The International Joint Conference on Neural Networks

Relationship Aware Context Adaptive Feature Selection Framework for Image Parsing

Basim Azam

b.azam@cqu.edu.au

Central Queensland University

Australia

Ranju Mandal

r.mandal@cqu.edu.au

Central Queensland University

Australia

Brijesh Verma

b.verma@cqu.edu.au

Central Queensland University

Australia

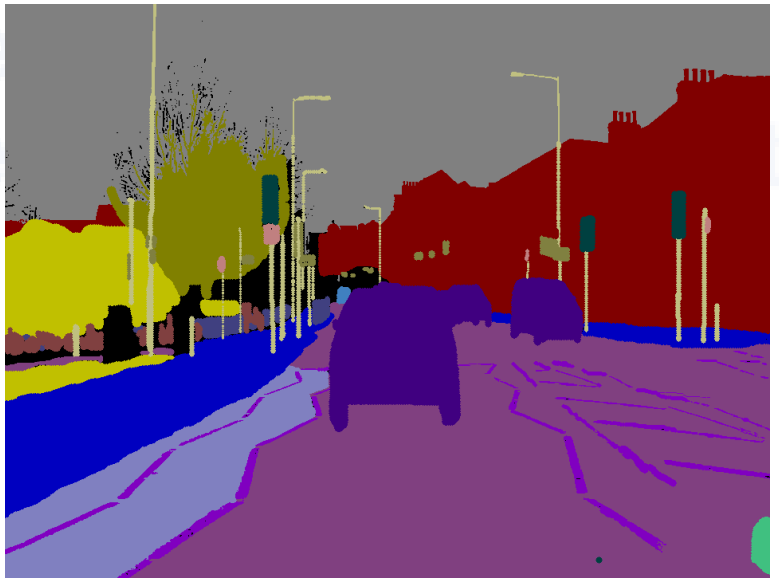
Outline

- Introduction
- Proposed Architecture
- Experiments
- Results
- Conclusion & Future Works

Introduction

Image Parsing

Image parsing refers to segmentation of an image into regions with object category labels such as tree, building, car and road.



Introduction

Applications

- Image parsing has a variety of applications and is being fundamentally applied in
  - Virtual reality
  - Traffic interpretability
  - Autonomous driving
  - Medical image processing
  - Remote sensing

Current Methods

Motivation

- The modern frameworks have shown capability for producing accurate pixel-wise labelling.
- CNN based approaches obtain a coarse label map my applying complex pixelwise convolution operations on input images.
- The context adaptability of Conditional Random Fields (CRF) leads to time consuming training inference and increased feature representation.
- It is highly desirable to learn critical features only in such high-dimensional feature space.

Proposed Approach

Overview

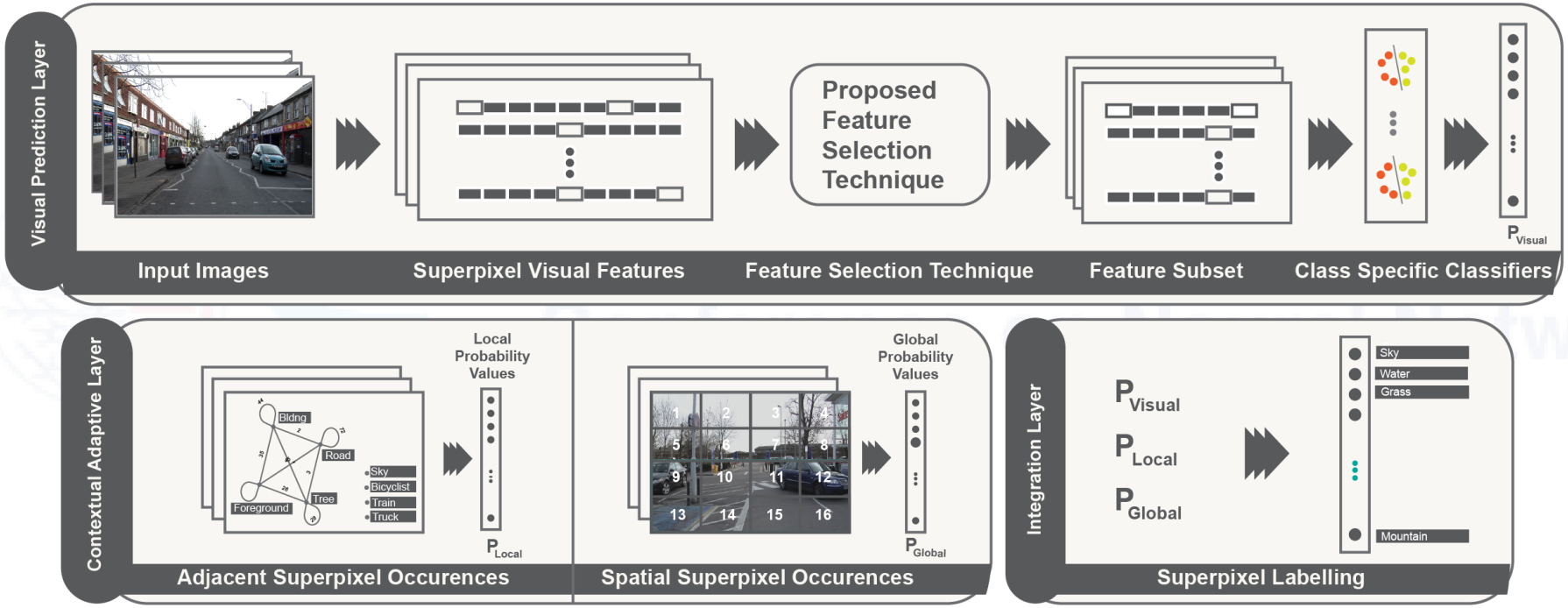
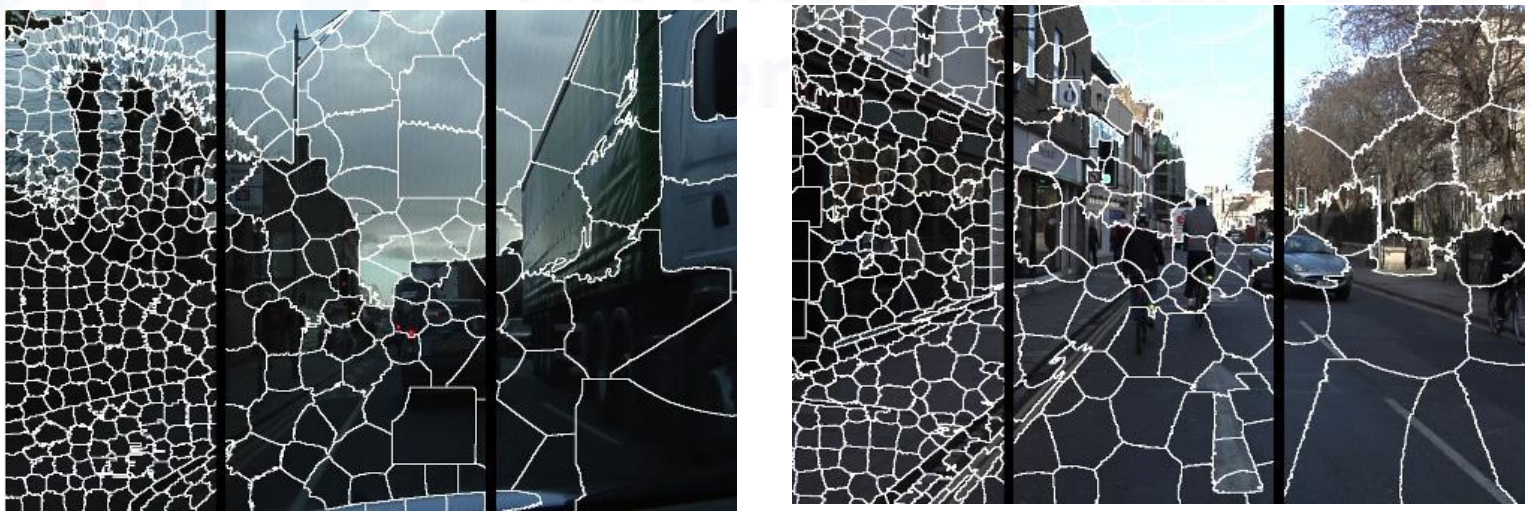


Image Parsing Framework. **Visual Prediction Layer**: this layer takes a subset of optimum (chosen by novel feature selection) super pixel visual features extracted from the image and trains class specific classifier to obtain probabilities for pixels. **Contextual Adaptive Layer:** CA layer considers the probability-based votes from adjacent superpixels and the spatial block object occurrences from the training data. **Integration Layer:** Integration layer renders final class labels for superpixels using the probabilities from VP layer and CA layer.

Visual Prediction (VP) Layer

Superpixel Computation

- The VP layer helps to compute class probabilities based on the visual features computed from the input training dataset.
- Initially, the input images is converted into set of superpixels using SLIC algorithm. For this study we have used superpixel values 256 and 512.



Images segmented using SLIC Superpixel algorithm of size 64, 512 and 1024 pixels.

Visual Prediction (VP) Layer

Visual Feature Computation

- The next step is to compute the visual features from these superpixel regions. The visual features computed from superpixels include
  - computation of color dissimilarities
  - change in geometry and texture
  - extraction of mean, standard deviation, mask and histograms of textons
  - the computation of SIFT descriptors
  - MR filter bank descriptors
- The overall length of the feature vector becomes 517 elements for each superpixel.

Visual Prediction (VP) Layer

Feature Selection

- The visual attributes computed may lead to overfitting problem, to avoid such problem we make use of *novel feature selection algorithm* proposed to chose a subset of affluent features.
- The attributes with higher ability to distinguish between classes are more important in terms of performance and instead of using every attribute only high-level attributes are chosen.
- The feature selection function considers mutual information between the features and the fisher criterion to rank the features in descending order.

$$Score_x = \lambda_1 * mutInf_x + \lambda_2 * Fisher_x$$

- A smaller subset of these highly related attributes are used to train the class-specific classifiers, ultimately affecting the time and computational complexities.

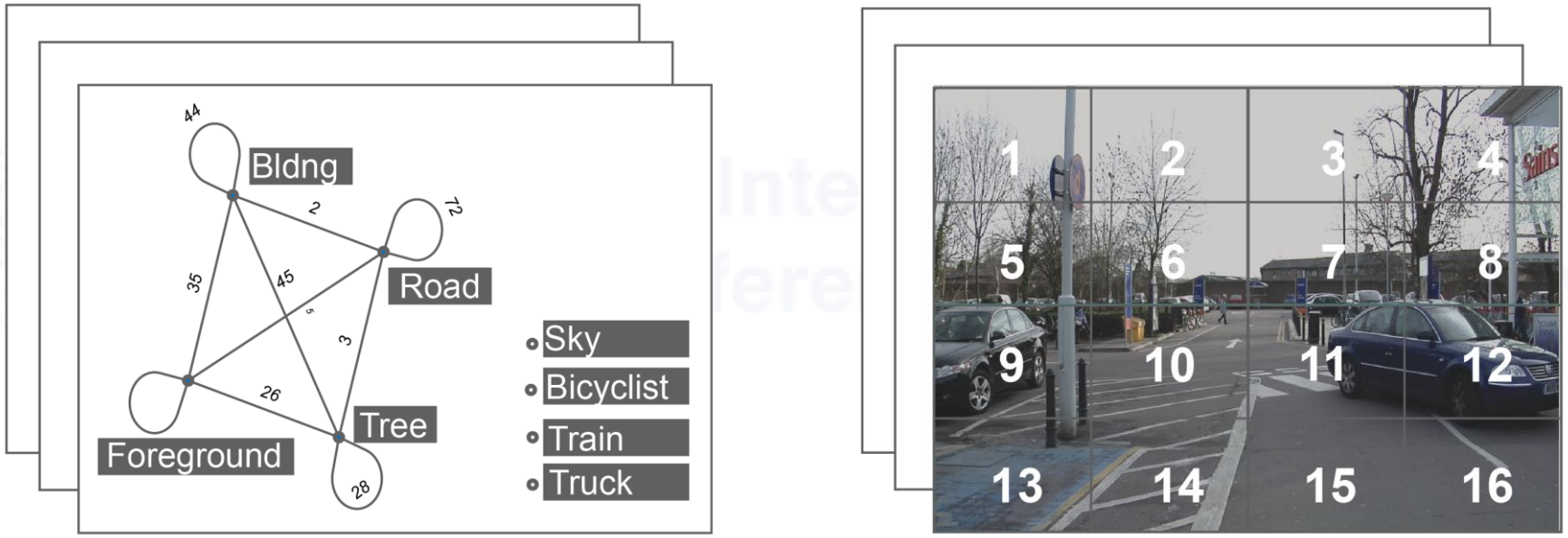
Visual Prediction (VP) Layer

Class-Specific Classifier

- We use the subset of highly important attributes to train 11 class-specific classifiers.
- The 11 models are trained for each of the classes in the CamVid dataset.
- The classifiers trained help compute the probability of superpixel belonging to the pre-defined semantic object category P_{visual}.

Context Adaptive (CA) Layer

Local & Global Occurrences



(a) **(b)**

(a) The adjacent superpixel vote computations and (b) global computation of votes for object occurrences.

Context Adaptive (CA) Layer

Local & Global Occurrences

- The CA layer assimilates the most likely class label for superpixels using the learned prior information of objects.
- The term contextual adaptation represents the information of the objects occurring within an image in varying ranges.
- CA layer computes the local occurrences by considering the adjacent superpixels and the global occurrence by computing the occurrence of object in a block B_k of superpixels.
- These local and global votes computed are then normalized to probability values respectively as P_{Local} and P_{Global}.

Integration Layer

- The weights of integration layer are computed using probability values estimated in VP layer (P_{visual})and CA layer (P_{Local} and P_{Global}).
- The Multilayer-Perceptron (MLP) is applied to learn these integration weights.

Dataset

Cambridge-driving Labeled Video Database (CamVid)

- CamVid dataset comprises of the road driving scene videos, including almost 10 minutes of best quality (970 x 720) footage.
- The scenarios consist of crowded scenes, varying illumination conditions and multiple types of roads.
- The 701 images annotated into 11 object categories have been used.
- The data is divided into 468 training and 233 test images.

The overview of the CamVid Database used in the study. Left: Percentage breakdown by category, Right: Visual sample images with labels

Results

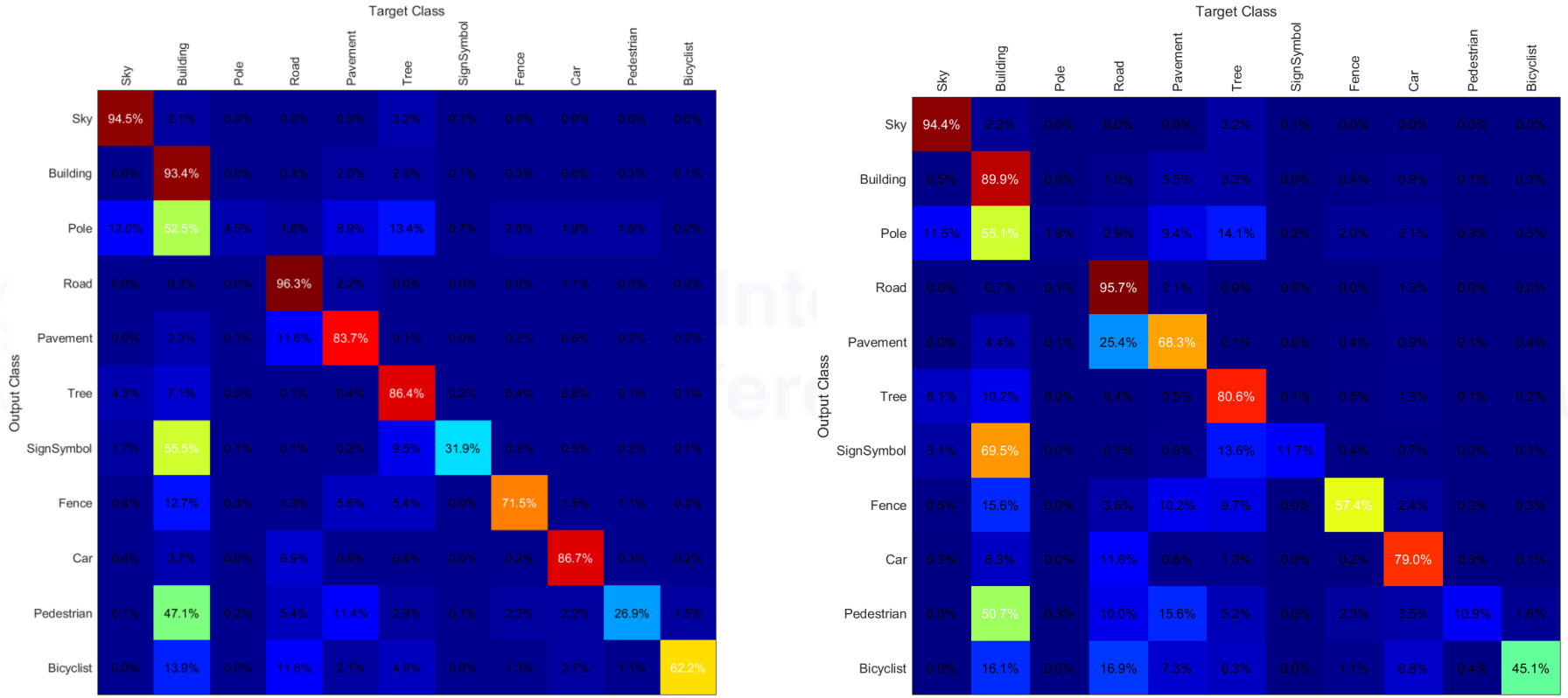
Table 1. Layer-wise Performance Evaluation in terms of Accuracy (%) for 512 superpixels

	VP Layer	CA Layer	Proposed Approach
512 Superpixels	68.25	87.61	**89.79**
Sky	95.79	92.66	**94.52**
Building	67.16	95.15	**93.44**
Pole	0.08	0	0
Road	95.52	98.22	**96.29**
Pavement	22.75	70.59	**83.69**
Tree	39.43	83.59	**86.42**
SignSymbol	0.06	0	**31.85**
Fence	26.76	61.43	**71.47**
Car	30.72	77.07	**86.73**
Pedestrian	11.83	0.02	**26.88**
Bicyclist	33.21	28.1	**62.16**

Table 2. Layer-wise Performance Evaluation in terms of Accuracy (%) for 256 superpixels

	VP Layer	CA Layer	Proposed Approach
512 Superpixels	66.57	81	**85.63**
Sky	96.6	88.62	**94.45**
Building	62.15	92.35	**89.93**
Pole	0.05	0	**0**
Road	97.05	98.22	**95.69**
Pavement	0	38.03	**68.3**
Tree	46.56	76.13	**80.64**
SignSymbol	0.04	0	**11.73**
Fence	25.69	32.92	**57.43**
Car	27.26	61.14	**79.04**
Pedestrian	27.24	0	**10.9**
Bicyclist	0.09	0.02	**45.06**

Results



Confusion matrix over the CamVid test set using Adaboost as class specific classifier and MLP in Integration Layer. Left: 512 Superpixels, Right:256 Superpixels

Results

Table 3. Quantitative Performance Comparisons with Previous Approaches on the CAMVID BG Dataset in terms of Accuracy (%), Proposed feature selection base architecture achieves competitive average and Global accuracy, It also performs Extremely well for various challenging classes (Building, Car, Tree, Fence, Bicyclist)

	Building	Tree	Sky	Car	Sign Symbol	Road	Pedestrian	Fence	Pole	Pavement	Bicyclist	Global	Average
Proposed Approach (ADB-MLP)	93.4	86.42	94.5	86.7	31.9	96.3	26.8	71.47	4.5	68.3	62.16	89.8	67.1
CNN-CRF [1]	84.3	65.3	95.6	74.6	0.4	93.5	25.6	32.3	13.8	85	54.3	72.9	56.8
SfM + Appearance [2]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	69.1	53.0
Super Parsing [3]	87.0	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70	19.4	83.3	51.2
Local Label Descriptors [4]	80.7	61.5	88.8	16.4	n/a	98.0	1.09	0.05	4.13	12.4	0.07	73.6	36.3
Boosting + Detector + CRF [5]	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	83.8	59.2
FCN+Comb [6]	79.7	77.2	85.7	86.1	45.3	94.9	45.9	69.0	25.2	86.2	57.9	88.7	63.8
ReSeg [7]	86.8	84.7	93.0	87.3	48.6	98.0	63.3	20.9	35.6	87.3	43.5	88.7	68.1

Conclusion

Conclusion

- A novel feature selection based deep learning architecture for image parsing is presented.
- The major novelties include
  - the introduction of short and long range contextual adaptive dependencies
  - the feature selection technique with adaptable parameters to the dataset.
  - It is demonstrated that the proposed architecture brings significant improvements to the global accuracy of 89.8% on CamVid Dataset

Future Works

- Evaluation of proposed approach on more benchmark datasets.
- The investigations to improve the proposed architecture by optimizing the learning parameters and fusion of layers.

References

1. J. M. Alvarez, Y. LeCun, T. Gevers, and A. M. Lopez, "Semantic road segmentation via multi-scale ensembles of learned features," in European Conference on Computer Vision, pp. 586-595, 2012.
2. P. Sturgess, K. Alahari, L. u. Ladický, and P. H. S. Torr, "Combining appearance and structure from motion features for road scene understanding," British Machine Vision Conference, pp. 1-11, 2009.
3. X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: Features and algorithms," in IEEE Conference on Computer Vision and Pattern Recognition pp. 2759-2766, 2012.
4. Y. Yang, Z. Li, L. Zhang, C. Murphy, J. Ver Hoeve, and H. Jiang, "Local label descriptor for example based semantic image labeling," in European Conference on Computer Vision pp. 361-375, 2012.
5. L. U. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr, "What, where and how many? Combining object detectors and CRFs," in European Conference on Computer Vision vol. 6314, pp. 424-437, 2010.
6. Y. Wu, T. Yang, J. Zhao, L. Guan, and J. Li, "Fully combined convolutional network with soft cost function for traffic scene parsing," in International Conference on Intelligent Computing, vol. 10361, pp. 725-731, 2017.
7. F. Visin et al., "ReSeg: A recurrent neural network-based model for semantic segmentation," IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 426-433, 2015.

Thank You.