# Region-based Deformable Fully Convolutional Networks for Multi-class Object Detection at Signalized Traffic Intersections

NVIDIA AICity Challenge 2017 Track 1

Shuo Wang, Koray Ozcan, and Anuj Sharma
Institute for Transportation, Iowa State University
2711 South Loop Drive, Ames, IA
shuowang@iastate.edu, koray6@iastate.edu, anujs@iastate.edu

***Abstract*—Multi-class object detection is critical for intelligent traffic monitoring applications in smart cities as well as connected autonomous vehicles. Although, numerous research works evaluate the performance of image processing algorithms for on-vehicle cameras, the body of research evaluating performance of image processing of stationary cameras located near intersections is limited. In this research, we use region-based deformable fully convolutional networks to detect 14 different object classes within images from traffic surveillance cameras. The object classes include vehicles, pedestrians, bicyclists and traffic signals. The goal of the NVIDIA AICity challenge is to provide accurate localization and correct classification of objects for stationary cameras mounted near signalized traffic intersections. Our proposed method scores mean average precision of 0.41, 0.37, and 0.34 for aic480, aic1080, and aic540 challenge datasets.**

***Keywords—multi-class object detection, surveillance camera, deformable convolutional networks, and deep learning***

## I. INTRODUCTION

Traffic surveillance of intersections and highways has been crucial for transportation monitoring and control. With millions of traffic surveillance cameras installed for traffic monitoring, the opportunities are increasing for deriving insights from cameras in order to make traffic flow safer and efficient. Therefore, detection and classification of different vehicles, pedestrians, bicyclists and traffic signals using the surveillance cameras is essential for future developments of any intelligent traffic applications. Currently the transportation industry uses different technologies, such as, radar based, inductive loop based, magnetic inductance based, image based etc., for detecting vehicles. The image based technology and radar based technology are considered non-intrusive as they can be mounted on the signal or street light poles without disturbing the road surface. Despite the benefits of being non-intrusive, the use of video detection technology has been marginal due to its inferior performance as compared to inductive loop detectors of magnetometers. Some of the common drawbacks cited by road managers about inefficiency of video detection is its poor performance under different lighting conditions, significant numbers of missed detections and false positives. This paper explores how the use of deep neural network based approach changes the abilities of video detection logic under real-world conditions.

In recent years, computer vision field has been transformed with the application of convolutional neural networks (CNN). One of the early applications of CNNs was on hand written number recognition by LeCun et al. in 1998 [1]. Not much traction was seen in following years, due to the lack of massive training data sets and processing power required for training CNNs. In 2012, Alexnet, a deep CNN model, developed by Alex Krizhevsky et al. [2] was introduced for the image classification task of Large Scale Visual Recognition Challenge (LSVRC) [3]. It significantly outperformed all the prior competitors and won the challenge by a large margin. The availability of large annotated datasets, from social media and easy access to parallel processing GPUs have made it possible to implement CNNs for practical applications. Since then, image classification has been improved by different kinds of CNN variations. In 2014, VGG networks from Oxford [4] demonstrated the benefits of using deeper models with more CNN layers. In 2015, ResNet [5] introduced the shorter paths between layers and pushed the CNN networks to over a hundred layers.

Object detection and classification performances have improved significantly with the evolution of CNNs. Region-based convolutional neural networks (R-CNN) [6] views detection as a classification problem. It first generates reginal proposals using unsupervised methods and then classifies each of the reginal proposals by separate CNNs. Although R-CNN provides high accuracy for object detection, the processing time for R-CNNs is long and that limits their usability for real-time applications. Fast R-CNN [7] improves the running speed by using ResNet, as a backbone, to extract features that are reused for all the regional classifiers. In Fast R-CNN, most of the CNN computations are shared within the backbone so that the running speed is no longer limited by CNN computations. Faster R-CNN [8] further improves the speed by replacing the unsupervised methods for generating regional proposals with another set of CNN layers, named, reginal proposal network (RPN) making it an end to end deep net model. Region-based fully convolutional neural networks (R-FCN) [9] carries on the idea of sharing CNN

layers from faster R-CNN and introduces a location sensitive maps on top of RPNs to share all the CNN computations. With R-FCN, the entire object detection is an end-to-end learning process and the processing speed is better than faster R-CNN.

Recently, another CNN variation was proposed that further improves the accuracy of deep classifiers. Deformable ConvNets [10] introduced adaptive-shaped convolutional filters by using an additional convolutional layer to learn the shape of the filter between two original convolutional layers. State-of-art results were presented using deformable ConvNets on both Faster R-CNN and R-FCN network pipeline. In this research, we train R-FCN with Deformable ConvNets for multi-class object detection task using images from traffic intersection surveillance cameras.

## II. Data Description

IEEE Smart World NVIDIA AI City Challenge [11] shared the largest data set on real-world signalized traffic intersection images. Participating teams collaboratively annotated more than 110,00 key-frames extracted from over 80 hours of traffic videos captured at various intersections within the cities of Santa Clara and San Jose in California, and Virginia Beach, Virginia. Rectangular bounding boxes and object labels are provided for 14 different classes including: car, SUV, small truck, medium truck, large truck, pedestrian, bus, van, group of people, bicycle, motorcycle, traffic signals – green, red, and yellow. The annotated images were processed into three datasets, namely aic480, aic1080 and aic540.

### *A. AIC480*

The aic480 dataset contains the images of size $720 \times 480$ captured from Virginia Beach, Virginia. There are four intersections for the dataset: 1. Princess Anne Rd & Lynnhaven Pkwy; 2. Great Neck Rd & First Colonial Rd; 3. Indian River Rd & Reon Dr; 4. Dam Neck Rd & London Bridge Rd. Fig. 1 shows the sample images from the aic480 dataset.

Fig. 1. Example images from the dataset aic480. The training set contains three intersections (annotations are shown on images): Princess Anne Rd & Lynnhaven Pkwy (first row), Great Neck Rd & First Colonial Rd (second row), Indian River Rd & Reon Dr (third row). The test set contains one intersection (no annotations provided for images): Dam Neck Rd & London Bridge Rd (last row).

### *B. AIC1080/AIC540*

The aic1080 dataset contains images of size $1920 \times 1080$ captured from Santa Clara and San Jose, California. There are three intersections included: 1. Stevens Creek Blvd & Winchester Blvd; 2. Walsh Ave & San Tomas Expy; 3. San Tomas Aquino Rd & Hamilton Ave. The aic540 dataset contains exactly the same images as the aic1080 dataset except all images are downscaled to size of $960 \times 540$. Fig. 2 shows the sample images from the aic1080 and the aic540 datasets.

### *C. Summary*

For each dataset, training images are provided with corresponding annotations while the released test images have no annotations. Table 1 summarizes the sample distribution in each dataset.

Fig. 2. Example images in datasets aic1080 and aic540. The training set contains two intersections (annotations are shown on images): Stevens Creek Blvd & Winchester Blvd (first row), Walsh Ave & San Tomas Expy (second row). Besides these two, the test set contains one more intersection (no annotations provided for images): San Tomas Aquino Rd & Hamilton Ave (last row).

TABLE I. DATASET SUMMARY

| AIC | Intersections | training samples | test samples |
|---|---|---|---|
| 480 | Princess Anne Rd & Lynnhaven Pkwy | 4115 | 0 |
| | Great Neck Rd & First Colonial Rd | 3376 | 0 |
| | Indian River Rd & Reon Dr | 3485 | 0 |
| | Dam Neck Rd & London Bridge Rd | 0 | 3372 |
| | total | 11016 | 3372 |
| 1080 /540 | Stevens Creek Blvd & Winchester Blvd | 3274 | 2410 |
| | Walsh Ave & San Tomas Expy | 75480 | 17408 |
| | San Tomas Aquino Rd & Hamilton Ave | 0 | 1800 |
| | total | 78754 | 21618 |

## III. METHODOLOGY

### A. RDFCN

We name our model Region-based Deformable Fully Convolutional Networks (RDFCN). RDFCN consists of two major components: R-FCN and Deformable ConvNets.

### B. R-FCN

In R-FCN, the location variance information is captured by position-sensitive score maps after the last convolution layer of the resnet-101 backbone. RPN provides the regional proposals. The final output layer directly takes the features from position-sensitive maps and reginal proposals to provide detections. All convolutional computations are shared in the object detection pipeline to achieve a fast processing speed.

### C. Deformable ConvNets

Deformable ConvNets expands the potential feature space by adaptive-shaped convolutional filters. The adaptive-shaped convolutional filters are achieved by introducing offsets to the sampling locations of a regular convolutional filter rather than doing a naïve square-shaped sampling around pixel location. The offsets are generated from the preceding feature maps via an additional convolutional layer which can be learnt using normal stochastic gradient descent (SGD) end-to-end. These additional convolutional filters expand the feature geometry in a natural and adaptive fashion and thus reduces the needs of data augmentation and improves overall detection accuracy.

### D. Transfer Learning

Transfer learning has been widely applied in deep learning and computer vision applications. A new task can benefit from previously well-trained models. The Microsoft COCO dataset [12] contains more than 2 million well-labeled objects of 80 different categories in more than 300,000 images. We used the RDFCN weights pre-trained on COCO dataset as the initialization of this detection task. Multiple transfer learning strategies were tested and eventually we choose to freeze the weights of ResNet-101 and RPN throughout the training and fine tune the location-sensitive score maps and the output layer.

## IV. EXPERIMENTAL RESULTS

The RDFCN models were trained using an Ubuntu 16.04 machine with 64 MB of RAM and one NVIDIA TITAN X GPU. Codes are open sourced and available on Github: https://github.com/NVIDIAAICITYCHALLENGE/AICity_Team6_ISU.

### A. AIC480

The model was trained on the training set of aic480 for 40 epochs. Fig. 3 shows the sample detections of the trained RDFCN model on the test set of aic480.

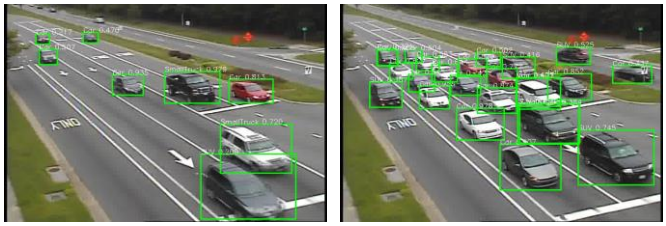

Fig. 3. Sample detections (green boundary boxes with class label and confidence score on top) of dam_neck_london_bridge in dataset aic480.

The precision-recall curves in Fig. 4 present the overall performance of our model on the test dataset of aic480. The average precision scores of each class are summarized in Table 2 below.

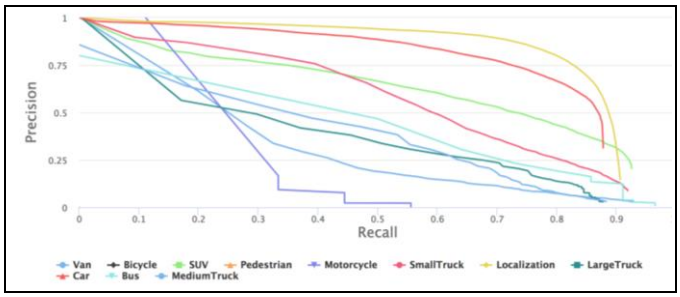


Fig. 4. Precision-recall curve of the submitted model on the test set of aic480.

### B. AIC1080/AIC540

The model was trained on the training set of aic1080 for 5 epochs. This aic1080 model was then trained on the training set of aic540 for 10 epochs. Fig 5 shows the sample detections of the trained RDFCN model on the test dataset of aic1080.

Fig. 5. Sample detections (green boundary boxes with class labels and confidence scores on top) of stevens_cr_winchester (first row),

walsh_santomas (second row) and san_tomas_hamilton (last row) in dataset aic1080.

The precision-recall curves in Fig. 6 show the overall performance of our model on the test dataset of aic1080 and aic540 from the online evaluation system of NVIDIA AICity challenge. The detailed average precision scores of each class are summarized in Table 2.

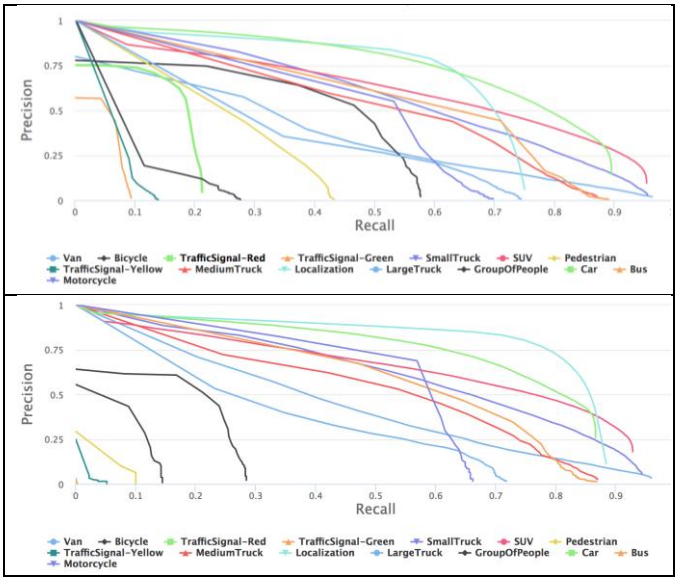


Fig. 6. Precision-recall curves of the submitted model on the test set of aic1080 (first row) and aic540 (second row).

TABLE II. MODEL PERFORMANCE SUMMARY

| | AIC480 | AIC1080[b] | AIC540 |
|---|---|---|---|
| Class | AP | AP | AP |
| Car | 0.76 | 0.7 | 0.69 |
| SUV | 0.61 | 0.6 | 0.63 |
| Small Truck | 0.56 | 0.57 | 0.6 |
| Median Truck | 0.37 | 0.47 | 0.47 |
| Large Truck | 0.35 | 0.3 | 0.32 |
| Pedestrian | 0 | 0.2 | 0.02 |
| Bus | 0.5 | 0.56 | 0.54 |
| Van | 0.29 | 0.36 | 0.41 |
| Group of People | N/A[a] | 0.06 | 0.06 |
| Bicycle | 0 | 0.37 | 0.15 |
| Motorcycle | 0.22 | 0.48 | 0.52 |
| Traffic Signal - Red | N/A[a] | 0.14 | 0 |
| Traffic Signal - Green | N/A[a] | 0.04 | 0 |
| Traffic Signal - Yellow | N/A[a] | 0.07 | 0 |
| Localization | 0.82 | 0.62 | 0.76 |
| mAP | 0.41 | 0.37 | 0.34 |

[a.] The test set of aic480 does not contain "group of people" and traffic signals.

[b.] By the due date of the challenge only the detection results on aic1080 were submitted and it was ranking the 2nd among the participants.

## V. CONCLUSTION

Based on the results obtained on the three datasets it can be seen that the performance of the model varies from class to class. Vehicle classes had significantly better performance when compared against bicycle or motorcycle classes. It should be noted the training data set had relatively smaller number of motorcycles and bicyclist and hence improving the training data set can improve the performance for those classes significantly. It should be also noted that traffic signal color detection problem had abysmal results and the reason was again the poor annotation quality for these classes. Another observation was that the test dataset itself had a number of mislabeled classes implying that these reported numbers of accuracy shouldn't be used in absolute sense but they should be used as comparison tools to compare other models or performance among different classes. As a recommendation for feature work, we would like to improve the training and test data sets to have a better understanding of absolute model performance.

## ACKNOWLEDGMENT

The authors are very thankful for the hard annotation work from Lakshay Ahuja, Pranamesh Chakraborty, Sambuddha Ghosal, Tingting Huang, Jacob Hess, Burak Kakillioglu, Gozde Ozcan, Logan M. Peters, Sandeep Rawat and Yu Zheng We are also thankful for the careful guidance from Dr. Soumik Sarkar.

## REFERENCES

[1] LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*86, no. 11 (1998): 2278-2324.

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems*, pp. 1097-1105. 2012.

[3] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge." *International Journal of Computer Vision* 115, no. 3 (2015): 211-252.

[4] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*(2014).

[5] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." In *Proceedings of the IEEE international conference on computer vision*, pp. 1026-1034. 2015.

[6] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587. 2014.

[7] Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

[8] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks." In *Advances in neural information processing systems*, pp. 91-99. 2015.

[9] Dai, Jifeng, Yi Li, Kaiming He, and Jian Sun. "R-fcn: Object detection via region-based fully convolutional networks." In *Advances in neural information processing systems*, pp. 379-387. 2016.

[10] Dai, Jifeng, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. "Deformable Convolutional Networks." *arXiv preprint arXiv:1703.06211* (2017).

[11] IEEE Smart World NVIDIA ai city challenge, 2017. http://smart-city-conference.com/AICityChallenge/

[12] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In *European conference on computer vision*, pp. 740-755. Springer, Cham, 2014