context-aware appearance modeling with Deep Convolutional Neural Networks. Chapter 3 and 4 present the appearance-aware context model with second-order and high-order potentials, respectively. Chapter 5 proposes the unified framework to integrating context-aware appearance model and appearance-aware context. The remainder of this dissertation is structured as follows.

In Chapter 2, we present an Adaptive Context Aggregation Network (ACAN) architectures for semantic segmentation. We present a method that learns context-aware representations of the visual object. These latent representations encode contextual information in a continuous vector space to help overcome major challenges in this task, including local appearance ambiguity. The proposed network automatically searches for the semantic context in an image that is related to the accurate semantic segmentation. Our frameworks contain multiple steps of reasoning of semantic segmentation. Experiments conducted on PASCAL VOC 2012 datasets demonstrate that the proposed ACANs significantly outperform the previous state-of-the-art approaches. The side-output of multiple reasoning illustrates that the ACAN progressively finds the semantic context that leads to the correct segmentations.

In Chapter 3, we propose a novel framework for modeling image-dependent con-

textual relationships using graph-based context model. This approach enables us to selectively utilize the contextual relationships suitable for an input query image. We introduce a context link view of contextual knowledge, where the relationship between a pair of annotated regions is represented as a context link on a similarity graph of regions. Link analysis techniques are used to estimate the pairwise context scores of all pairs of unlabeled regions in the input image. Our system integrates the learned context scores into a Markov Random Field (MRF) framework in the form of pairwise cost and infers the semantic segmentation result by MRF optimization. Experimental results on object class segmentation show that the proposed graph-based context model outperforms the current state-of-the-art methods.

Chapter 4 presents a novel nonparametric approach for semantic segmentation using high-order semantic relations. Conventional context models mainly focus on learning pairwise relationships between objects. Pairwise relations, however, are not enough to represent high-level contextual knowledge within images. In this paper, we propose semantic relation transfer, a method to transfer high-order semantic relations of objects from annotated images to unlabeled images analogous to label transfer techniques where label information are transferred. We first define semantic tensors representing high-order relations of objects. Semantic relation transfer problem is then formulated as semi-supervised learning using a quadratic objective function of the semantic tensors. By exploiting the low-rank property of the semantic tensors and employing Kronecker sum similarity, an efficient approximation algorithm is developed. Based on the predicted high-order semantic relations, we reason semantic segmentation and evaluate the performance on several challenging datasets.

In Chapter 5, We combine the likelihood probability with adaptive context aggregation networks from Chapter 2 and the prior probability with second-order and

higher-order context from Chapter 3, 4 for accurate semantic segmentations. Finally, we conclude the dissertation in Chapter 6.

Chapter 2

Adaptive Context Aggregation Network

2.1 Introduction

Recent advances on semantic segmentation are achieved by applying Deep Convolutional Neural Networks (DCNNs) to pixel-wise classification on images [15]. To efficiently utilize DCNNs for semantic segmentation, Fully Convolutional Neural Networks (FCNNs) [16, 14] have been advocated instead of patch-by-patch or region-based classification. Specifically, repurposing convolutional neural network architectures [17, 5, 18] developed for image classification achieves state-of-the-art performance. The success of network fine-tuning shows that the Convolutional Neural Networks have general representation powers on various computer vision tasks. In consequence, DCNNs have achieved state-of-the-art performance in a broad array of vision problems, including object detection [19, 20], semantic segmentation [14], and human pose estimation [21].

Meanwhile, as shown in many previous works on semantic segmentation [22, 23, 24, 25, 1, 26, 27, 28, 29, 30], capturing contextual knowledge of the image is important to achieve a high class accuracy and a visually pleasing labeling. In a DCNNs

architecture, context is implicitly considered to some degree by feeding a sufficiently large surrounding input patch into DCNNs and predicting the labels of the pixels in the input patch simultaneously. However, individual responses at the final layer of FCNNs of each patch are not locally and globally consistent. To address this problem, Chen *et al*. [16] employs refinement step using the conventional Conditional Radom Fields (CRFs) followed by DCNNs. Also, jointly training DCNNs and CRFs are proposed [31, 32]. They independently learn unary/pairwise potential of CRFs using DCNNs [32] or marginalize the probability as a regularized loss.

To this end, we propose a novel approach in which deep convolutional neural network architectures in which appropriately capture semantic context information recursively in this chapter. In our approach, the proposed network has multiple side-output with deep supervision [33, 34] and each side-output helps to parse the next side-output by properly sampling semantic context information. Adaptive context sampling module is proposed by using convolutional spatial transformer network for learning latent context space. The proposed method allows us to efficiently and effectively aggregate context information from networks. This context-aware representation can be generalized to any problems with convolutional neural networks.

The key contributions of this chapter include: (1) The use of a latent space representation of contextual information; (2) A novel adaptive context aggregation network for semantic segmentation; and (3) A convolutional spatial transformer networks for learning the adaptive context aggregation.

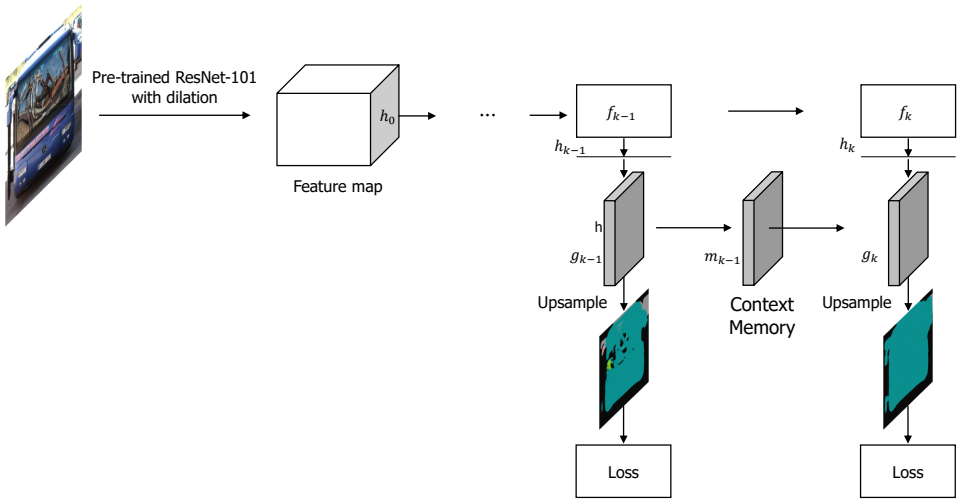


Figure 2.1: Our Network Structure: Our adaptive context aggregation network adds several feature layers to the end of pretrained networks. Side-output layers are inserted after each feature layers. Deep supervision is imposed at each sid-eoutput layer, guiding the side-outputs. Convolutional spatial transformer layer samples the context adaptively for accurate semantic segmentation of the next side-output layer.

2.2 Related Works

We now review related works on semantic segmentation and recent Fully Convolutional Network (FCNs) architectures for semantic segmentation tasks.

Before the deep learning era, much of the works on semantic segmentation is based on hand-crafted features (*e.g*. SIFT, HoG, *etc*.) with the swallow classifiers such as Support Vector Machines, Random Forests, and Boosting. Although these works have proven its feasibility, the performance was unsatisfactory in practice.

In early deep learning systems on semantic segmentation tasks, the central paradigms are replacing hand-craft features to deep features. Mainly, deep learning based classi-

fiers were applied for classifying the regions from the conventional region proposals or superpixel. Farabet *et al*. [15] proposed hierarchical features for superpixels based scene labeling. Girshick *et al*. [35] presented region-based convolutional neural networks for object detection and semantic segmentation. These region-based systems outperform the previous hand-crafted features, but still largely depends on the regions generated from the images.

Recently, in conjunction with successful Deep Convolutional Neural Networks architectures [17, 5, 18, 36] for image classification, pixel-level dense prediction tasks such as semantic segmentation achieve a great breakthrough by reusing and fine-tuning the classification networks [14]. However, image classification networks usually have the low-resolution prediction. To overcome this problem, for object detection tasks, shift-and-stitch approach [37] is proposed to make the image classification networks dense. To address this issue and generate a dense prediction, Long *et al*. [14] incorporated bilinear deconvolution layer and skip connection at the end of the final score map to achieve the same size of the score map with input images. The dilated convolution approaches [16] also called astrous [16] or sparse kernel convolution [38] directly output a middle-resolution score map and apply the dense CRF method [39] to refine boundaries by leveraging color information. It is a generalization of shift and stitch techniques of which used for densifying the classification result for pixel-level classification. [16]. Extending this approaches, simulating the dense CRF with recurrent layers [40, 41] proposed for end-to-end learning. Meantime, in [42, 43], deconvolution layers presented to up-sample the low-resolution features. To enlarge the receptive field of neural networks, methods of [44] proposed to use dilated convolution for context aggregation. Liu *et al*. [45] use global average pooling with L2 normalization to improve segmentation result.

Meanwhile, context plays a crucial role for scene understanding. Many approaches have been proposed to impose the contextual consistency. [46].

2.3 Proposed Method

In this Section, we present the details of our framework. Our network consists of two major networks: embedding networks and deeply supervised context aggregation networks. The network structure of the adaptive context aggregation network is outlined in Figure 2.1. The embedding net extract image feature for later inference. Next, deeply supervised context aggregation net recursively produce semantic segmentation results the while gathering semantic context from the previous segmentation results.

2.3.1 Embedding Network

The embedding network takes the input image and represents it as a feature map. We begin by transforming state-of-the-art classification architectures. We use a pretrained Residual Networks (ResNet) [17] that won ILSVRC16. We pick the 101-layer Residual Networks (ResNet-101) due to the limitation of the GPU memory size. We remove the final classification layer and the average pooling layer, and choose the feature maps from the last remaining convolution layer. However, the repeated combination of striding in ResNet-101 reduces the spatial resolution of the features map to 1/32 of the input image. The most archtectures of ImageNet winners [17, 5, 18] suffer from the downscaling of the feature maps. To alleviate this problem, we use dilated convolution [44], also called atrous convolution [16] or sparse kernel convolution [38]. The dilated networks strategy allows us to compute the feature maps of any layer at any desirable resolution without losing the size of receptive fields. Following the conventional ap-

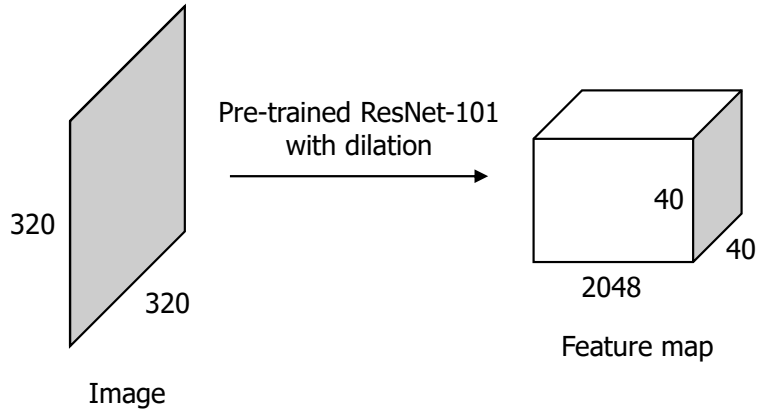


Figure 2.2: Feature map extraction using pre-trained ResNet-101 with dilation.

proaches [16], we densify the resulting feature maps of ResNet-101 by a factor of 8.

Mathematically, the image feature map f_0 is extracted from a raw image I using ResNet-101:

$$f_0 = CNN_{ResNet-101}(I). \tag{2.1}$$

We firstly take the images at a resolution of 320×320 pixels, and obtain the features of the dimension of $40 \times 40 \times 2048$ as shown in Figure 2.2. 40×40 is the number of regions in the image and 2048 is the feature dimension of each region.

2.3.2 Deeply Supervised Context Aggregation Network

In the proposed deeply supervised context aggregation network, we formulate semantic segmentation as the multiple steps of reasoning. We predict the side-output recursively with the help of the aggregated semantic context from the previous side-output.

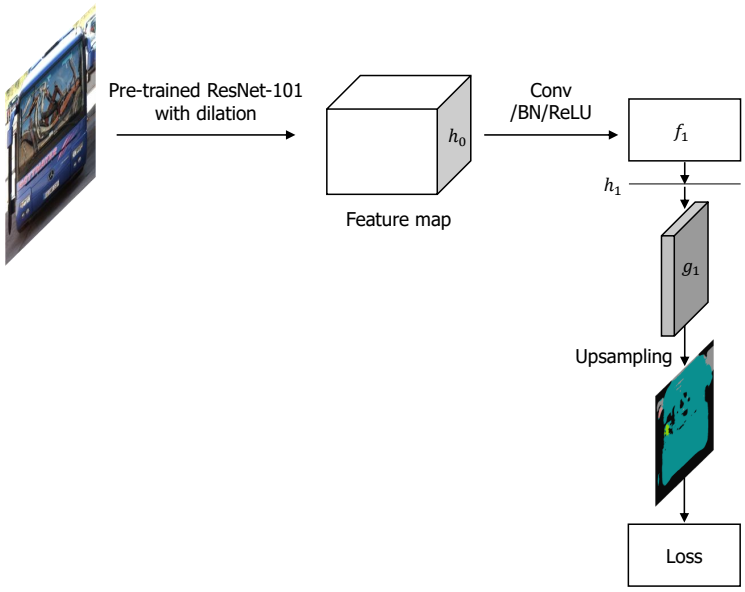


Figure 2.3: The illustration of the process of computing the first side-output g_1.

The deep supervision strategy also can improve the directness and transparency of the hidden layer learning process as pointed out in [33].

Formally, given the image feature maps f_0, deeply supervised context aggregation net predicts the K outputs $g_1, g_2, ...g_K$ sequentially. We first take f_0 and compute the first side-output g_1:

$$f_1 = max(0, bn(W_1 * f_0)) \tag{2.2}$$

$$h_1 = max(0, bn(H_1 * f_1)) \tag{2.3}$$

$$g_1 = G_1 * h_1 + b_1, \tag{2.4}$$

where the $*$ operator represents a convolution, W_1, H_1, G_1 is the weight matrices, b_1 denotes the bias matrix, $max(0, \cdot)$ corresponds to a ReLU operator, and $bn(\cdot)$ denotes

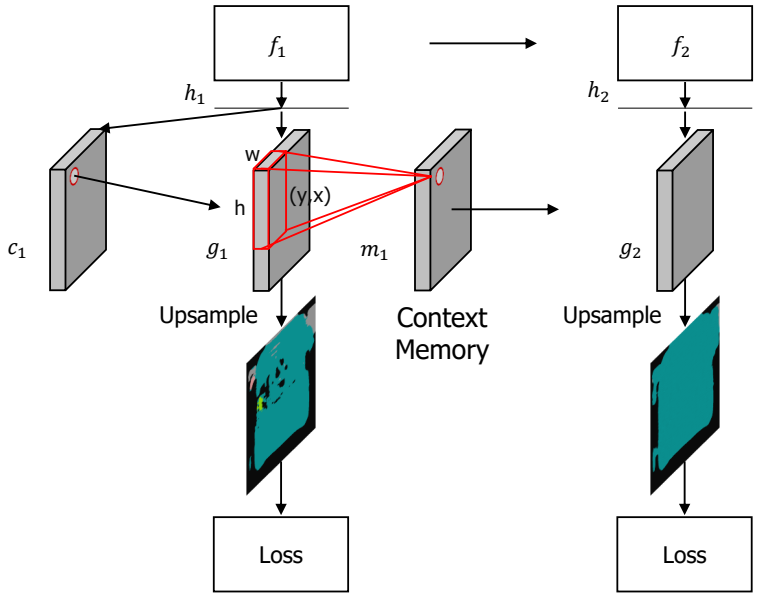


Figure 2.4: The illustration of the process of computing the first context memory m_1 and the second side-output g_2.

the batch normalization operator. f_1, h_1 denotes hidden layer values. This process is illustrated in Figure 2.3.

The first side-output is obtained without the help of adaptive context aggregation, but for the second side-output, we aggregate the semantic context for each neuron and use this information additionally. From each context windows, we adaptively sample the semantic context and embeds it to the feature space. Using the first side-output g_1, we compute context memory feature map m_1. To compute m_1 we utilize the convolutional spatial transformer network [47] extending the original spatial transformer network [48]. We take the input h_1 for generating context windows parameters and g_1 to sample the appropriate context.

$$c_1 = C_1 * h_1 + b_1^c \tag{2.5}$$

$$s_1 = \tau_{c_1}(g_1) \tag{2.6}$$

$$m_1 = max(0, bn(M_1 * s_1)) \tag{2.7}$$

where C_1, M_1 is the weight matrices, b_1^c denotes the bias matrix, and $\tau_{c_1}(\cdot)$ represents convolutional spatial transformer operator. c_1, s_1 denotes context windows parameters and sampled context, respectively. This process is illustrated in Figure 2.4.

The computation of the second side-output is consequently as follows:

$$f_2 = max(0, bn(W_2 * f_1)) \tag{2.8}$$

$$h_2 = max(0, bn(H_2 * f_2)) \tag{2.9}$$

$$g_2 = G_2 * [h_2, m_1] + b_2, \tag{2.10}$$

where the W_2, H_2, G_2 is the weight matrices, b_2 denotes the bias matrix, and $[\cdot, \cdot]$ represents feature concatenations. This process is illustrated in Figure 2.4. Generally, the k-th side-output g_k with the following formula:

$$f_k = max(0, bn(W_k * f_{k-1})) \tag{2.11}$$

$$h_k = max(0, bn(H_k * f_k)) \tag{2.12}$$

$$g_k = G_k * [h_k, m_{k-1}] + b_k, \tag{2.13}$$

where the W_k, H_k, G_k is the weight matrices and b_k denotes the bias matrix.

Semantic adaptive context aggregation with fixed size of windows can help to segment an object as shown in Table 2.1. However, which relative position and scale of semantic context are helpful for each neuron is not known generally. Therefore, we

Table 2.1: Performance comparison of our algorithm on PASCAL VOC 2012 validation dataset.

	Mean IoU	Pixelwise accuracy
Baselin FCN	72.7	93.5
Adaptive context agg. without STN	74.4	93.8
Adaptive context agg. STN	**75.0**	**93.9**

train a sptial transformer networks [48] for each neuron and propose a convolutional version to generate the transformation of side predictions by deep supervision. As shown in our experiments, this is especially important for properly capture the context knowledge.

The proposed transformer network takes input from lower feature maps and applies independent spatial transformation. The parameters of sampler is convolutionally trained. Since the independent transformation, the transformed semantic contexts can be larger or smaller than nearby neuron.

Then, we learn a mapping into a feature space where contexts are represented. The latent mapping of context windows concatenated with output of convolution layers and produce the predictions. The mapping function is learnt to optmize only with the supervised loss of interest for our task, semantic segmentation.

2.4 Experiments

To show the effectiveness of our approach, we perform extensive experiments on two public datasets, PASCAL VOC 2012 and SiftFlow dataset. The segmentation accuracy is measured by the intersection-over-union (IoU) score, and the pixel accuracy and the

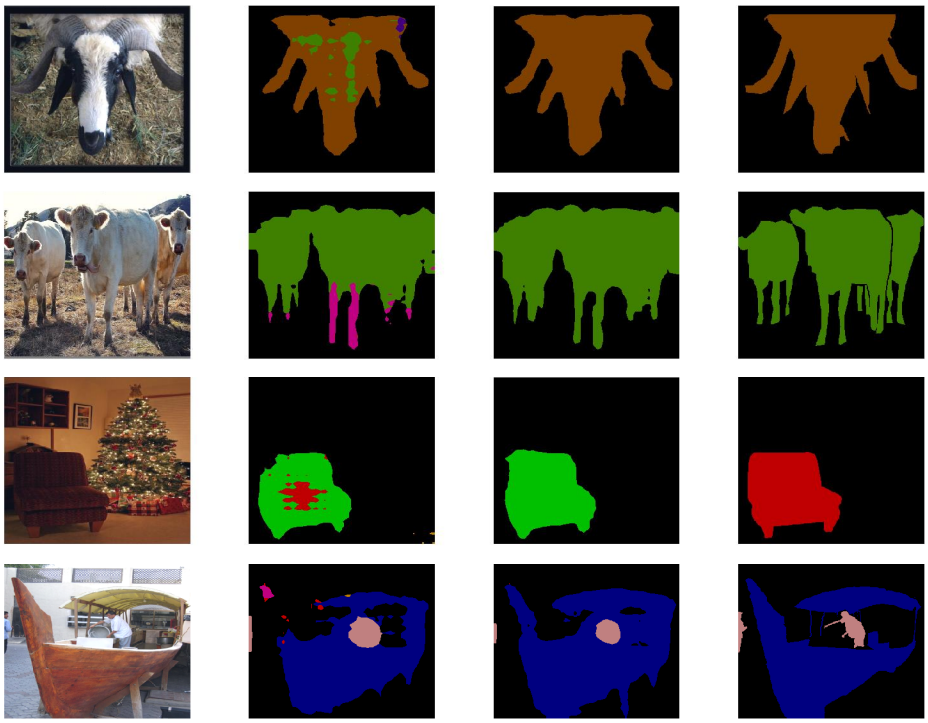
Table 2.2: Performance comparison of our algorithm on PASCAL VOC 2012 test dataset.

	Mean IoU
GCRF [49]	73.2
DPN [50]	74.1
Piecewise [32]	75.3
Adaptive context agg. STN	**77.2**

mean class accuracy over all categories.

Implementation. We apply data augmentation during training. In particular, we perform random scaling (ranging from 0.5 to 2.0), random cropping, random rotation (ranging from -10 to 10) and horizontal flipping. We consider the momentums of batch normalization layer of residual networks [17] as a constant and fix the learning rate of scale and bias of the layer. The learning rate of the pre-trained layers starts with to 0.001, and the learning rate of new layers are set to 0.01 initially, futher multiplying the learning rate by 0.1 every 30k iterations. We also apply test-time multiscale evaluation and use the minibatch of 16 images. We have implemented the proposed methods by extending the MatConvNet framework [51].

Baseline FCN: Our baseline networks is based on the Dilated version of Fully Convolutional Networks (FCNs). The original FCNs is trained using VGG net [18] and use skip-connection instead of the dilated filters, but in our experiments, we append the 7×7 convolutional layer to predic scores for each of the object classes on the feature maps in Section 2.3.1. Hence, it works similar to the *DeepLab* proposed in [16].



(a) Input image (b) Baseline FCN (c) Adaptive context agg. (d) Ground truth

Figure 2.5: Representative results from the PASCAL VOC 2012 validation dataset. (a) Input images. (b) Baseline FCN results. (c) The proposed adaptive context aggregation network. (d) Ground truth.

2.4.1 PASCAL VOC 2012 dataset

PASCAL dataset is proposed by Everingham *et al*. [12] which includes 20 object categories and on background class. The dataset is split into a training set, a validation set, and a test set, with 1464, 1449, and 1456 images each. We compare our framework with Baseline FCN with mean IoU (intersection-over-union) score and pixel accuracy score. Furthermore, we compare the performance of proposed frameworks without some of the components as shown in Table 2.3. The selected prediction examples are

Table 2.3: Effect of each components in the proposed method. In this experiments, we train only with original 1464 training images and use single scale testing.

	Mean IoU	Pixelwise accuracy
Baseline FCN	67.3	92.3
Deep supervision	68.9	92.7
Adaptive context agg. without STN	69.8	92.9
Adaptive context agg. with STN	**70.6**	**93.0**

shown in Figure 2.5.

As shown in Table 2.1 and 2.2, our system achieves an overall mean IoU scores 77.2% on the *test* set and outperforms baseline FCN in *val* set.

Component analysis: In Table 2.3, we analyze each components of the proposed methods. Continually adding the components we observe performance increasing. In this experiments, we do not use additional augmentation of the data proposed by [52] and single-scale evaluation is used.

2.4.2 SIFT Flow dataset

We also compare the proposed method on The SIFT Flow dataset as shown in Table 2.4. The SIFT Flow dataset provided by Liu *et al*. [2] consists of 2,688 images of outdoor scenes. The dataset provides ground truth labels hand-annotated by LabelMe users. Liu *et al*. [2] split this dataset into 2,488 training images and 200 test images, and selected top 33 object categories as semantic labels. For comparison, the same training/test split is used as in [2, 4].

Class Balancing: Since the SIFT Flow dataset has heavily unbalanced class distri-

Table 2.4: Evaluation of the proposed algorithm on SIFT Flow dataset. Our algorithm achieves state-of-the-art performance in mean class accuracy metric.

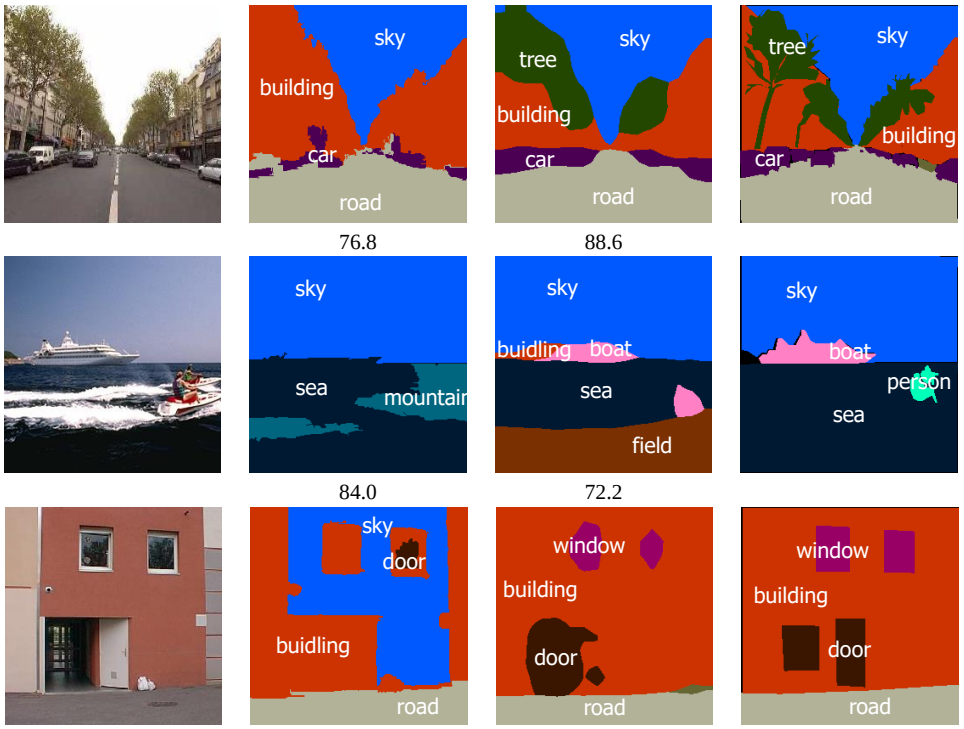
	Pixel accuracy	Mean class accuracy
Farabet *et al*. [15]	78.5	50.8
Long *et al*. [14]	85.2	51.7
Adaptive context agg. without STN	85.6	55.3
Adaptive context agg. with STN	**85.8**	**56.0**

bution, we compensate the loss with natural class frequency in the dataset. The loss function of each pixel is normalized by the frequency of the ground truth labels as follows:

$$w = \frac{1}{log(\frac{P_c}{median_i(P_i)} + 1)}, \tag{2.14}$$

where P_i is # of pixels of class i and c is the ground truth label of the pixel.

As shown in Table 2.4, our system achieves an overall pixel-level accuracy of 85.8% and a per-class accuracy of 56.0%.

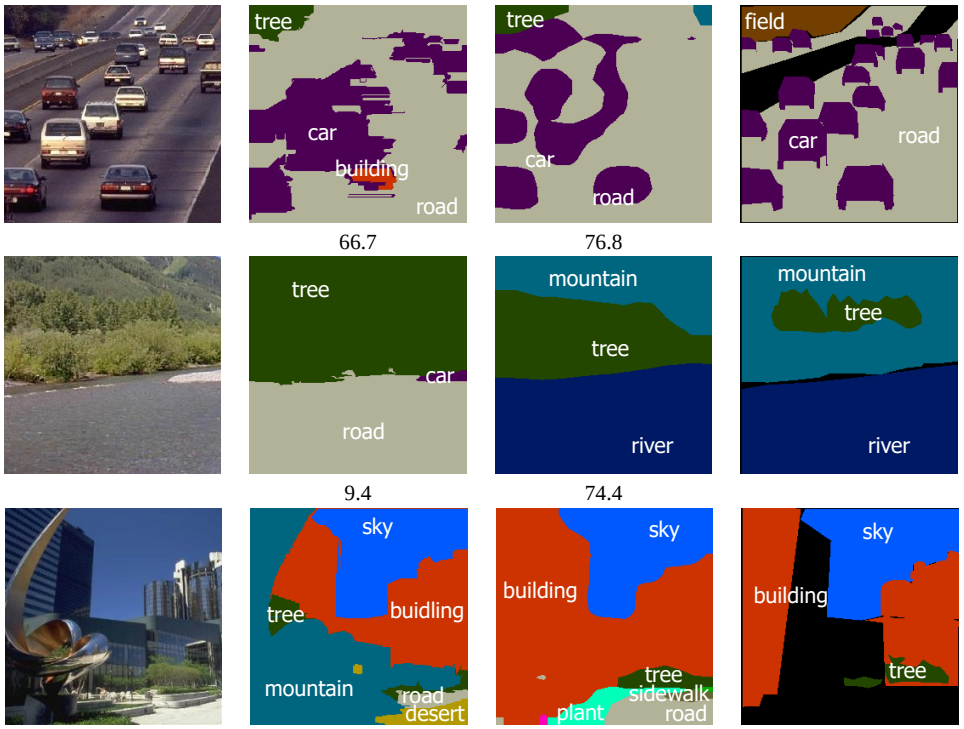


(a) Input image (b) SuperParsing (c) Adaptive context agg. (d) Ground truth

Figure 2.6: Representative results from the SIFT Flow dataset. (a) Input images. (b) The output of SuperParsing. (c) The output of adaptive context agregation networks. (d) Ground truth. The number below the image shows pixelwise accuracy.

2.5 Summary

In this chapter, we have proposed adaptive context aggregation module with convolutional neural network for semantic segmentation. In this framework, context information adaptively crawled while multistep reasoning process. This Adaptive output aggregation can be further applied to video object and image segmentation.



(a) Input image (b) SuperParsing (c) Adaptive context agg. (d) Ground truth

Figure 2.7: Representative results from the SIFT Flow dataset. (a) Input images. (b) The output of SuperParsing. (c) The output of adaptive context agregation networks. (d) Ground truth. The number below the image shows pixelwise accuracy.

Chapter 3

Second-order Semantic Relationships

3.1 Introduction

Recent works [22, 23, 24, 25, 1, 26, 28, 29, 30] have shown that employing contextual information is extremely helpful for resolving this problem. There are various sources of context including scene [25, 30], semantic [28, 29], scale [23, 28], and spatial relation [53, 54]. Recently, many researchers have highlighted the importance of pairwise relationships between objects [53, 54, 1, 28, 29]. This relationship is commonly represented by high-level statistics such as the object class co-occurrence which captures semantic context between object classes. For example, building and road are likely to co-occur in an image. To incorporate object relationships, traditional approaches often model such relations as local interactions between pixels or regions. To produce the final labeling result, the obtained object relationships are combined with pre-learned unary potentials which are usually learned based on the visual features of the objects. This scenario, separately learning contextual relationships and visual appearance, has been successfully used to solve the scene understanding problem.

However, this system tends to prefer frequently appeared objects to enforce object

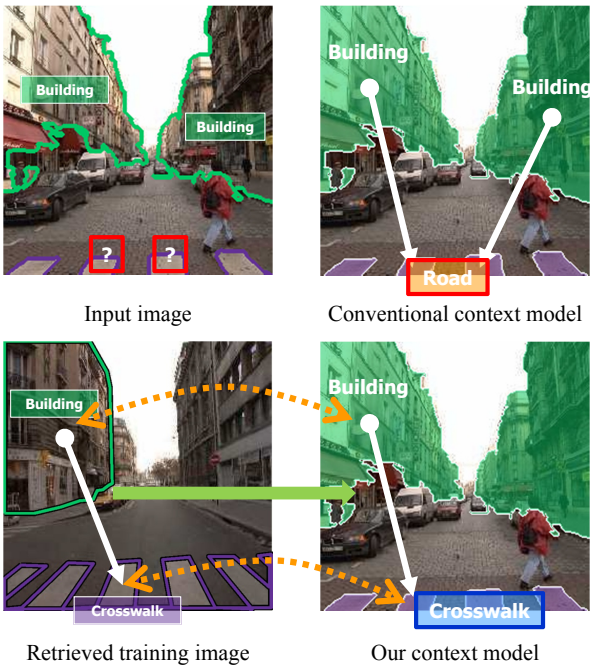


Figure 3.1: Comparison of our context model to a conventional context model based on co-occurrence statistics. We appropriately establish the object relationship depend on the visual appearance as well as the contextual relation from the matched similar scene.

label agreement according to semantic relevance. For example, consider the example illustrated in Figure 3.1, where the ground truth label of the unknown regions is *crosswalk*. Notice that the regions labeled as *building* enforce the unknown regions to be labeled as *road* because building and road are more strongly correlated than building and crosswalk. Furthermore, as pointed out in [55], conventional context models are not invariant to the number of pixels/regions that an object occupies, which makes the small objects likely to be eliminated. Our key idea is to utilize context relationships adaptively according to the visual appearance of objects to correctly label such unknown regions.

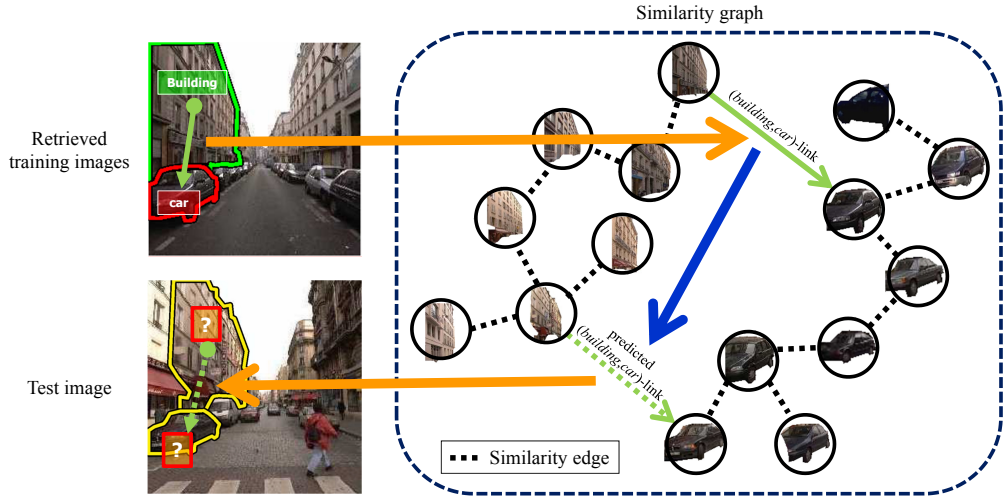


Figure 3.2: Illustration of our approach. The contextual relationship between the pair of the annotated regions is represented as (*building*,*car*)-link between the two corresponding nodes on the similarity graph. No link is constructed between the two regions from the test image because they are unlabeled. By applying link analysis techniques [3], our system predicts the strength of (*building*,*car*)-link between them based on node similarity.

In this work, we present a novel approach for properly capturing the contextual relationships between two regions by considering the content of an input image. One difficulty is that learning such relations between all pairs of regions across whole object classes is computationally challenging. To overcome this problem, we propose a novel nonparametric exemplar-based context model. This nonparametric context model consists of a bunch of *context exemplars* which are basically annotated region pairs extracted from similar training images to the input image. From these context exemplars, we provide the novel interpretation of context exemplars in a context link view and relate the problem of learning contextual relationships to the link prediction problem

on a similarity graph of regions. The configuration of the similarity graph naturally reflects visual similarity between regions. On this similarity graph, all context exemplars can be compactly encoded in the form of context links. Moreover, the similarity graph is usually sparse, so computation of learning contextual relationships can be done very efficiently. Although traditional context models for high-level object interaction is staying in refinement of local labeling result with co-occurrence statistics, our context transfer approach provides additional clue for reasonable image labeling as illustrated in Figure 3.3.

The key contributions of this section are as follows. (1) We establish a novel context link view of contextual knowledge. (2) In this view, we formulate the problem of learning object relationships as graph-based link prediction problem which can be efficiently solved via state-of-the-art link analysis techniques [56, 3]. (3) Our system is nonparametric and exemplar-based, and therefore does not need to see whole training images to build a context model. Hence, it easily scales to large datasets with the tremendous number of images and object classes. Our system can also infer contextual relationships even from a single training image.

The rest of the chapter is organized as follows. In Section 3.2, we review some relevant works. Section 3.3 presents our context model. Section 3.4 describes a region labeling algorithm using our context model. Section 3.5 provides experimental results and related discussion, followed by a conclusion in Section 3.6.

3.2 Related Work

There are two different types of approaches to object relationships. The first type focuses on the neighborhood interactions that captures the relation of two classes be-

tween nearby pixels/regions. To obtain it, various approaches have been proposed such as simple continuity preference [2], training classifier over pairwise features [23], and penalty term using co-occurrence statistics [4, 57]. However, the adjacent interactions is limited to modeling local properties of the image. Nevertheless, many existing non-parametric scene parsing methods [2, 4, 57] have employed neighboring relationships due to the scalability. The second type, on the other hand, models high-level relationships among objects by considering both long range and neighboring dependencies. This context model is typically represented by co-occurrence statistics or spatial relationships between object classes. Ravinovich *et al*. [29] incorporated co-occurrence statistics into the fully connected Conditional Random Field (CRF). Galleguillos *et al*. [53] proposed exploiting the information of relative location such as *above, beside,* or *enclosed* between object classes. Gould *et al*. [54] designed a more complex and informative relative location prior among object classes. Parikh *et al*. [28] differently learned co-occurrence statistics according to location and scale information. However, all these existing global context models rely on pixels/regions label prediction and are unable to incorporate visual appearance information effectively during context learning stage.

Jain *et al*. [1] proposed adaptively predicting "what" object relationships to consider and "how" to evaluate these relationships based on local and global image features. They learnt class-specific pairwise feature weights in a nonparametric manner, but they only consider simple relative position, overlap, and brightness. Different to Jain *et al*. [1], our approach relies on context link, allowing us to model complex object relationships directly associating to object classes.

Perhaps one of the most similar works to our approach is that of Malisiewicz and Efros [58]. They developed the Visual Memex graph with similarity and contextual

edges. In contrast to [58], we build the memex at query time only using matched images on global similarity level. Furthermore, our system reasons the strength of contextual relationships between regions, while [58] only predicts the category of a hidden object with some provided objects. This paves new promising way of representing and embedding higher-level semantic contextual relationships among objects in scene parsing and understanding.

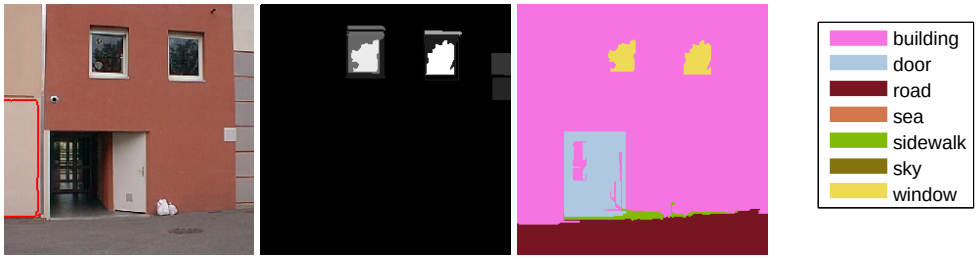
3.3 Our Approach

3.3.1 Overview

For a query image, we first retrieve its best matched similar scenes in a large dataset using global descriptors analogous to several nonparametric scene parsing methods [2, 59, 4]. All pairs of the annotated regions in the retrieved scenes can be defined and exploited as *context exemplars*. A context exemplar is composed of a pair of regions and a pair of the corresponding object classes. It represents that a region with its particular object class supports the paired region to have its corresponding object class. For example, in Figure 3.2, the contextual relation from the region labeled as *building* to the region labeled as *car* forms a context exemplar. This means, when the former region is labeled as *building*, the latter region would be labeled as *car*. Note that this context exemplar can capture the global interaction between regions and is not limited to the local adjacent interaction.

Our goal is to estimate how much each region pair of the query image is consistent with the context exemplars from the retrieved images. For this, we first construct the similarity graph in which unlabeled regions from the test image and the annotated regions from the matched scenes are regarded as nodes. Each context exemplar is then

(a) Query image (b) Oversegmentation (c) Local labeling (d) Baseline MRF



(e) Sample region (f) Learned context (g) Final labeling

Figure 3.3: Illustration of image parsing using our context transfer approach. Top row shows the failure of the baseline context model: (a) the given query image, (b) the over-segmented regions, (c) the local labeling result based on local features using implementation [4], and (d) the CRF labeling result with baseline context model. Since the incorrect local labeling dominates the final performance, it is not easy to identify the window regions correctly. Bottom row: (e) assume that the selected region (encircled with red line) is *building*, (f) the example of the learned contextual consistency scorse by our context transfer approach how much each region will be *window* with the given region (e) (normalized for visualization), and (g) the CRF final labeling result combining local labeling in (c) with the learned contextual consistency scores. The explicitly learned contextual consistency scores successfully corrects the final result making the window regions appear.

encoded as a link between two nodes with the corresponding object classes on the similarity graph as illustrated in Figure 3.2. By applying the label propagation technique, a kind of semi-supervised learning method, the links between all nodes of the query image are constructed with their associated scores. Note that this label propagation method was originally proposed to solve the node classification problem [60]. After that, many researchers [56, 3] extended it to predict the relations among the nodes. In this work, we follow the approach of [3] because it is efficient compared to other methods [56]. Finally, the learned context scores are incorporated into the MRF framework for final labeling.

3.3.2 Retrieval System

A confident image set for the input test image is first extracted from a large training dataset because it is not scalable to consider all context exemplars from whole labeled training images. What we expect to have in the retrieval set are similar objects with consistent spatial arrangement compared to the test image. Hence, retrieval is done not only for computational efficiency but also for more informative region-based context learning.

Four different types of global image features are used: color histogram, spatial pyramid [61], gist [62], and tiny image [63]. For each feature, top-scored $T/4$ images according to the ranking scores are collected and used as the retrieval set similar to [4]. Having the best matches from each of the global features allows us to take into account various examples of scene context with the different views. All pairs of annotated regions in this retrieved set will form the context exemplars and serve as the source of region-level context learning.

3.3.3 Graph Construction

The k-nearest neighbor *similarity graph* is constructed between regions from both the test image I and the corresponding retrieved image set $\hat{\mathcal{I}}$. Each image is segmented into a number of regions based on the fast graph-based segmentation algorithm [64], and then each region is described by its appearance using selective shape, location, texture, color, and appearance features same as in [4]. The similarity graph is defined as a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V}$ is a set of vertices that contains a set of regions $S_U = \{s_1, ..., s_M\}$ from the test image I and a set of regions $S_L = \{s_{M+1}, ..., s_N\}$ from the retrieved images $\hat{\mathcal{I}}$. Each vertex is connected to its k-nearest neighbor. A weight $w_{ij} \in \mathcal{W}$ is assigned to an edge $e_{ij} \in \mathcal{E}$, and is defined by the following similarity measure comparing two regions s_i and s_j based on Gaussian kernel:

$$w_{ij} = \prod_{H_k \in \mathcal{H}} \exp\left(-\frac{\|H_k(s_i) - H_k(s_j)\|}{\sigma_{H_k}}\right), \tag{3.1}$$

where $H_k(s_i)$ is the feature vector of the k-th type for s_i, $\mathcal{H}$ represents the set of features arranged in Table 3.1, σ_{H_k} denotes the standard deviation of H_k, and all features are equally weighted.

3.3.4 Context Exemplar Description

In this step, the contextual relationships within the retrieved scenes are extracted in the form of contextual exemplars. Instead of counting co-occurrence or voting spatial arrangement between object classes, we simply extract all pairs of the annotated regions from the retrieved scenes and represent each pair as a context exemplar with the corresponding pair of object classes. More formally, given a set of classes $\mathcal{C} = \{c_1, c_2, ..., c_K\}$ (e.g. *sky, building, ..*) containing all existing classes in the correspond-

Types	Feature Name	Dim.
Shape	Relative area	1
	Shape mask over its bounding box	64
	Relative width/height	2
Location	Shape mask	64
	Relative top height	1
Texture	Texton hist	100
	Dilated texton hist.	100
SIFT	SIFT hist.	100
	Dilated SIFT hist.	100
Color	RGB color mean	3
	RGB color std	3
Appearance	Gist over its bounding box	320

Table 3.1: Selected region features for constructing the similarity graph.

ing retrieved image set $\hat{\mathcal{I}}$, the set of context exemplars for each class pair (c_a, c_b) is represented as

$$M^{ab} = \{(s_i, s_j) : G(s_i) = c_a, G(s_j) = c_b, s_i, s_j \in \hat{\mathcal{I}}_l\}, \tag{3.2}$$

where $s_i, s_j \in \hat{\mathcal{I}}_l$ represents two regions s_i and s_j in the same image $\hat{\mathcal{I}}_l$ included in the retrieved image set $\hat{\mathcal{I}}$ and $G(s_i)$ represents the ground truth class of region s_i. Note that the order of all pairs (s_i, s_j) should be preserved since each context exemplar is assumed to have direction. Hence, based on region pair (s_i, s_j) labeled as (c_a, c_b), two context exemplars $(s_i, s_j) \in M^{ab}$ and $(s_j, s_i) \in M^{ba}$ are constructed. We hold $\mathcal{M} = \{M^{11}, M^{12}, M^{13}, ..., M^{KK}\}$ for all object class pairs and this contains

the whole contextual relationships within the retrieved image set $\hat{\mathcal{I}}$ without loss of information.

Our key observation is that a context exemplar $(s_i, s_j) \in M^{ab}$ can be viewed as a directional (c_a, c_b)-type link between two nodes s_i and s_j on the similarity graph. We will refer to this link as the (c_a, c_b)-link. To transform all context exemplars into context link form, let $\mathcal{F}$ denote the set of $N \times N$ matrices with nonnegative entries. A matrix $\mathbf{F}^{ab} \in \mathcal{F}$ associates to (c_a, c_b)-links and $[\mathbf{F}^{ab}]_{ij}$ represents the strength of (c_a, c_b)-link between two nodes s_i and s_j. The strength close to 1 means high confidence of the existence of a link. On the other hand, the strength close to 0 means the absence of a link. We define $\mathbf{Q}^{ab} \in \mathcal{F}$ to represent the observed (c_a, c_b)-links within the retrieved images such that

$$[\mathbf{Q}^{ab}]_{ij} = \begin{cases} 1 & \text{if } (s_i, s_j) \in M^{ab} \\ 0 & \text{otherwise} \end{cases} . \tag{3.3}$$

Now we have a set of context link $\mathcal{Q} = \{\mathbf{Q}^{11}, \mathbf{Q}^{12}, \mathbf{Q}^{13}, ..., \mathbf{Q}^{KK}\}$.

3.3.5 Context Link Prediction

Link prediction problem is a task of predicting how likely a link exists in a network. In this work, we consider a problem of predicting (c_a, c_b)-link among the pairs of nodes of S_U based on Q^{ab} consistent to the configuration of the similarity graph. For this, we adopt semi-supervised link propagation approach using node similarity similar to [56]. We directly propagate (c_a, c_b)-links in Q^{ab} to the pairs of nodes of S_U and estimate the strength of them. We assume that all $\mathbf{Q}^{ab}$ is uncorrelated to each other, therefore, context link prediction problem can be solved by K^2 independent link propagation problems. We drop the ab suffix for clarity.

However, directly applying the approach of [56] to our context link prediction

Procedure 1 Proposed Context Learning Algorithm

Input: Query image I

Output: Learned context scores $L(s_i, c_i, s_j, c_j)$

1: Retrieve the scene-level similar image set $\hat{\mathcal{I}}$

2: Generate superpixels S_U of the query image I

3: Construct the similarity graph $\mathbf{W}$ of regions S_U from I and S_L from $\hat{\mathcal{I}}$

4: Derive the matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ in which $\mathbf{D}$ is a diagonal matrix with its (i, i)-element equal to the sum of the i-th row of $\mathbf{W}$.

5: Extract the context exemplars $\mathcal{M}$ from $\hat{\mathcal{I}}$

6: Build the context link $\mathcal{Q}$

7: **for** each object class pair (c_a, c_b) **do**

8: Initialize $\mathbf{F}_c(1) = \mathbf{F}_r(1) = \mathbf{0}$

9: (Column-wise link propagation)

10: Iterate $\mathbf{F}_c(t+1) = (1-c)\mathbf{L}\mathbf{F}_c(t) + c\mathbf{Q}^{ab}$ until convergence

11: (Row-wise link propagation)

12: Iterate $\mathbf{F}_r(t+1) = (1-c)\mathbf{F}_r(t)\mathbf{L} + c\hat{\mathbf{F}}_c$ until convergence where $\hat{\mathbf{F}}_c$ indicates the limit of $\{\mathbf{F}_c(t)\}$

13: Assign $L(s_i, c_i = c_a, s_j, c_j = c_b) = [\hat{\mathbf{F}}_r]_{ij}$ where $1 \leq i, j \leq M$ and $\hat{\mathbf{F}}_r$ denotes the limit of $\{\mathbf{F}_r(t)\}$

14: **end for**

problem is impractical because it requires $O(N^4)$ times for a link propagation. Thus, we follow the strategy of the constraint propagation for spectral clustering [3]. We decompose the link propagation problem into two independent label propagation sub-problems. First, the j-th column $\mathbf{Q}_{\cdot j}$ serves as an initial configuration of two-class label propagation problem with respect to s_j. We will refer this process as a column-

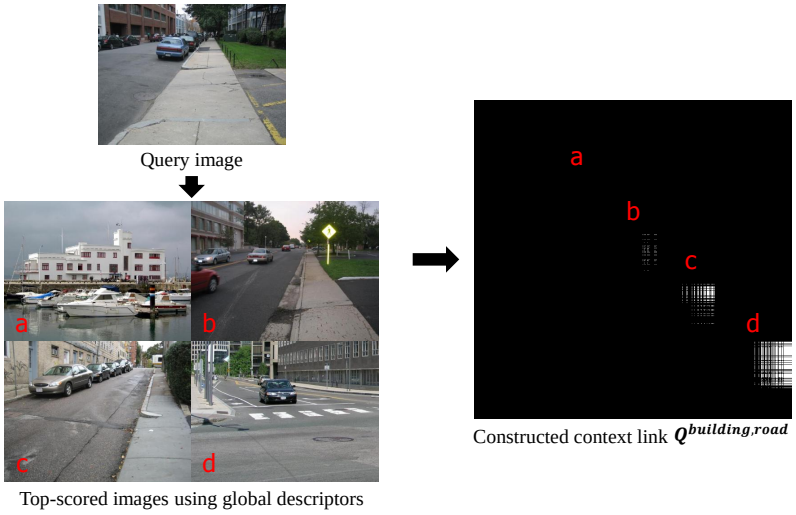


Figure 3.4: Example of constructed context link $Q^{building,road}$ from the annotated regions of the top-ranked $T = 1$ retrieved scenes. Since building is appeared but no road is presented in image a, no context link Q^{ab} is built.

wise link propagation. The work of Zhou *et al.* [60] is employed to solve the label propagation problem with respect to s_j. All columns of $\mathbf{Q}$ are handled separately and the converged configuration $\hat{\mathbf{F}}_c$ (Step 10) is obtained. In practice, we observed that the columns of $\mathbf{Q}_{.j}$ within a retrieved image are exactly same as shown in Figure 3.4. Therefore, only T, the number of retrieved images, of column-wise link propagation is required not N ($T \ll N$).

Next, the i-th row of $[\hat{\mathbf{F}}_c]_{i.}$ is set as an initial configuration of two-class label propagation problem with respect to s_i. This is a row-wise propagation which works similar to the column-wise propagation. Practically, only what we want to obtain is the link information within the query image. Hence, row-wise link propagation with ($M < i \leq N$) is not necessary. After convergence of the row-wise iteration (Step 12), the strength of (c_a, c_b)-link between two nodes s_i and s_j within the query image I is

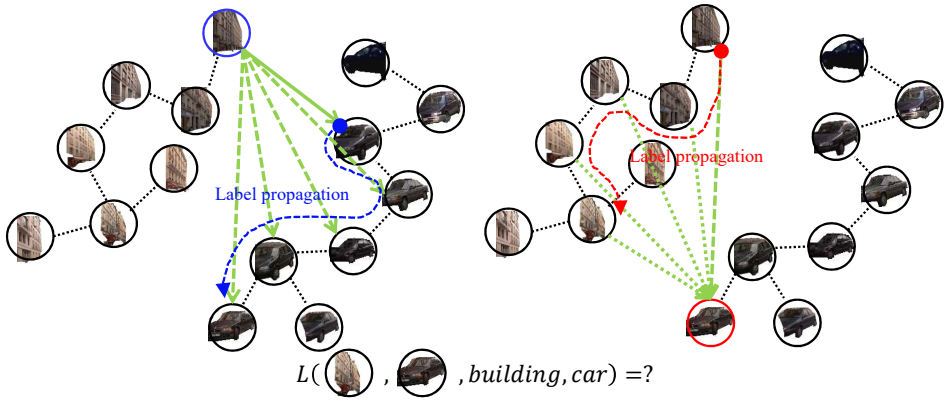


Figure 3.5: The process of two stage context link prediction. Two individual label propagation approximate link prediction process.

obtained.

Learning is independently performed for each Q^{ab} and repeated K^2 times. Each context learning is solved $O(kN^2)$ times on the k-nearest neighbor similarity graph ($k \ll N$) [3]. Therefore, the overall complexity of learning the context scores using our approach is $O(K^2N^2)$.

3.4 Inference

To assign labels to a set of regions S_U, the learned context scores $L(s_i, c_i, s_j, c_j)$ are incorporated to the fully connected MRF model. The fully connected model is proved to be effective for encoding the object interactions [53, 28, 29]. Similar to that of [53, 28, 29], we define the energy function of object class labels $\mathcal{C} = \{c_1, c_2, .., c_K\}$

Table 3.2: Performance comparison of our algorithm on Jain *et al*. [1] dataset and SIFT Flow dataset [2]. Per-pixel rates and average per-class rates in parentheses are presented.

	Jain *et al*. [1] Dataset	SIFT Flow dataset [2]
Jain *et al*. [1]	59.0 (-) [1]	-
Chen *et al*. [65]	75.6 (45) [65]	-
Liu *et al*. [2]	-	74.75 (-) [2]
Tighe and Lazebnik [4]	-	76.82 (29.38) [4]
Baseline classifier	77.62 (49.45)	73.35 (29.04)
Baseline MRF	76.48 (47.13)	74.08 (26.87)
Our (without ψ_i)	76.35 (45.72)	71.51 (30.84)
Our (with ψ_i)	**80.14** (**53.25**)	**77.14 (32.29)**

as:

$$\mathcal{J}(\mathbf{c}) = \sum_{i=1}^{M} \psi_i(c_i) + \lambda \sum_{i,j=1}^{M} \phi_{ij}(c_i, c_j), \tag{3.4}$$

where M is the number of regions in the test image I. The data term $\psi_i(c_i)$ represents the negative logarithm of the probability of class c_i given the region s_i. To obtain $\psi_i(c_i)$, we train discriminative classifiers from training dataset using visual features [4]. The smoothness term $\phi_{ij}(c_i, c_j)$ indicates pairwise contextual cost between the regions learned by our approach. This can be written as

$$\phi_{ij}(c_i, c_j) = -\log(\frac{1}{Z} L(s_i, c_i, s_j, c_j)), \tag{3.5}$$

where $Z = \sum_{i=1}^{M} \sum_{c_i}^{K} L(s_i, c_i, s_j, c_j)$ is the normalization constant. Notice that the energy function is controlled by λ, which is the influence of the learned context scores.

To minimize the MRF energy function, we applied α-expansion algorithm [66, 67] using the Quadratic Pseudo-Boolean Optimization (QPBO) algorithm [68, 69] which is publicly available [1].

3.5 Experiements

In this section, we report experimental results on two challenging datasets: the dataset of Jain *et al*. [1] and SIFT Flow dataset [2]. We evaluate the performance of the learned context scores and compare the accuracy of our approach both to a baseline and to recent state-of-the-art results. In each experiments, we evaluate four different models: a baseline classifier without MRF model; a baseline MRF with convential co-occurrence prior; our approach without unary potential; our approach. Our implementation is in MATLAB based on the available SuperParsing code [2]. We fix the parameters of our system with $T = 16, k = 10, c = 0.9, \lambda = 1$ in all experiments.

Baseline MRF: We evaluate the performance of the proposed approach against a conventional co-occurrence based model for object interaction. Following the most successful approaches [29], we incorporate the object class co-occurrence as local interaction into the fully connected MRF model. Hence, we design a baseline MRF model that has different form of the smoothness term to our model as

$$\phi_{ij}(c_i, c_j) = -\log(\frac{P(c_i|c_j) + P(c_j|c_i)}{2}) \times \delta[c_i \neq c_j], \tag{3.6}$$

where $P(c_i|c_j)$ is the empirical probability of classes c_i and c_j co-occurring in the training images.

Jain *et al*. [1] Dataset: Jain *et al*. [1] dataset contains total 350 images randomly se-

[1] http://pub.ist.ac.at/ vnk/software.html

[2] http://www.cs.unc.edu/ jtighe/Papers/ECCV10/index.html

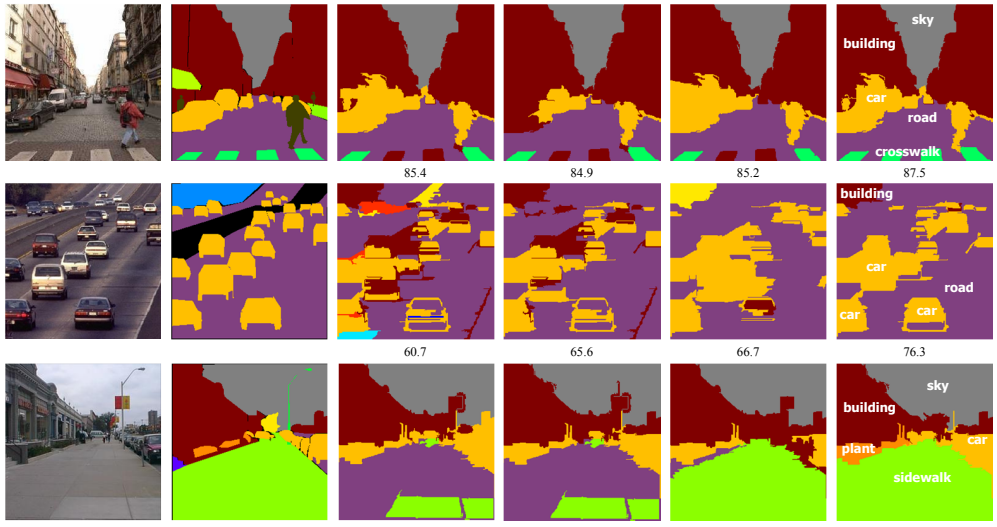


Figure 3.6: Representative results from the SIFT Flow dataset. Column (a) shows the query image to be labeled and Column (b) represents the ground truth of (a). Column (c), (d), (e), and (f) show the prediction of the baseline classifier, baseline MRF models, SuperParsing [4], and our approach, respectively. The numbers under each image indicates pixel-wise accuracy (%) on that image. Crosswalk is appeared in the first row, building is removed without smoothing in the second row, and sidewalk and plant are recovered in the last row. Obviously, implausible baseline classifier results are appropriately corrected based on the learned context scores. These figures are best viewed in color.

lected from LabelMe [70] dataset with 19 classes (250 training and 100 test images). We train boosted decision tree classifiers [25] for computing ψ_i terms. Per-pixel and per-class rates are presented in Table 3.2. Our system has an overall pixel-wise accuracy of 80.14% and a class-wise accuracy of 53.25%. We achieve pixel-wise 5% and class-wise 8% improvement over state-of-the-art performance [65]. Compared to the

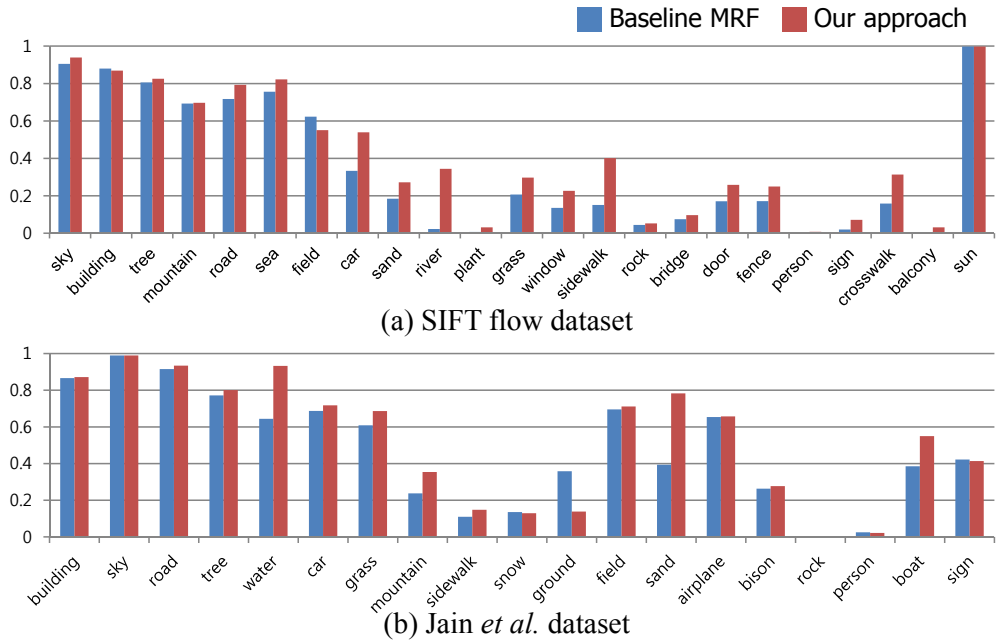
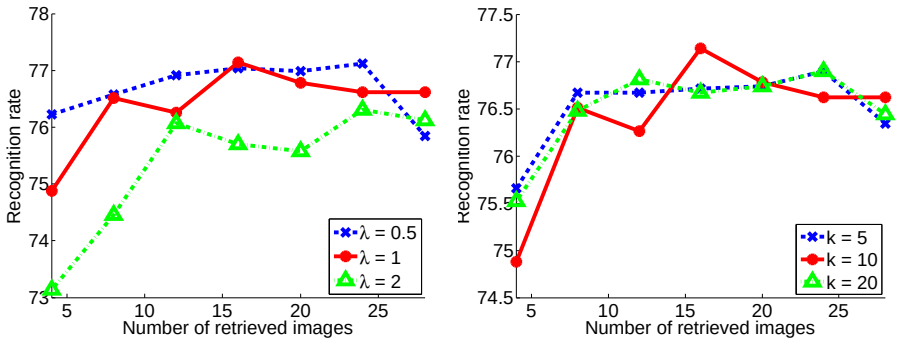


Figure 3.7: The per-class recognition rate of our system compared with baseline MRF on (a) SIFT flow dataset and (b) Jain *et al*. [1] dataset. Note that categories has 0% accuracy are not shown in (a).

baseline MRF, our approach improves overall per-pixel rates by about 4% and this result clearly shows the advantage of our approach. More importantly, baseline MRF drops per-class rates since the conventional context models smooth away smaller object classes. On the other hands, our approach does not suffer from such a problem and even improves per-class rates by 3%.

SIFT Flow dataset: The SIFT Flow dataset provided by Liu *et al*. [2] consists of 2,688 images of outdoor scenes. The dataset provides ground truth labels hand-annotated by LabelMe users. Liu *et al*. [2] split this dataset into 2,488 training images and 200 test images, and selected top 33 object categories as semantic labels. For comparison, the same training/test split is used as [2, 4]. To obtain ψ_i terms, we employ nonparametric



(a) (b)

Region feature	Rate (%)
SIFT	75.27 (28.29)
+ Texture	75.62 (29.28)
+ Location	76.49 (30.65)
+ Shape	76.82 (30.68)
+ Appearance	76.80 (30.97)
+ Color	77.14 (32.29)

(c)

Figure 3.8: (a): Recognition rate as a function of the number of the retrieved images T and the influence of our model λ. (b): Recognition rate as a function of the number of the retrieved images T and the k of the visual similarity graph. (c): Feature evaluation on the SIFT Flow dataset.

nearest-neighbor classifiers [4, 57]. As shown in Table 3.2, our system achieves an overall pixel-level accuracy of 77.14% and a per-class accuracy of 32.29%. Figure 3.7 (a) shows that our per-class rate on the SIFT Flow dataset is significantly better than that of the baseline MRF.

Next, we validate our system by varying the parameters including the number of retrieved images T, the feature combination, k of k-nearest neighbor, and the influence

Table 3.3: Average computation time in second.

	Jain *et al*. Dataset	SIFT Flow Dataset
Image size	640×480 (few exception)	256×256
Average N (# of regions)	4243	1005
Average K (# of object class)	15	11
	Time (second)	
Graph Construction	23.92	4.22
Context Link Prediction	155.51	18.09
Inference	8.83	0.79

of context scores λ. First, we fix $k = 10$, use all features, and plot the recognition rate as a function of T in Figure 3.8 (a) with different λ. The recognition rate increases as more retrieved images are used. However, the recognition rate slightly drops continue to add retrieved images. Additionally, it is observed that strongly enforcing contextual consistency increases ambiguities and degenerates the performance. The maximal performance is achieved when $T = 16$ and $\lambda = 1$. Second, we fix $\lambda = 1$, use all features, and plot the recognition rate as a function of T in Figure 3.8 (b) with different k. Clearly, appropriate number of retrieved images is needed to achieve accurate context consistency. The maximal performance is achieved when $T = 16$ and $k = 10$. Finally, Figure 3.8 (c) shows recognition rates with region features added consecutively. Notice that it is arranged in order of increasing per-class rate and the SIFT histogram is the strongest feature in our system similar to the result of [4].

The computation time of our algorithm is shown in Table 3.3. All experiments were run on a standard PC with 3.0 GHz Intel quadcore CPU and 8 GB RAM. The

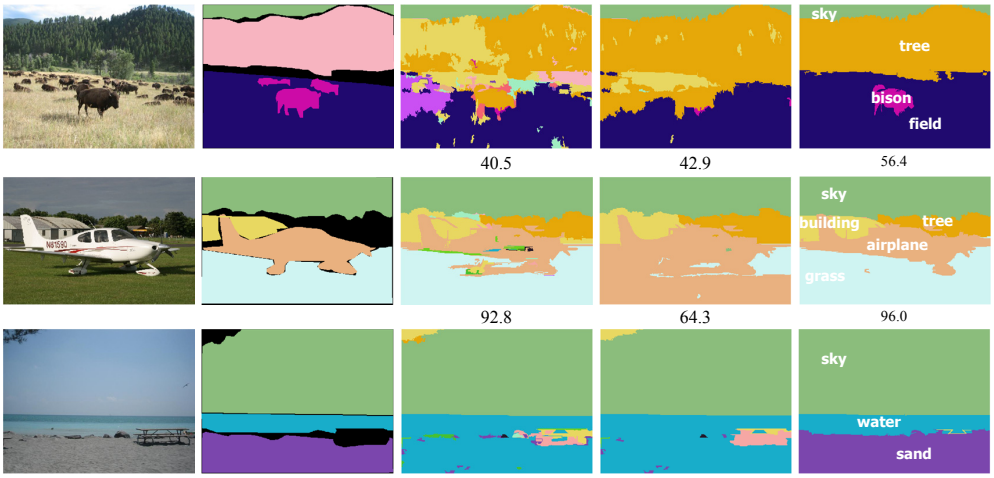
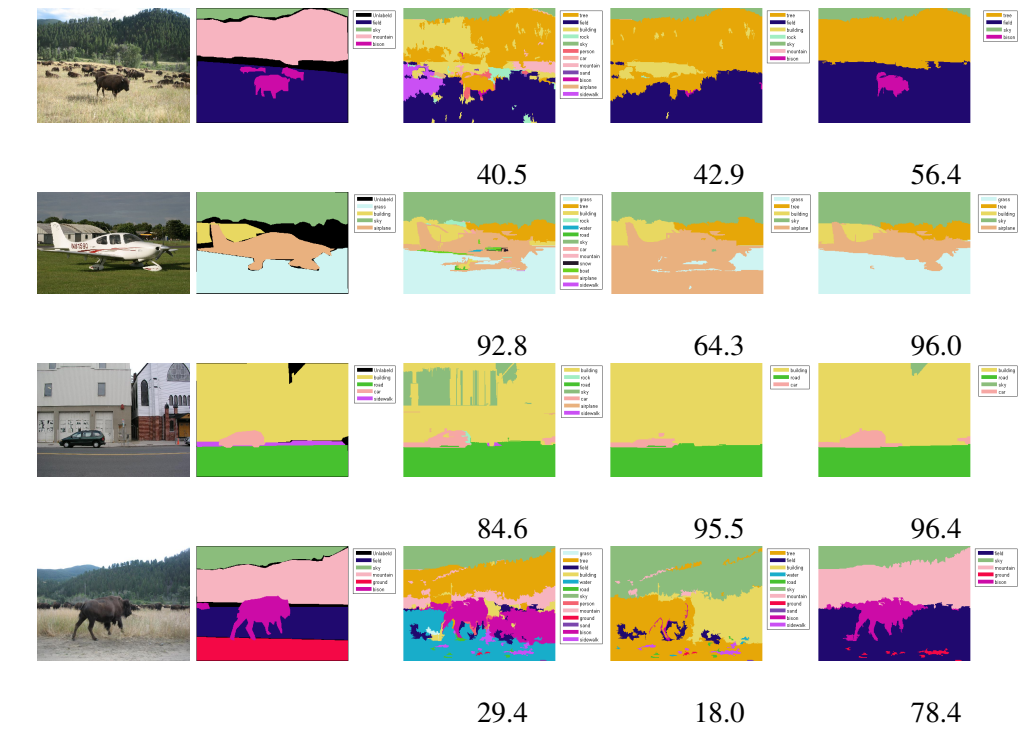


Figure 3.9: Example results from Jain *et al*. dataset. Column (a) shows the query image to be labeled and Column (b) represents the ground truth of (a). Column (c), (d), (e) shows the prediction of the baseline classifier, SuperParsing [4], and our approach with unary potential, respectively. The numbers under each image indicates pixel-wise accuracy on that image.

proposed method was implemented as a MATLAB code without any parallelization efforts. For both datasets, we fixed our parameters to $T = 16, k = 10, \lambda = 1$. It means that total $T + 1 = 17$ images are used to construct a similarity graph. Since our algorithm requires $O(K^2N^2)$ times, increasing K and N makes our algorithm significantly slow.



Original image Groundtruth Baseline (ψ_i only) SuperParsing Our approach

Figure 3.10: Scene labeling results on Jain *et al*. Dataset against SuperParsing and our approach.

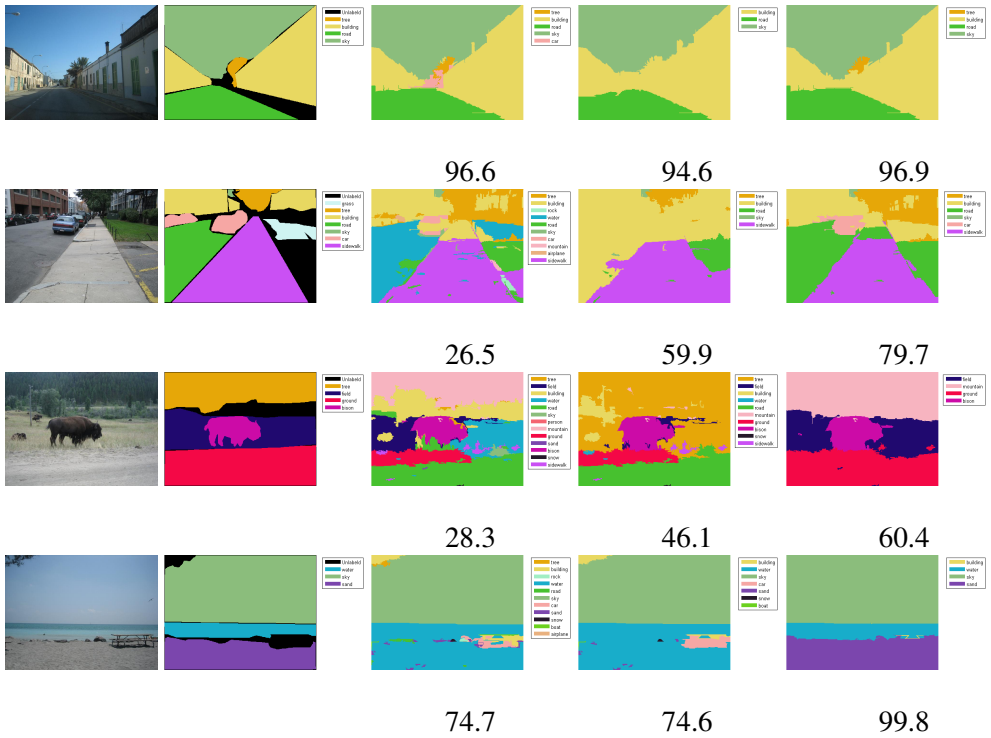


Figure 3.11: Scene labeling results on Jain *et al*. Dataset against SuperParsing and our approach.

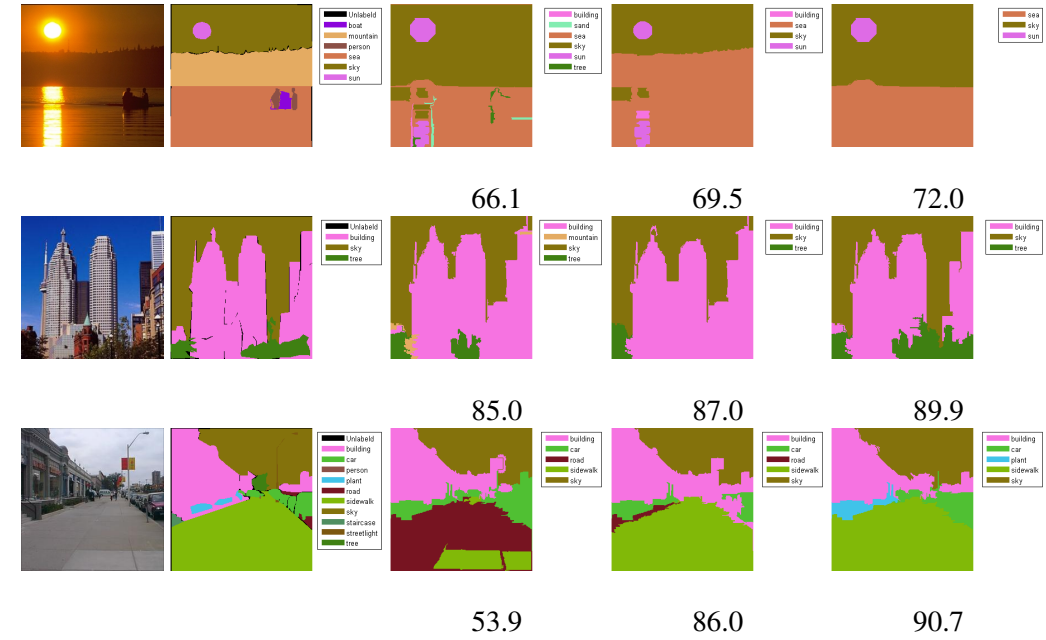


Figure 3.12: Scene labeling results on SIFT Flow Dataset against SuperParsing and our approach.

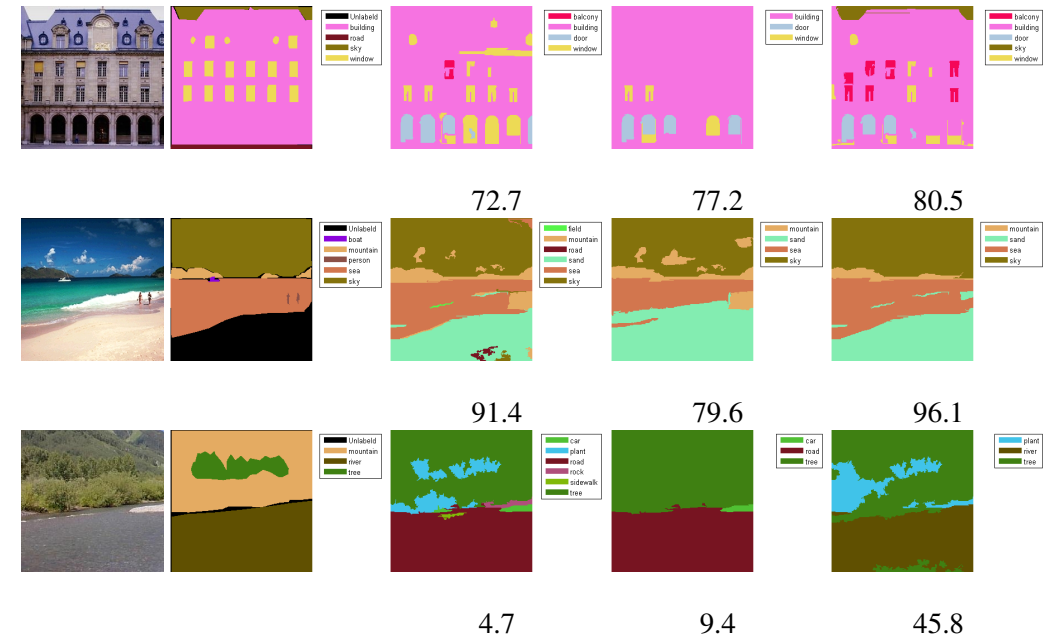


Figure 3.13: Scene labeling results on SIFT Flow Dataset against SuperParsing and our approach.

3.6 Summary

We have presented a nonparametric exemplar-based context model in which object relationships are explicitly captured. A graph-based context representation is proposed to efficiently transfer contextual relationships from training images to a query image. This allows jointly modeling visual appearance and context. Our novel approach helps to overcome the limitation of conventional context models relying on object label agreement and gives richer appearance-based context information. Moreover, the learned object relationships can be incorporated into any region-based scene labeling approaches as an additional cue. One of the main limitations of our model is that it considers all relations between regions as equally important. Clearly, there might be implausible or unimportant context exemplars, but our model cannot eliminate them. Our future work is to overcome this problem and extend our system to the multiple segmentation framework.

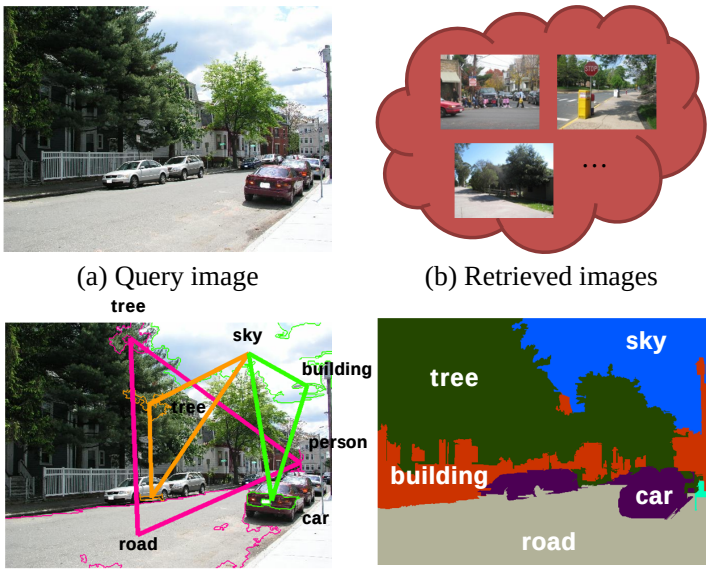
Chapter 4

High-order Semantic Relationships

4.1 Introduction

Recently, with the increasing availability of large image collections of hand-labeled images, nonparametric label transfer approaches for this problem have attracted many computer vision researchers and shows very good performance [65, 71, 2, 4, 57, 72, 73]. Compared to conventional parametric semantic segmentation methods [74, 23, 55, 75], these approaches do not need training model parameters, hence, they can be scalable to large datasets with an unknown number of object categories. Typical label transfer approaches start by retrieving similar images for a given test image. After that, they establish dense correspondence between two images and then warp labels from the matched annotated images to the test image. In spite of good performances, these approaches sometimes produce unsatisfactory results because they do not explore high-level contextual knowledge within the annotated images. Obviously, high-level semantic relationships between objects within the annotated image are very important cues to successful semantic segmentation.

To this end, recent approaches have advocated the use of nonparametric context



(a) Query image (b) Retrieved images

(c) Estimated high-order relation (d) Semantic segmentation

Figure 4.1: For a query image (a), our system finds the matched similar images (b) from a large dataset using global scene descriptors. The high-order semantic relations are transferred from the annotated images (b) to the query image (a). (We densely estimate high-order semantic relation across the image, but this figure displays only a few top scored relations for visualization purposes.) We then infer semantic segmentation (d) using estimated semantic relation (c).

models [1, 27]. These learn pairwise relationships between objects using global scene features and local features. However, these methods use only pairwise relationships to model high-level semantic relationships. Since natural images typically contain more than three object categories, pairwise relations are not enough to represents high-level information within images.

In this chapter, we develop a novel nonparametric approach for semantic segmentation by incorporating high-order semantic relations. Specifically, similar to several

label transfer methods [71, 2, 4, 72], we first find a set of small retrieved images from training images. Our goal is to transfer high-order semantic relations of annotated objects from each matched image to the query image. Since it is not feasible to obtain dense pixel-wise high-order semantic relations, we utilize "superpixel" regions obtained by oversegmentation of the query image. We define semantic tensors to represent the higher-order semantic relations of regions. We approach the problem of transferring the high-order semantic relations by defining a quadratic objective function of the semantic tensors. To optimize our objective function, we develop an efficient approximate algorithm based on Kronecker sum similarity and low-rank property of semantic tensors. To integrate our predicted semantic tensor into a semantic segmentation system, a fully connected Markov random field optimization is employed.

The key contributions of this chapter include: (1) The use of high-order semantic relations for semantic segmentation; (2) A novel tensor-based representation of high-order semantic relations; and (3) A quadratic objective function for learning the semantic tensor and an efficient approximate algorithm.

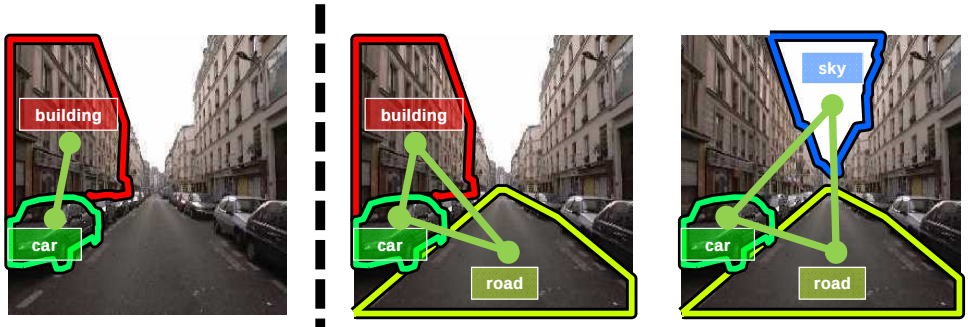
The chapter is organized as follows. We review some relevant works in Section 4.2. In Section 4.3, we introduce high-order semantic relation transfer algorithm and explain in detail. Section 4.3.3 presents a semantic segmentation method through semantic relation transfer. The experimental results are given in Section 4.5. Finally, in Section 4.6, we discuss our approach.

4.2 Related work

We now review related works on label transfer approaches and nonparametric context models. The problem of label transfer was first addressed recently by Liu *et al*. [2].

They first retrieved similar images using GIST matching [62] and constructed pixel-wise dense correspondence between each retrieved image and test image using SIFT flow [76]. They then transferred the annotations based on dense correspondence and reasoned semantic segmentation. Following the idea of label transfer [2], Zhang *et al*. [72] employed partial matching between the test image and the retrieved images to use partial similarity between images. Gould and Zhang [77] constructed Patch-MatchGraph to reduce complexity of retrieval step. Chen *et al*. [71] proposed supervised geodesic propagation to guide label transfer. Tighe and Lazebnik [4, 57] considered superpixel-level matching to transfer label information. However, all of these approaches are restricted to transferring label information from matched images. Although Liu *et al*. [2] claimed that the label transfer approach naturally embeds contextual information in the retrieval/alignment procedure, it is hard to tell how much contextual knowledge will help or what the effects will be.

On the other hand, recent nonparametric context models [1, 27] for semantic segmentation employed contextual relationships between objects to achieve more accurate results. Jain *et al*. [1] learned which contextual relationships should be considered and predicted features weight for each relation in a nonparametric manner. Myeong *et al*. [27] formulated a data-driven context learning problem as a graph-based context link prediction problem. Since our semantic tensor can be viewed as a generalization of the context link [27], their work is most similar to our own. However, there are several important differences with respect to our work. First, they only considered pairwise object relationships. On the contrary, our method focuses on high-order (mostly third-order) semantic relations, allowing us to model complex contextual relationships. For example, triplet-wise semantic relations can be found such as (*sky*,*car*,*road*) by our method as illustrated in Figure 4.2. These relations become important when consid-



(a) Pairwise semantic relation (b) Third-order semantic relations

Figure 4.2: An example of pairwise and high-order semantic relations. The third-order semantic relations (b) can model complicated high-level semantic knowledges within an image compared with the pairwise semantic relation (a).

ering complicated scenes with many object classes. Second, we develop a quadratic objective function for the high-order semantic relation transfer problem. However, Myeong *et al*. [27] did not show how their context link prediction works mathematically.

High-order models are not well studied in the context of semantic segmentation. Kohli *et al*. [78] introduced high-order model to enforce label consistency among regions. However, their high-order model is not related to high-level semantic knowledge. To our knowledge, there are no prior works explicitly considering high-order contextual relationships between objects in the literature on semantic segmentation.

4.3 The high-order semantic relation transfer algorithm

4.3.1 Problem statement

We consider two images I^1 and I^2; the first one is not annotated whereas the second one is densely labeled with the corresponding object class. We assume that two images are closely-related in which the similar objects are present and that objects roughly maintain their high-order relation. We define high-order semantic relation transfer problem as a task to predict high-order relation between unlabeled regions in I^1 based on annotated regions in I^2. For simplicity, we will focus on third-order relations from now.

Let $\mathcal{S} = \{S^1, S^2\}$ be a set of superpixels generated by segmenting the respective images. n^1 and n^2 is the number of segments in S^1 and S^2, respectively, and $N = n^1 + n^2$ is the total number of segments. $\mathcal{C} = \{c_1, c_2, ..., c_K\}$ is a given set of object classes. Third-order semantic relations among region triplets $(s_i, s_j, s_k) \in \mathcal{S} \times \mathcal{S} \times \mathcal{S}$ is defined as a set of $N \times N \times N$ third-order tensors $\mathbb{X} = \{\mathcal{X}^{111}, \mathcal{X}^{112}, \mathcal{X}^{113}, ..., \mathcal{X}^{KKK}\}$. We refer to each tensor $\mathcal{X}^{\alpha\beta\gamma} \in \mathbb{X}$ as a *semantic tensor*. A semantic tensor $\mathcal{X}^{\alpha\beta\gamma}$ denotes third-order semantic relations among region triplets on object class triplet $(c_\alpha, c_\beta, c_\gamma)$. Each element of $\mathcal{X}^{\alpha\beta\gamma}$ is defined as

$$[\mathcal{X}^{\alpha\beta\gamma}]_{ijk} = x_{ijk}^{\alpha\beta\gamma}. \tag{4.1}$$

The variable $x_{ijk}^{\alpha\beta\gamma}$ indicates confidence score of how likely the region triplet (s_i, s_j, s_k) would be labeled as $(c_\alpha, c_\beta, c_\gamma)$, respectively. $x_{ijk}^{\alpha\beta\gamma}$ is close to 1 if the assigned object class triplet $(c_\alpha, c_\beta, c_\gamma)$ is reliable. On the other hand, $x_{ijk}^{\alpha\beta\gamma}$ is close to 0 if the assigned object class triplet $(c_\alpha, c_\beta, c_\gamma)$ is unreliable.

Next, we define another set of $N \times N \times N$ tensor representing the observed third-order semantic relations within the image I^2. Similar to $\mathbb{X}$, we define $\mathbb{Y} =$

$\{\mathcal{Y}^{111}, \mathcal{Y}^{112}, \mathcal{Y}^{113}, ..., \mathcal{Y}^{KKK}\}$, and represent each element of $\mathcal{Y}^{\alpha\beta\gamma}$ as

$$y_{ijk}^{\alpha\beta\gamma} = \begin{cases} 1 & \text{if } G(s_i) = c_\alpha, G(s_j) = c_\beta, G(s_k) = c_\gamma, \\ & (s_i, s_j, s_k) \in S^2 \\ 0 & \text{otherwise} \end{cases}, \tag{4.2}$$

where $G(s_i)$ denotes the ground truth class of region s_i and $(s_i, s_j, s_k) \in S^2$ indicates that three regions s_i, s_j, and s_k are from the same image I^2. Since there are no semantic relations within S^1 and across images, all $y_{ijk}^{\alpha\beta\gamma}$ is 0 for $(s_i, s_j, s_k) \notin S^2$. In practice, each $\mathcal{Y}^{\alpha\beta\gamma}$ can be compactly generated from label vectors. Let $\mathbf{y}^\alpha$ be a column vector of length N, where $[\mathbf{y}^\alpha]_i = y_i^\alpha$ is 1 if region s_i belongs to object class c_α; and 0 otherwise. Then each element of $\mathcal{Y}^{\alpha\beta\gamma}$ can be generated by

$$y_{ijk}^{\alpha\beta\gamma} = y_i^\alpha y_j^\beta y_k^\gamma. \tag{4.3}$$

Eq. (4.3) can be rewritten as

$$\mathcal{Y}^{\alpha\beta\gamma} = \mathbf{y}^\alpha \circ \mathbf{y}^\beta \circ \mathbf{y}^\gamma. \tag{4.4}$$

The symbol "$\circ$" denotes the vector outer product. Since $\mathcal{Y}^{\alpha\beta\gamma}$ can be represented as the outer product of three vectors, $\mathcal{Y}^{\alpha\beta\gamma}$ is a *rank-one* tensor [79]. This rank-one property of $\mathbb{Y}$ is one of key aspects to approximate the following objective function.

4.3.2 Objective function

Now, the third-order semantic relation transfer problem can be regarded as the problem of estimating the magnitude of confidence scores $x_{ijk}^{\alpha\beta\gamma}$ for all superpixel triplets (s_i, s_j, s_k) and for all object class triplets $(c_\alpha, c_\beta, c_\gamma)$ based on $\mathbb{Y}$. We assume that there is no interaction between the semantic tensors. Hence, we separately deal with the third-order semantic relations transfer problem with respect to $\mathcal{Y}^{\alpha\beta\gamma}$. For simplicity, we drop the $\alpha\beta\gamma$ suffix from now.

Following the idea of link propagation [56], we want to enforce that two similar region triplets are likely to have the same confidence score. Thus, we design the quadratic objective function with respect to $\mathcal{Y}$ as

$$F(\mathcal{X}) = \frac{1}{2} \sum_{i,j,k,l,m,n}^{N} w_{ijk,lmn}(x_{ijk} - x_{lmn})^2 + \lambda \sum_{i,j,k}^{N} (x_{ijk} - y_{ijk})^2, \tag{4.5}$$

where $w_{ijk,lmn}$ is the triplet-wise similarity between two region triplets (s_i, s_j, s_k) and (s_l, s_m, s_n) and $\lambda > 0$ is the regularization parameter. The first term of Eq. (4.5) is the continuity constraint that two triplets should have the same confidence score if two triplets are similar. The second term is the unary constraint that each region triplet x_{ijk} tends to have their target values y_{ijk}. The cost function defined as pairwise and unary terms is a generalization of the cost function for label propagation [60].

Now, we rewrite Eq. (4.5) using tensors. For that, let $\mathbf{L}$ be an $N^3 \times N^3$ matrix called a *Laplacian matrix* defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{W}, \tag{4.6}$$

where $w_{ijk,lmn}$ is rewritten as similarity matrix $\mathbf{W}$ of size $N^3 \times N^3$ and $\mathbf{D}$ is a diagonal matrix whose diagonal elements are $[\mathbf{D}]_i = \sum_j^{N^3} [\mathbf{W}]_{ij}$. Using $\mathbf{L}$, Eq. (4.5) can be reformulated as

$$F(\mathcal{X}) = \frac{1}{2} \mathbf{vec}(\mathcal{X})^T \, \mathbf{L} \, \mathbf{vec}(\mathcal{X}) + \lambda (\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\mathcal{Y}))^2, \tag{4.7}$$

where $\mathbf{vec}(\mathcal{X})$ is the vector constructed by concatenating the mode-1 fibers of the tensor $\mathcal{X}$ [79].

Differentiating Eq. (4.7) with respect to $\mathbf{vec}(\mathcal{X})$, and set to 0, we can get $\mathcal{X}$ that minimizes Eq. (4.7),

$$\frac{\partial F(\mathcal{X})}{\partial \mathbf{vec}(\mathcal{X})} = \mathbf{L} \, \mathbf{vec}(\mathcal{X}) + \lambda \mathbf{vec}(\mathcal{X}) - \lambda \mathbf{vec}(\mathcal{Y}) = 0 \tag{4.8}$$

It can be transformed into

$$(\mathbf{L} + \lambda\mathbf{I})\mathbf{vec}(\mathcal{X}) = \lambda\mathbf{vec}(\mathcal{Y}), \tag{4.9}$$

where $\mathbf{I}$ indicates identity matrix of size $N^3 \times N^3$. Since $\mathbf{L}+\lambda\mathbf{I}$ is positive definite, the linear equation (4.9) can be solved by matrix inversion. However, computing inverse matrix of size $N^3 \times N^3$ is not realistic in practice.

4.3.3 Approximate algorithm

In this section, we present an efficient optimization scheme for the proposed objective function. Since providing all of the N^6 elements of the triplet-wise similarity matrix $\mathbf{W}$ is intractable, we consider constructing $\mathbf{W}$ using the segments-wise similarity matrix $\mathbf{W}_S$ the same as [56]. As described in Section 4.5, $\mathbf{W}_S$ is defined as similarity between two superpixels. Recommended by [56], we define $\mathbf{W}$ based on *Kronecker sum similarity*. Hence, $\mathbf{L}$ can be re-represented as

$$\mathbf{L} = \mathbf{L}_S \oplus \mathbf{L}_S \oplus \mathbf{L}_S, \tag{4.10}$$

where $\oplus$ indicates the Kronecker sum and $\mathbf{L}_S$ is defined as $\mathbf{L}_S = \mathbf{D}_S - \mathbf{W}_S$ and $\mathbf{D}_S$ is a diagonal matrix whose diagonal elements are $[\mathbf{D}_S]_i = \sum_j^N [\mathbf{W}_S]_{ij}$. Using Eq. (4.10), the objective function (4.5) can be expressed as

$$F(\mathcal{X}) = \frac{1}{2}\mathbf{vec}(\mathcal{X})^T\mathbf{vec}(\mathcal{X} \times_1 \mathbf{L}_S + \mathcal{X} \times_2 \mathbf{L}_S + \mathcal{X} \times_3 \mathbf{L}_S) + \lambda(\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\mathcal{Y}))^2, \tag{4.11}$$

where $\times_n$ represents n-mode product of tensor [79]. Inspired by [80, 3], we approximate the objective function in three optimization steps:

$$\dot{\mathcal{X}} = \arg\min_{\mathcal{X}} \frac{1}{2}\mathbf{vec}(\mathcal{X})^T\mathbf{vec}(\mathcal{X} \times_1 \mathbf{L}_S) + \lambda(\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\mathcal{Y}))^2 \tag{4.12}$$

$$\ddot{\mathcal{X}} = \arg\min_{\mathcal{X}} \frac{1}{2}\mathbf{vec}(\mathcal{X})^T\mathbf{vec}(\mathcal{X} \times_2 \mathbf{L}_S) + \lambda(\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\dot{\mathcal{X}}))^2 \tag{4.13}$$

$$\hat{\mathcal{X}} = \arg\min_{\mathcal{X}} \frac{1}{2}\mathbf{vec}(\mathcal{X})^T\mathbf{vec}(\mathcal{X} \times_3 \mathbf{L}_S) + \lambda(\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\ddot{\mathcal{X}}))^2. \tag{4.14}$$

That is, we sequentially estimate the semantic tensor for each mode product term. In a similar way to Eq. (4.9), we can obtain linear system equation for each optimization step.

$$\mathcal{X} \times_1 (\mathbf{L}_S + \lambda\mathbf{I}_S) = \lambda\mathcal{Y} \tag{4.15}$$

$$\mathcal{X} \times_2 (\mathbf{L}_S + \lambda\mathbf{I}_S) = \lambda\dot{\mathcal{X}} \tag{4.16}$$

$$\mathcal{X} \times_3 (\mathbf{L}_S + \lambda\mathbf{I}_S) = \lambda\ddot{\mathcal{X}}, \tag{4.17}$$

where $\mathbf{I}_S$ indicates identity matrix of size $N \times N$. For solving each linear equation, let us consider Eq. (4.15), 1-mode tensor product of Eq. (4.15) can be expressed in terms of unfolded tensors.

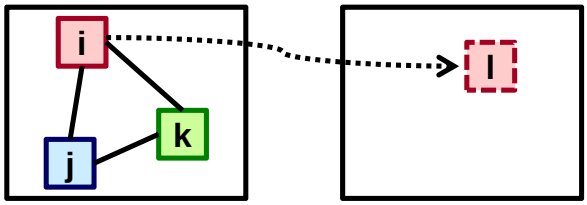
$$(\mathbf{L}_S + \lambda\mathbf{I}_S)\mathbf{X}_{(1)} = \lambda\mathbf{Y}_{(1)}, \tag{4.18}$$

where $\mathbf{X}_{(1)}$ denotes the mode 1 matricization of a tensor $\mathcal{X}$ (see [79] for more details). Remind that $\mathcal{Y}$ is rank-one, $\mathcal{Y}$ can be written as in matricized form [79],

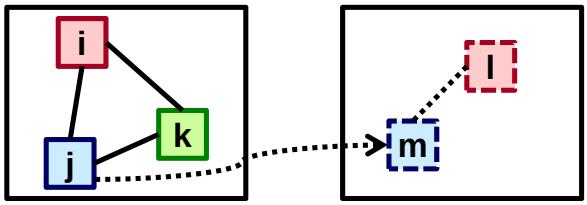
$$\mathbf{Y}_{(1)} = \mathbf{y}^{\alpha}(\mathbf{y}^{\gamma} \circ \mathbf{y}^{\beta})^T. \tag{4.19}$$

Hence, $\dot{\mathcal{X}}$ can be efficiently computed by

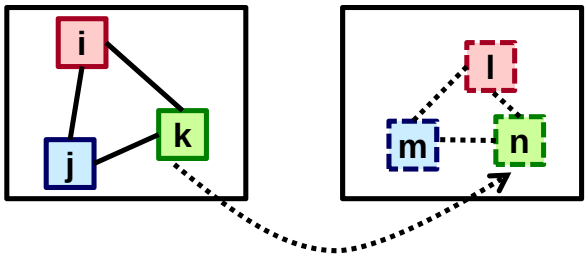
$$\dot{\mathbf{X}}_{(1)} = (\mathbf{L}_S + \lambda\mathbf{I}_S)^{-1}\lambda\mathbf{y}^{\alpha}(\mathbf{y}^{\gamma} \circ \mathbf{y}^{\beta})^T. \tag{4.20}$$



(a) 1-mode product



(b) 2-mode product



(c) 3-mode product

Figure 4.3: Illustration of the proposed approximate algorithm. The algorithm (b) first find similar region s_l with respect to s_i while fixing s_j and s_k, (c) then find similar region s_m with respect to s_j while fixing s_l and s_k, (d) and finally find similar region s_n with respect to s_k while fixing s_l and s_m.

We continue to solve for $\ddot{\mathcal{X}}$ and $\hat{\mathcal{X}}$ similarly. Then we can obtain the approximate solution of the objective function (4.5) as follows.

$$\hat{\mathcal{X}} = [(\mathbf{L}_S + \lambda \mathbf{I}_S)^{-1} \lambda \mathbf{y}^{\alpha}] \circ [(\mathbf{L}_S + \lambda \mathbf{I}_S)^{-1} \lambda \mathbf{y}^{\beta}] \circ [(\mathbf{L}_S + \lambda \mathbf{I}_S)^{-1} \lambda \mathbf{y}^{\gamma}]. \quad (4.21)$$

Note that $\hat{\mathcal{X}}$ also can be represented as the outer product of three vectors, $\hat{\mathcal{X}}$ is a *rank-one* tensor. In Figure 4.3, this procedure summarizes schematically. We independently

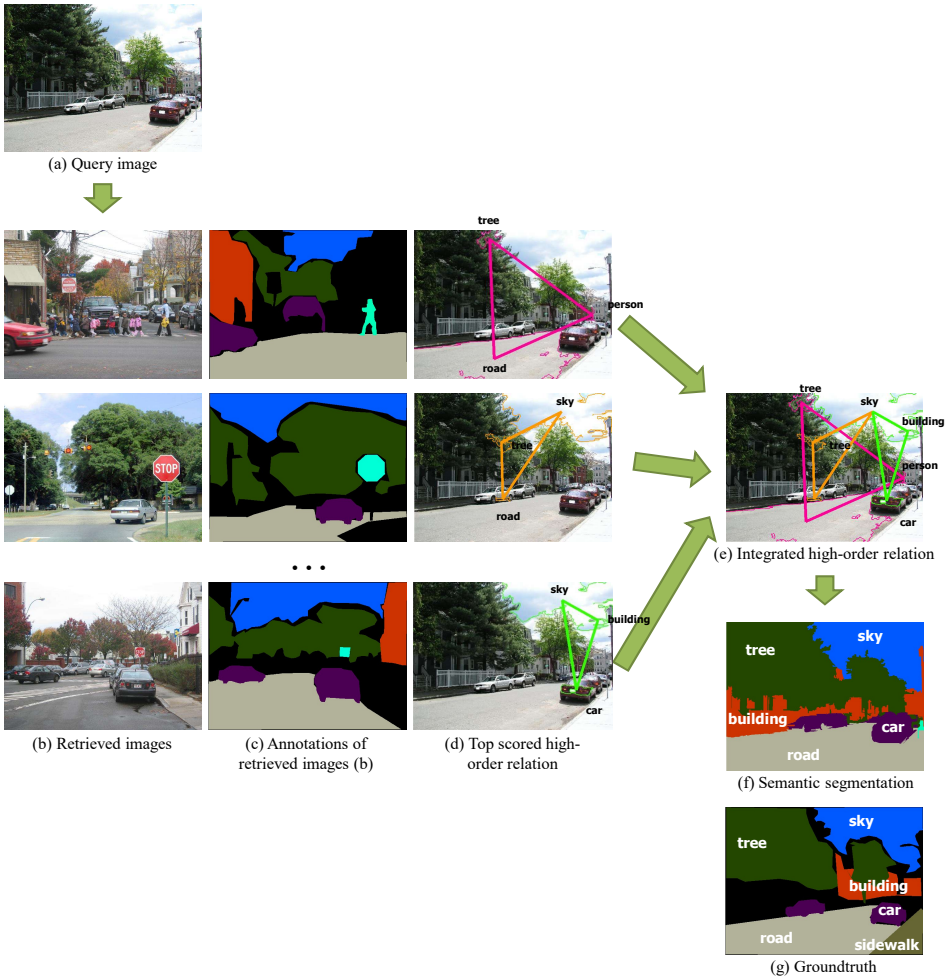


Figure 4.4: System overview. For a query image (a), we first retrieve the matched similar scenes (b). We predict the third-order semantic relations (d) by transferring semantic relations from each annotated image (c) to the query image (a). We aggregate semantic relations (e) from multiple semantic relation candidates (d) and generate semantic segmentation (f). (g) is the ground-truth annotation of (a).

transfer each $\mathcal{Y}^{\alpha\beta\gamma}$, hence, this procedure repeats K^3 times. Finally, we can get the set of the predicted semantic tensors $\hat{\mathbb{X}} = \{\hat{\mathcal{X}}^{111}, \hat{\mathcal{X}}^{112}, \hat{\mathcal{X}}^{113}, ..., \hat{\mathcal{X}}^{KKK}\}$.

4.4 Semantic segmentation through semantic relation transfer

Now that we have the semantic relation transfer algorithm from annotated images to unlabeled images, we can infer semantic segmentation using estimated semantic tensors.

4.4.1 Scene retrieval

Recall that we assume that each pair of images I^1 and I^2 roughly agree on the spatial layout of objects. Hence, it is essential to to extract closely-related images from large dataset with respect to a query image for successful semantic relation transfer. Unreliable semantic tensors can be predicted between two unrelated images. To find similar images, we first retrieve M candidate images by color histogram, GIST matching [62], and spatial pyramid [61] from the training dataset. This candidate image set will be used to transfer its high-order semantic relations into the query image.

4.4.2 Inference

After performing the scene retrieval in section 4.4.1, we transfer high-order semantic relations from each candidate image to the query image and obtain multiple sets of predicted semantic tensors $\{\mathbb{X}\}_{u=1:M}$. Our goal is to assign object class for each region in the query image. To integrate the sets of predicted semantic tensors with a conventional unary and pairwise potential, we build high-order fully connected Markov

random field model. The energy function is defined as

$$E(\{l_i\}) = \sum_{i}^{n^1} E^D(l_i) + \sum_{(i,j)\in E} E^P(l_i, l_j) + \sum_{i,j,k}^{n^1} E^H(l_i, l_j, l_k), \tag{4.22}$$

where $l_i \in \{1, ..., K\}$ is the index of object class for region s_i. Since we want to label the regions in the query image, the energy function is only defined on the regions of image I^1. The first term is data term which represents the negative logarithm of the probability of class l_i given the region s_i. The second term is smoothness term which encourage two neighboring regions to have the same label. These two terms are typically used to conventional nonparametric scene parsing approaches [2, 4, 57].

Table 4.1: Performance comparison of our algorithm on the three challenging datasets. Per-pixel recognition rates and average per-class recognition rates in parentheses are presented.

	Jain *et al.* [1]	LMO [2]	Polo [72]
Jain *et al*. [1]	59.0 (-)	-	-
Liu *et al*. [2]	-	74.8 (-)	-
Tighe and Lazebnik [4]	-	76.8 (29.4)	87.9 (76.1) [72]
Zhang and Quan [72]	-	-	89.8 (82.5)
Chen *et al*. [65]	75.6 (45)	-	-
Myeong *et al*. [27]	80.1 (53.3)	**77.1 (32.3)**	-
Gould and Zhang [77]	-	-	**94.2 (91.7)**
Proposed (max)	81.5 (51.2)	76.1 (28.9)	89.1 (80.6)
Proposed (sum)	**81.8** (**54.4**)	76.2 (29.6)	88.3 (79.3)

However, it is nontrivial how to integrate the sets of predicted semantic tensors to

semantic segmentation framework. Hence we develop two third-order clique potential E^H_{max} and E^H_{sum}. The first high-order potential E^H_{max} take the form

$$E^H_{max}(l_i = c_\alpha, l_j = c_\beta, l_k = c_\gamma) = -\log(\max_u\{\hat{x}^{\alpha\beta\gamma}_{ijk}\}_u). \tag{4.23}$$

The first clique potential E^H_{max} take maximum confidence score among M number of candidate scores for region triplet (s_i, s_j, s_k) and for object triplet $(c_\alpha, c_\beta, c_\gamma)$. This means that we only consider the strongest one from the set of relation candidates. The second high-order potential E^H_{sum} have the form

$$E^H_{sum}(l_i = c_\alpha, l_j = c_\beta, l_k = c_\gamma) = -\log(\sum_u^M\{\hat{x}^{\alpha\beta\gamma}_{ijk}\}_u). \tag{4.24}$$

Meanwhile, the second clique potential E^H_{max} takes summation of M number of confidence scores. This potential picks average scores from the set of relation candidates. These two potential will be examined in the experimental section.

It is very important to effectively minimize the energy function (4.22), but efficient order reduction techniques such as [81] cannot be used due to space and time complexity. Hence, we apply multistart simulated annealing algorithm.

4.5 Experiements

In this section, we (1) evaluate our method's semantic segmentation performance and compare against pairwise semantic segmentation [27] and (2) analyze integration of our predicted semantic tensors. Now, we validate our approach with three challenging datasets: the dataset of Jain *et al*. [1], LabelMe Outdoor (LMO) dataset [2], and Polo dataset [72]. We evaluate on all sets, but focus additional analysis on the LMO dataset since it has the largest number of categories. Table 4.1 summarizes our semantic segmentation accuracy compared with the state-of-the-art methods. Proposed

(max) indicates the accuracy of the semantic segmentation with the max high-order term Eq. (4.23). Proposed (sum) represents performance with the sum high-order term Eq. (4.24).

Implementation details. Our implementation is based on the framework of Tighe and Lazebnik [4, 57]. We use the algorithm of Felzenszwalb and Huttenlocher [64] for segmentation, and fix the parameters $\sigma = 0.8, K = 200, min = 100$ on all sets. To form superpixel-wise weight $\mathbf{W}_S$, we use several types of descriptors $a_k(s_i)$ for regions s_i: shape, texture, color, and appearance from [4]. Along with appearance features, we integrate geometric position $g(s_i)$ (row+column) of the center of the region s_i. Hence, each elements of W_S are computed as

$$[\mathbf{W}_S]_{ij} = e^{-\sum_k \frac{\|a_k(s_i)-a_k(s_j)\|}{\sigma_{a_k}} - \frac{\|g(s_i)-g(s_j)\|}{\sigma_g}} \tag{4.25}$$

where $a_k(s_i)$ is the feature vector of the k-th type for s_i and σ_{a_k} denotes the standard deviation of a_k. Note that we densely obtain the weight between regions, it means that a region is connected to all the other regions with the corresponding weights. We fix the parameter of the objective function $\lambda = 10$. To compute E^D, we employ the nonparametric superpixel parsing [4] for the LMO dataset and the boosted decision tree classifier [25] for the other datasets. As a pairwise term E^P, we adopt simple Potts model.

Evaluation metric. We use both pixel-wise measure and class-wise measure to quantify the accuracy. The former rates total proportion of correctly labeled pixels, while the latter indicates the average proportion of correctly labeled pixels in each object class.

19-Class Jain *et al.* [1] dataset. Jain *et al*. [1] randomly collects 350 images of size 640×480 from LabelMe [70] with 19 classes. This dataset is splitted into 250 training

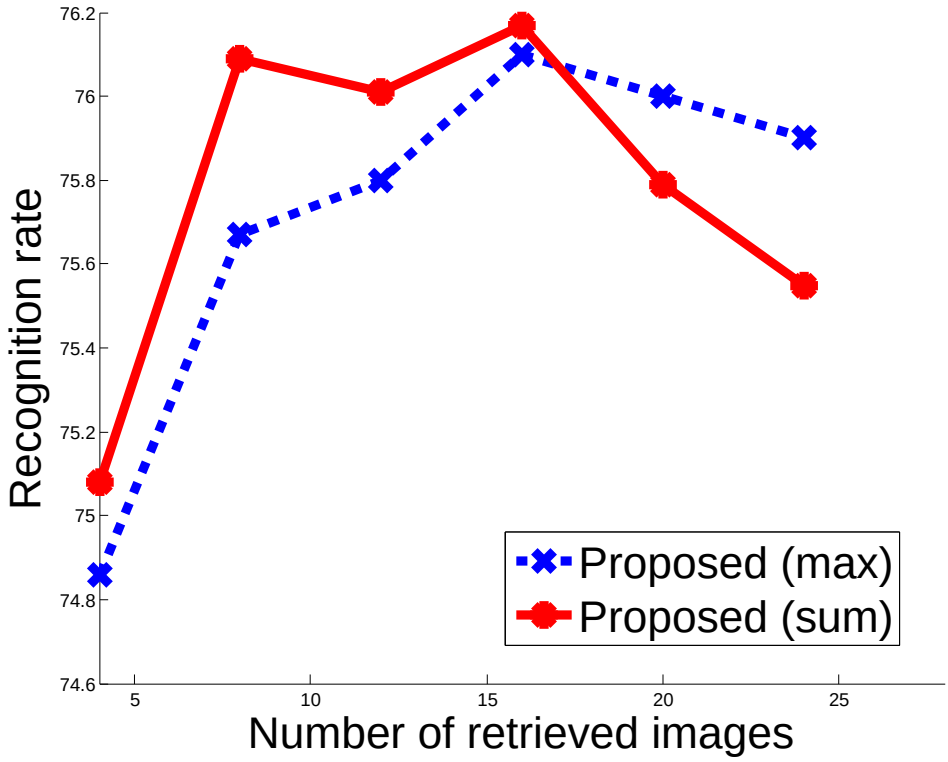


Figure 4.5: Recognition rate of two different high-order potential as a function of the number of the retrieved images M on the LMO dataset.

images and 100 test images. The number of similar images M is set to be 16. The semantic segmentation accuracy on this dataset is 81.8%.

This is relatively good dataset to evaluate high-order semantic relations. The size of the images is large enough and there are a lot of objects within an image. We achieve the state-of-the-art performance on this dataset and obtain promising results.

33-Class LabelMe Outdoor (LMO) dataset. This dataset provided by Liu *et al.* [2] contains total 2,688 images of size 256×256 from LabelME [70] with 33 object categories. Liu *et al.* [2] randomly split this dataset into 2,488 training images and 200 test

images. For qualitative comparison with [2, 27, 4], we use the same training/test split. We set the number of similar images M to 16. The semantic segmentation accuracy of the proposed method on this dataset is 76.2%.

Our results are below the state-of-the-art methods. We think that this is due to many images from this dataset with one or two object classes. The number of test images containing less than two object classes is 43 out of 200. It seems that complex contextual models such as the proposed method are not crucial to improve performance on this dataset.

6-Class Polo dataset. The polo dataset consists of 320 images from the web with keyword polo. Zhang *et al*. [72] annotated each image into six categories:*sky, horse, person, ground, tree, grass*. We set the number of similar images M to 20.

Our results are under the state-of-the-art methods. One reason is that context is not much important since all images have almost the same object classes. The other reason is the state-of-the-art method use complex pixel-wise model, on the other hand, we works on relatively simple region level.

Max vs. Sum. We design two different high-order potential for incorporating the set of the predicted semantic tensors. As shown in Figure 4.5, sum potential, taking summarization of candidates confidence scores, provides more better semantic segmentation results at some point. On the other hand, max potential, taking maximum of candidates confidence scores, is more robust to the number of retrieved images M. As gradually adding retrieved images, wrong matched images become larger and the performance of sum potential decreases faster.

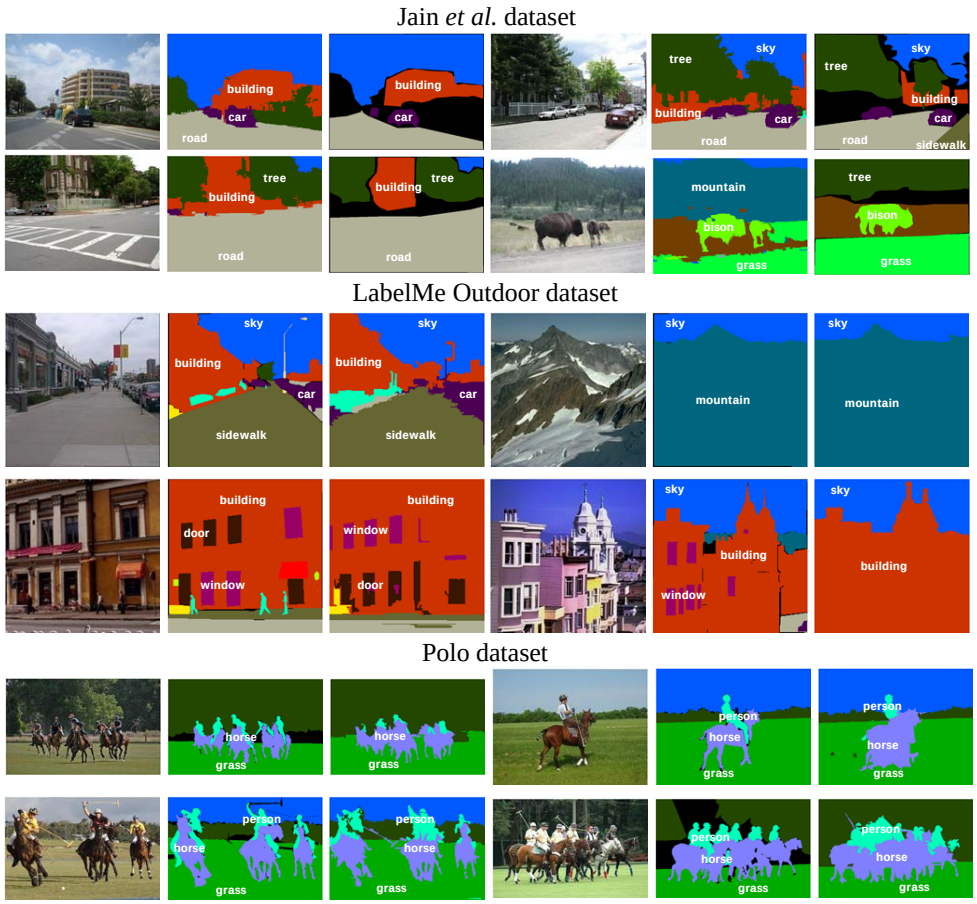


Figure 4.6: Example results from different datasets. The query images, ground truth, and results from our proposed (sum) are shown. Best viewed in color.

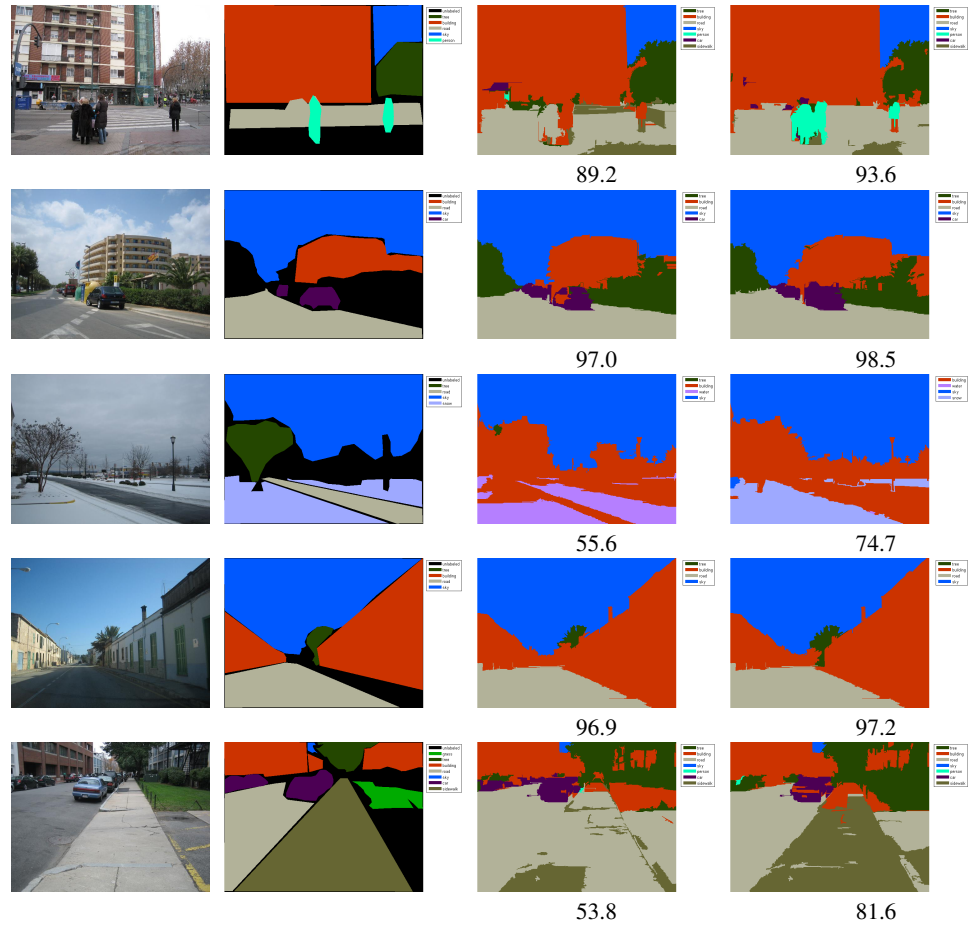


Figure 4.7: We report additional semantic segmentation results on Jain *et al*. Dataset against pairwise model of Myeong *et al*.

4.6 Summary

We have presented a novel approach to learn high-order semantic relations of regions in a nonparametric manner. We cast the high-order semantic relation transfer problem as a quadratic objective function of semantic tensors and propose an efficient approximate algorithm. We develop a novel semantic tensor representation of the high-order semantic relations. While we have presented this representation in the context of semantic segmentation, it can be applicable to various computer vision problem including object detection, scene classification, and total scene understanding.

Chapter 5

Multiscale CRF formulation

5.1 Introduction

Our final goal of this chapter is combining the context-aware appearance and appearance-aware context. In Chapter 2, we obtained the object class score map of each pixel with appearance model enhanced by semantic context. Meanwhile, we extracted the contextual relationships between two regions by utilizing the visual appearances in Chapter 3. Furthermore, we obtained the high-order contextual relationships between three regions similarly in Chapter 4. In each chapter, we generate the semantic segmentation results and compare them with the state-of-the-art methods. At first glance, each model seems to work independently. However, the result of the appearance model reinforced by the semantic context is represented by the probability of the pixels, and the semantic context information enhanced by the visual appearances is represented by the high-order potentials between the regions. In this chapter, we unify the two representations in a single optimization framework.

However, the challenge for this task is that the likelihood functions from adaptive context aggregation networks are pixel-wise, on the other hands, the prior probability

obtained from graph-based context model is region-wise. To address this problem, we propose the use of Multiscale Conditional Random Field (mCRF) to infer with cliques potential over regions and pixels.

5.2 Proposed Method

In this section, the proposed model for integrating the learned appearance in Chapter 2 and the learned context in Chapter 3, 4 will be discussed. Our goal is to simultaneously produce plausible multi-class labeling results $\mathcal{L}^p = \{l_n^p\}_{n=1,...,N}$ and $\mathcal{L}^s = \{l_m^s\}_{m=1,...,M}$ with the given a test image I with N pixels $X^p = \{x_n^p\}_{n=1,...,N}$ and M over-segmented regions $X^s = \{x_m^s\}_{m=1,...,M}$ at spatial scale s. The over-segmentation region X^s is generated by the graph-based segmentation algorithm [64].

In our Multiscale Conditional Random Field (mCRF), the conditional probability of the label field L^p, L^s given the observation X^p, X^s by combining conditional distributions that capture different statistical structure at pixels and regions is defined as follows, similar to [82, 83]:

$$P(L^p, L^s|X^p, X^s) \propto P(L^p|X^p)P(L^s|X^s), \tag{5.1}$$

where $P(L|X)$ is the posterior distributions of L over X. Estimating maximum a posterior (MAP) soltion is formulated as finding L^{p*}, L^{s*} that maximizes the following:

$$L^{p*}, L^{s*} = argmax_{L^p,L^s} P(L^p, L^s|X^p, X^s). \tag{5.2}$$

The posterior distribution $P(L^p, L^s|X^p, X^s)$ is a *Gibbs* distribution from the Hamersley Clifford thereom and can be rewritten as:

$$P(L^p, L^s|X^p, X^s) \propto exp(-\sum_{c\in C} \phi_c(X^p, X^s)), \tag{5.3}$$

where C represents all cliques and ϕ_c are the potential functions over the random variables $\{X^p, X^s\}$. We takes the negative logarithms to the probability function and the *Gibbs* energy can be written as:

$$E(L^p, L^s) \propto \sum_{c \in C} \phi_c(X^p, X^s). \tag{5.4}$$

The energy term associates low energy to the right values and high energy to the wrong values.

5.2.1 Multiscale Potentials

Our Multiscale Conditional Random Field (mCRF) are characterized by the following energy functions defined on the pixels and the regions.

A. For the pixel layer potentials,

$$E^p(X^p) = \sum_n E^p_{unary}(x^p_n) + \sum_n E^p_{pairwise}(x^p_n, x^p_{n'}) \tag{5.5}$$

$$+ \lambda \sum_n E^p_{region}(x^p_n, X^s) \tag{5.6}$$

where the clique potentials $E^p_{pairwise}(x^p_n, x^p_{n'})$ between two pixels x_n and $x_{n'}$ is defined based on the color difference as follows:

$$E^p_{pairwise}(x^p_n, x^p_{n'}) = \begin{cases} exp\left(-\frac{\|g_n - g_{n'}\|}{\sigma_g}\right) & n' \in \mathcal{N}_n \\ 0 & \text{otherwise} \end{cases}, \tag{5.7}$$

where g_n indicates the color value at pixel x^p_n in Lab color space and σ_g is a constant that controls the strength of the weight. It provides us with a numerical measure for the label similarity between two neighboring pixels. The set $\mathcal{N}_n$ is the neighboring pixels at pixel x_n, usually 4 neighborhoods. The unary potential $E^p_{unary}(x^p_n)$ is defined by the negative log likelihoods using deep context aggregation networks.

The term $E^p_{region}(x^p_n, X^s)$ in Equation 5.6 for only pixel labels is the higher-order region consistency by which a pixel labels should be similar to its corresponding region labels. Since the regions quite often contain pixels belonging to multiple labels, it partly enforces the label consistency of regions with a weight λ. In contrast to other segmentation algorithms which use the hard label consistency in regions on the assumption that all pixels constituting a particular region belong to the same label, our work uses this soft label consistency constraint by nonparametric learning from the test image.

B. For the region layer potentials,

$$E^s(X^s) = \sum_n E^s_{unary}(x^s_n) + \sum_n E^s_{pixels}(x^s_n, X^p) \tag{5.8}$$

$$+ \sum_n E^s_{pairwise}(x^s_n, x^s_{n'}) + \sum_n E^s_{third}(x^s_n, x^s_{n'}, x^s_{n''}) \tag{5.9}$$

where the weight $E^s_{pairwise}(x^s_n, x^s_{n'})$ between two regions x^s_n and $x^s_{n'}$ is defined based on the learned context scores in Chapter 3 and $E^s_{third}(x^s_n, x^s_{n'}, x^s_{n''})$ is obtained from the learned third-order context scores in Chapter 4.

Pairwise Terms:

The terms $E^p_{pairwise}(x^p_n, x^p_{n'})$ in (5.6) is the label-continuity pairwise constraints that two neighboring elements (pixels or regions) in the neighborhood system should have the same label if their colors are similar. On the other hand, the term $E^s_{pairwise}(x^s_n, x^s_{n'})$ in (5.9) is the high-order semantic context information extracted in Chapter 3.

Unary Terms:

The terms $E^p_{unary}(x^p_n)$ in (5.6) and $E^s_{unary}(x^s_n)$ in (5.9) are the unary potentials. For the initial pixel likelihoods, deep context memory networks in Chapter 2 is used. For

the initial region likelihoods, region-classifiers is used in Chapter 3 and 4.

The term $E^s_{pixels}(x^s_n, X^p)$ in (5.9) for only region likelihoods is another refined unary constraint whereby a region labels should be similar to the weighted average of inner pixel labels. It gives the effect of refining the region likelihoods from more informative pixel likelihoods, since it can not guarantee that the reliable regions are always extracted under complex region boundaries.

5.2.2 Non Convex Optimization

It is very important to effectively minimize the energy function in Equation (5.6) and (5.9), but efficient order reduction techniques such as [81] cannot be used due to space and time complexity. Hence, we apply multi-start simulated annealing algorithm for the-order relationships. Without the third-order term the inference can be achieved by the Quadratic Pseudo-Boolean Optimization (QPBO) algorithm [68, 69].

5.3 Experiments

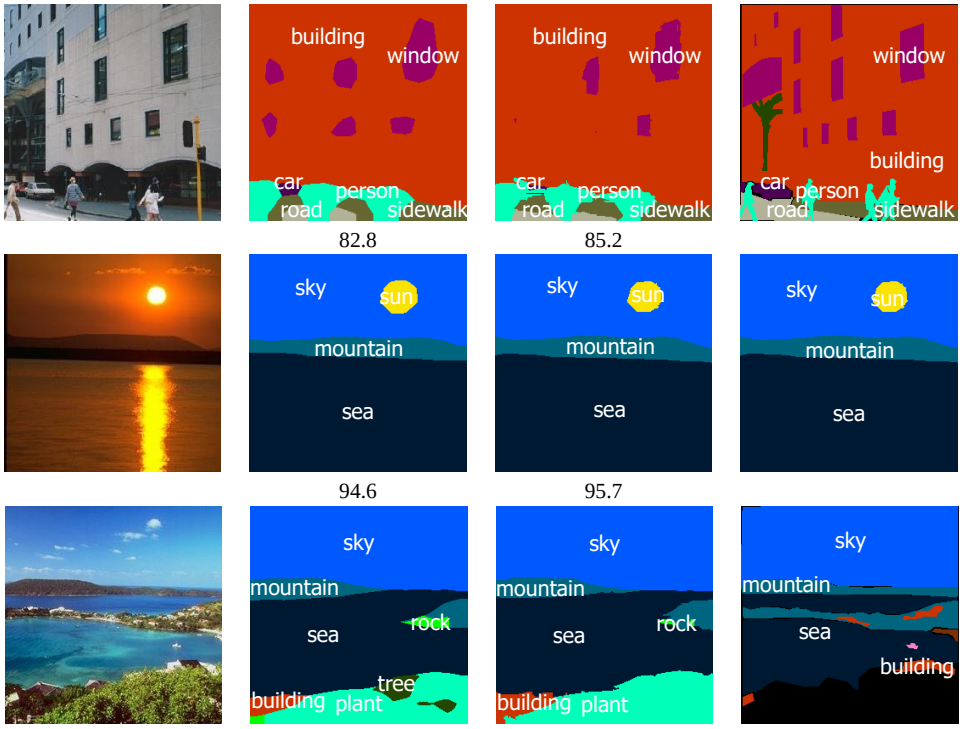
5.3.1 SiftFlow dataset

The SIFT Flow dataset provided by Liu *et al*. [2] consists of 2,688 images of outdoor scenes. The dataset provides ground truth labels hand-annotated by LabelMe users. Liu *et al*. [2] split this dataset into 2,488 training images and 200 test images, and selected top 33 object categories as semantic labels. For comparison, the same training/test split is used as [2, 4]. As shown in Table 5.1, our system achieves an overall pixel-level accuracy of 86.4% and a per-class accuracy of 56.4%. However, third-order information does not show the best performance due to the limited inference tools. For the second-order case, we use the QPBO optimization instead MCMC inference for

Table 5.1: Performance comparison of our algorithm on Sift Flow dataset.

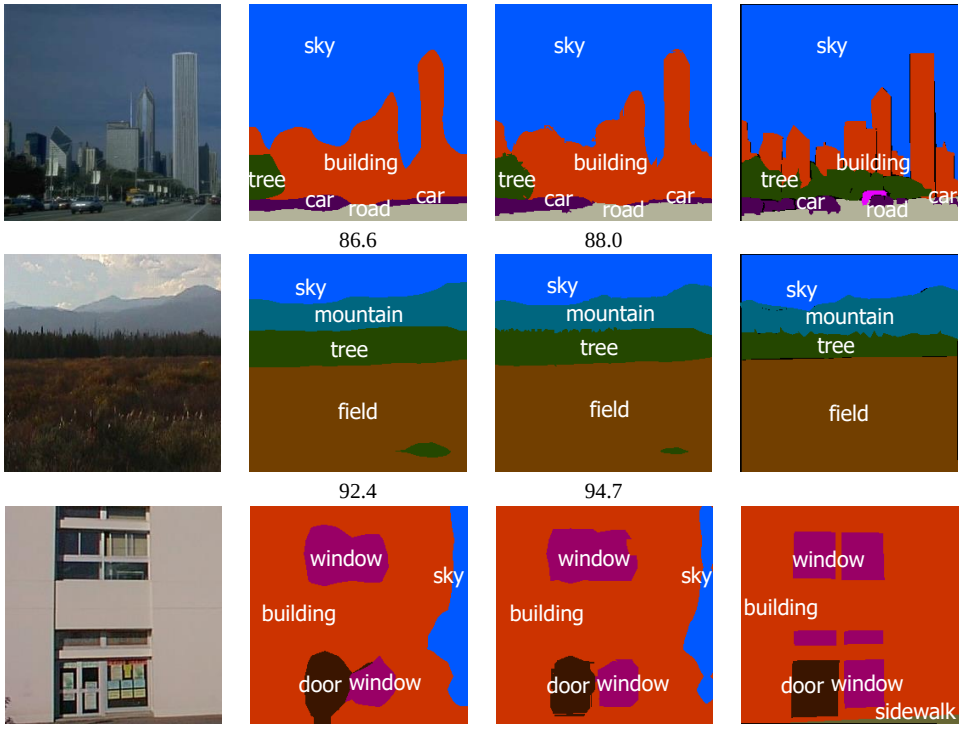
	Pixel accuracy	Mean class accuracy
Adaptive context agg.	85.8	56.0
Adaptive context agg. + second-order	**86.4**	**56.4**
Adaptive context agg. + high-order	86.1	56.2

third-order term.



(a) Input image (b) Adaptive context agg. (c) Adaptive context agg. + second-order (d) Ground truth

Figure 5.1: Representative results from the SIFT Flow dataset 1. (a) Input images. (b) The output of adaptive context aggreagation networks. (c) The output of adaptive context agregation networks with second-order context. (d) Ground truth. The number below the image shows pixelwise accuracy.



(a) Input image
(b) Adaptive context agg.
(c) Adaptive context agg. + second-order
(d) Ground truth

Figure 5.2: Representative results from the SIFT Flow dataset 2. (a) Input images. (b) The output of adaptive context aggreagation networks. (c) The output of adaptive context agregation networks with second-order context. (d) Ground truth. The number below the image shows pixelwise accuracy.

Chapter 6

Conclusion

6.1 Summary of the dissertation

In this dissertation, the semantic segmentation framework with jointly modeling visual appearance and semantic context have been presented. In particular, the problem of handling visual ambiguity in appearance modeling and enriching semantic context in context modeling is addressed. It is apparent that simultaneously learning appearance and semantic context plays an important role for accurate scene understanding, but the conventional approaches have rarely been applied to semantic segmentation frameworks due to the absence of the suitable representation model. The proposed methods in this dissertation make appropriately representations model in which appearance can be enhanced and enriched by context, and vice versa. To make an accurate final semantic segmentation result, context-aware appearance model and adaptive semantic context model are integrated into a Multiscale Conditional Random Fields framework.

Semantic segmentation system with adaptive context aggregation networks is presented in Chapter 2. Specifically, the use of external adaptive context aggregation have proposed and recursively refine the segmentation result. Considering that many com-

puter vision tasks can be interpreted as dense prediction tasks similar to semantic segmentation, the adaptive context aggregation networks is a general framework for many computer vision tasks. As recursive understanding of scene progress, we can observe that the how context helps the understanding of a scene. In Chapter 3 and 4, semantic context learning with appearance information have presented. Image-dependent semantic context is not easy to learn in the conventional frameworks, but we present an efficient and effective framework with exemplar-based and graph-based context model. The second-order relationships play a crucial role in accurately estimating the semantic segmentation. Furthermore, these high-order relationships have the rich contextual knowledge of each image which cannot be learned by conventional classifiers. In Chapter 5, we consider the combined optimization framework for accurate semantic segmentation.

6.2 Future Works

The presented adaptive context aggregation networks can be generalized to not only to classification problems but also regression problem. Low-level image regression problem such as image super-resolution, deblurring, and optical flow problem can benefit from the proposed network model. The major difference of image regression problem is that the loss function for this problem is L2 loss function which is different from the usual classification loss such as hinge and softmax loss function. Furthermore, more general context aggregation model for semantic segmentation that can be applied to any types of visual recognition need to be investigated.

In the current study, a semantic context model which relies on graph and exemplar. However, due to the limitations of the computing resource, such association-based ap-

proach needs to be boosted for general applicability. Reducing complexity in semantic context modeling can have many potential applications in object recognition literature.

Bibliography

[1] A. Jain, A. Gupta, and L. S. Davis, "Learning what and how of contextual models for scene labeling," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2010.

[2] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] Z. Lu and H. H. Ip, "Constrained spectral clustering via exhaustive and efficient constraint propagation," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2010.

[4] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2010.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] J. L. Mundy, "Object recognition in the geometric era: A retrospective." in *Toward Category-Level Object Recognition*, ser. Lecture Notes in Computer Science, J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, Eds., vol. 4170. Springer, 2006, pp. 3–28.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[10] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[12] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision (IJCV)*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 8, pp. 1915–1929, 2013.

[16] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[19] R. Girshick, "Fast r-cnn," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.

[20] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence (TPAMI), vol. 38, no. 1, pp. 142–158, 2016.

[21] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[22] C. Galleguillos, B. McFee, S. Belongie, and G. R. G. Lanckriet, "Multi-class object localization by combining local contextual interactions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[23] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2009.

[24] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2008.

[25] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *International Journal of Computer Vision (IJCV)*, vol. 75, no. 1, pp. 151–172, Oct. 2007.

[26] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2005.

[27] H. Myeong, J. Y. Chang, and K. M. Lee, "Learning object relationships via graph-based context model," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[28] D. Parikh, C. L. Zitnick, and T. Chen, "From appearance to context-based recognition:dense labeling in small images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[29] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2007.

[30] A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in *Advances in Neural Information Processing Systems (NIPS)*, 2005.

[31] L.-C. Chen, A. G. Schwing, A. L. Yuille, and R. Urtasun, "Learning deep structured models," in *Proceedings of International Conference on Machine learning (ICML)*, 2015.

[32] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[33] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[34] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[37] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2014.

[38] H. Li, R. Zhao, and X. Wang, "Highly efficient forward and backward propagation of convolutional neural networks for pixelwise classification," *CoRR*, vol. abs/1412.4526, 2014.

[39] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[40] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.

[41] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[42] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, vol. abs/1511.00561, 2015.

[43] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[44] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2016.

[45] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *CoRR*, vol. abs/1506.04579, 2015.

[46] H. Myeong and K. M. Lee, "Tensor-based high-order semantic relation transfer for semantic scene segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[47] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[48] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[49] R. Vemulapalli, O. Tuzel, M.-Y. Liu, and R. Chellapa, "Gaussian conditional random field network for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[50] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[51] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of ACM International Conference on Multimedia*, 2015.

[52] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011.

[53] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[54] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, "Multi-class segmentation with relative location prior," *International Journal of Computer Vision (IJCV)*, vol. 80, no. 3, pp. 300–316, 2008.

[55] L. Ladicky, C. Russell, P. Kohli, and P. Torr, "Graph cut based inference with co-occurrence statistics," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2010.

[56] H. Kashima, T. Katoy, Y. Yamanishiz, and MasashiSugiyama, "Link propagation: A fast semi-supervised learning algorithm for linkprediction," in *SIAM International Conference on Data Mining*, 2009.

[57] J. Tighe and S. Lazebnik, "Understanding scenes on many levels," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011.

[58] T. Malisiewicz and A. A. Efros, "Beyond categories: The visual memex model for reasoning about objectrelationships," in *Advances in Neural Information Processing Systems (NIPS)*, December 2009.

[59] B. C. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman, "Object recognition by scene alignment," in *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[60] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[61] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[62] A. Oliva and A. Torralba, "Building the gist of a scene: the role of global image features inrecognition," in *Progress in Brain Research*, 2006.

[63] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[64] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision (IJCV)*, vol. 59, no. 2, pp. 167–181, Sep. 2004.

[65] X. Chen, A. Jain, A. Gupta, and L. S. Davis, "Piecing together the segmentation jigsaw using context," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[66] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pat-*

tern Analysis and Machine Intelligence (TPAMI), vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[67] V. Kolmogorov and R. Zabih, "What energy functions can be minimizedvia graph cuts?" *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 26, no. 2, pp. 147–159, Jan. 2004.

[68] E. Boros and P. L. Hammer, "Pseudo-boolean optimization," *Discrete Applied Mathematics*, vol. 123, no. 1-3, pp. 155–225, Nov. 2002.

[69] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer, "Optimizing binary mrfs via extended roof duality," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[70] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *International Journal of Computer Vision (IJCV)*, vol. 77, no. 1-3, pp. 157–173, May 2008.

[71] X. Chen, Q. Li, Y. Song, X. Jin, and Q. Zhao, "Supervised geodesic propagation for semantic label transfer," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2012.

[72] H. Zhang, T. Fang, X. Chen, Q. Zhao, and L. Quan, "Partial similarity based nonparametric scene parsing in certain environment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[73] H. Zhang, J. Xiao, and L. Quan, "Supervised label transfer for semantic segmentation of street scenes," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2010.

[74] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2012.

[75] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision (IJCV)*, vol. 81, no. 1, pp. 2–23, Jan. 2009.

[76] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "Sift flow: Dense correspondence across different scenes," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2008.

[77] S. Gould and Y. Zhang, "PatchMatchGraph: Building a graph of dense patch correspondences for label transfer," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2012.

[78] P. Kohli, L. Ladický, and P. H. Torr, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision (IJCV)*, vol. 82, no. 3, pp. 302–324, May 2009.

[79] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.

[80] Z. Fu, Z. Lu, H. H. S. Ip, Y. Peng, and H. Lu, "Symmetric graph regularized constraint propagation," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2011.

[81] H. Ishikawa, "Higher-order clique reduction in binary graph cut," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[82] X. He, R. S. Zemel, and M. A. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[83] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

초 록

본 논문에서는 임의의 영상 내에 존재하는 물체들을 정확히 분할하고 인식하는 의미론적 영상 분할 기법에 대해 다룬다. 기존의 일반적인 의미론적 영상 분할 기법은 두 가지 주요 요소, 시각적인 특징과 의미론적 맥락 정보에 기반을 두고 있다. 색, 에지, 모양 등과 같은 시각적인 특징 정보는 장면으로부터 객체를 추론하기 위한 주요한 정보이다. 하지만 실제 장면에서 나타나는 객체의 형상은 조명, 질감, 가리워짐 및 시점과 같은 영상 획득 조건의 영향을 받기 때문에 이미지 데이터로는 객체의 다양성을 완전히 포착할 수 없는 경우가 많다. 따라서, 영상 내의 다른 물체의 존재나 위치 정보로 대표되는 의미론적 맥락 정보가 의미론적 영상 분할 작업에서 시각 정보의 모호성을 해결하는데 매우 중요하다. 의미론적 맥락을 활용한 최신의 의미론적 영상 분할 기법들은 객체들 사이의 상호 작용의 모델링을 통해 잘못 인식된 영역들을 수정한다. 그러나 기존의 표현 학습 방법으로는 시각적인 형상과 의미론적 맥락을 동시에 학습할 수 없어 각각을 독립적으로 학습하는데 그치고 있다. 따라서 본 논문에서는 시각적 특징을 의미론적 맥락과 함께 학습할 수 있는 맥락 인식 기반 표현 방법을 제안하였다.

본 논문의 첫 부분에서는 맥락 기반 형상 표현 방법에 대하여 다룬다. 제안하는 적응형 맥락 집합 네트워크 (Adaptive context aggregation network) 는 여러 단계의 추론 과정 안에서 적절하게 의미론적 맥락 정보를 추출할 수 있도록 설계되어 있다.

둘째, 시각적 형상 특징을 활용하여 의미론적 맥락 정보를 강화하는 방안을 고안하였다. 그래프 및 예시 기반의 맥락 모델은 객체의 시각적 특징에 적합하게 객체의 상호 작용을 학습한다. 마지막으로, 맥락 기반 형상 표현 모델과 형상 기반 맥락 모델을 통합하여 정확한 영상 분할 결과를 얻기 위한 다시점 마르코프 랜덤 필드를 제안하였다. 실험을 통해 제안하는 기법이 높은 정확도로 의미론적 영상 분할을 수행함을 보였다.

주요어: 컴퓨터 비젼, 물체 인식, 의미론적 영상 분할, 맥락 인식

학번: 2009-20799