# Billboard Detection in the Wild

Miss Sayali Avinash Chavan, Dr. Dermot Kerr, Prof. Sonya Coleman, Mr. Hussein Khader
*Intelligent Systems Research Centre*
*School of Computing, Engineering & Intelligent Systems*
*University of Ulster*
*Northern Ireland, United Kingdom*

**Abstract**

Advertising has a huge impact on modern life hence its analysis is very important. Billboard detection in a scene is very challenging given the outdoor position of billboards and the changing nature of a board's size, scale, and the angle at which it is viewed by oncoming traffic or pedestrians. Hence the requirement to detect the billboard and determine the visibility to consumers is a very difficult task. In this paper, we propose a system which will not only detect a billboard but also classify the different types of billboard panels. There exists a number of different types of billboard such as Street Furniture, Roadside, in-Mall, or Spectacular to name a few, however here we focus solely on Street Furniture and Roadside. For this, the Tensorflow object detection API is used with the Single Shot Multibox Detector (SSD) architecture. SSD is chosen because of its high-speed computation and ability to eliminate false-positive cases. In this paper, we demonstrate SSD's detection performance by fine-tuning hyperparameters and illustrate this using a dataset of billboards in the wild.

**Keywords:** Object Detection, Single Shot Detector (SSD), Deep Learning, Billboard, Image Processing

## 1. Introduction

Billboards are the only media which cannot escape the attention of the pedestrians and drivers of any moving vehicle. They are also considered a more reliable source of advertising than advertisements seen online [Borisova, O. and Martynova, A. 2017]. Advertisements influence a person's mind and this ultimately increases the profitability of the company who has placed them. Traditionally, outdoor impressions have been measured using traffic counts and daily circulations to calculate total reach and quantify consumer viewing of billboard advertising. More recently, with the advancement of Spatial and GIS information, along with SDKs and millions of applications, location based mobile data are widely used to measure audience impressions and give insight. While eye-tracking technology offers huge advancements in outdoor impression measurement, considering the high cost of equipment and wearables combined with the sample size, its use still poses difficulties around accurate measurement of billboard visibility. Visibility factors take the physical visibility characteristics of billboards into consideration which are independent of an individual's visual attention and are therefore more suited for obtaining an overall measurement of billboard visibility [Wilson, R. T. and Casper, J. 2016].

When dealing with detection of objects or structures outdoors or 'i*n the wild*', we need to consider factors such as the object location, the viewing angle of the billboard, whether the object is illuminated, if it is static or digital signage, if a car is in motion then what is the contact zone, occlusion by trees or many other factors. Visibility also can be obstructed by the weather conditions, glare from any light source, relative position of object and observer distance, brightness of background etc. All such factors, which can be considered as Visibility Adjustment Indices (VIAs), play an important role in determining the impact and value of an advertisement impression, which will affect the marketable value of a particular billboard at a particular location. For example, if a billboard is partially covered by a tree from the pedestrians on the other side of the road or it is not visible due to illumination issues, this could potentially reduce it's value to an advertiser. In the case of high billboard visibility such as placement of a unit beside traffic lights, it will most likely be viewed by drivers as well as pedestrians, and therefore will have increased value [Wilson, R. T. and Casper, J. 2016]. Hence the VAIs play a key role in advertising billboard value. In this paper we present a dataset which contains static billboard images captured to include a range of these possible visual variations. We then determine the billboard's location by manually annotating a bounding box, and subsequently training a convolutional neural network to automatically recognise the billboard. Detection performance is evaluated and visual performance demonstrated.

## 2. State of Art

Computer vision is commonly used for understanding the external world through the use of algorithms. Common applications include understanding image data in order to identify and classify objects, object tracking, vehicle position monitoring, lane tracking and night time lane marking recognition [Li, Y. et al. 2016]. Much research has also focussed on developing algorithms to model human attention and saliency; saliency detection is an automatic process of locating the key parts of an image without any prior knowledge. Issues with current saliency approaches are that models use contrast and colour for low-level saliency cues [Krishna, O. and Aizawa, K. 2018].

Computer vision techniques have been previously used to specifically detect advertising boards. Such approaches have been based on Canny edge detection and morphological operations in order to determine the rectangular area of a billboard. However, many of these approaches used datasets containing minimal background noise and therefore are unable to generalise for the detection of unknown billboards in a range of dynamic scenes and environments [Rahmat, R.F. et al. 2019]. Another study demonstrated that to automatically detect regions that may be billboards, planar object detection can initially be used to locate and describe such objects. Planar object detection involves gathering individual object level information from an image then classifying which one of those objects is a billboard [Liang, P. et al. 2018]. In [Watve, A.K. and Sural, S. 2007] the focus was on the detection of advertising billboards on a soccer field, and this was achieved using the Fast Fourier Transform (FFT). The approach considered field detection, baseline detection, occlusion by players, image rectification, advertising board height detection and advertising board extraction using image data collected with a high resolution camera to provide the required clarity and high frequency colour intensities. The research in [Cai, G., Chen, L. and Li, J. 2003] focussed on advertising detection in sport TV using Hough transforms and geometric features of text to extract information from live high quality images and showed promising results. Another example is detecting advertisements from buildings in order to determine if they are in accordance with rules and regulations using computer vision techniques such as image rectification and segmentation in order to find the coordinates of the billboard object [Bochkarev, K. and Smirnov, E. 2019].

All these existing approaches are based on localisation and classification processes typically at pixel level. This process is very slow and has localisation problems when there are multiple objects in the scene. Additionally, considering real time use with low quality, blurry or occluded images, these methods have several limitations. There are various feature detectors available, however they are not capable of handling large amounts of data in real time applications. Hence Convolutional Neural Networks have become popular for working with large image datasets and high speed object detection and recognition [Shi, W., Bao, S. and Tan, D. 2019].

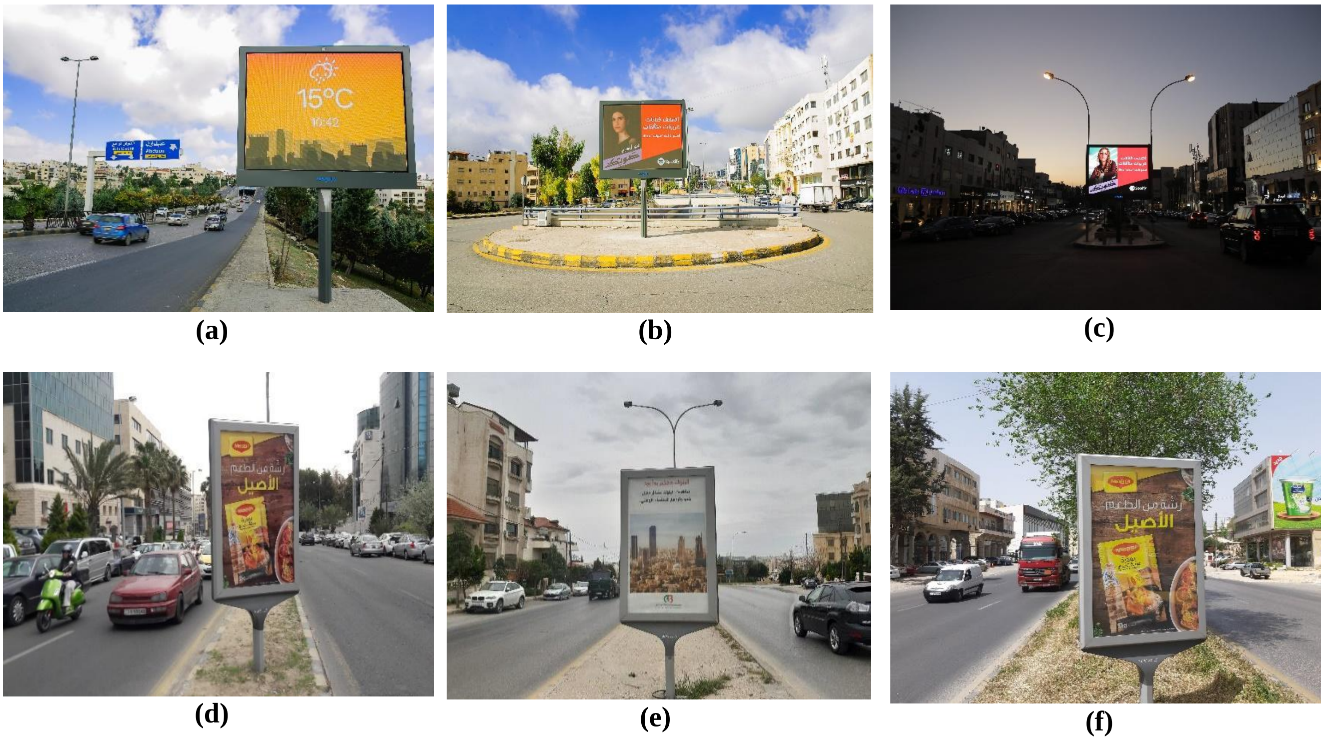
When neural networks are built using a number of convolutional layers in its model, these are known as Convolutional Neural Networks (CNNs). CNNs detect patterns in the image data and produce highly accurate predictions which are often measured as a percentage of correct classifications. Previous work using CNNs specifically for object detection has seen them applied to a wide range of applications ranging across medical applications, robotics, industry, wildlife detection, and geo-tagging. Most of the modern neural network architectures are derivatives of the famous ImageNet competition on supervised learning in computer vision in 2010, for example AlexNet, VGG, GoogLeNet, NiN, DenseNet and ResNet [Alom, M. Z. et al. 2018]. The best feature of a CNN is the capability to use transfer learning where the pre-trained model transfers the weights of its learned network to initiate the process of fine-tuning for another (unseen) dataset. There is a wide range of existing pre-trained neural network models like YOLO,  SSD MobileNet, Faster R-CNN ResNet, and R-FCN ResNet which use various open-source frameworks such as TensorFlow Object detection API, PyTorch, Microsoft cognitive toolkit, Keras, OpenCV, and DDN Library [Rahmat, R. F. et al. 2019]. The ADNet architecture is specifically designed to detect advertising instances from video frames and uses Microsoft COCO dataset to train its network [Hossari, M. et al. 2018]. One of the most recent examples of research based on billboard detection includes a comparative study of a SSD model vs YOLO (You Only Look Once) which showed promising results [Morera, Á. et al. 2020]. Hence we will utilise SSD in this study.

# 3. Methodology

The chosen methodology focusses on using a large dataset of annotated billboard images to train a convolutional neural network in order to recognise different classes of billboard. This section describes the dataset, the architecture of SSD and the experimental setup.

## 3.1 Dataset

The given dataset (see examples in Figure 1) consists of high-resolution images of real-world billboards with various background scenes. For each image the billboard is positioned in a different geographical location, contains differing advertisement content, the billboard is subject to various changes in orientations, varying positions from where the image was taken, has been captured over a wide range of times and thus is subject to daily and seasonal variations. Within each image there may be multiple background objects such as roads, pedestrians, vehicles, buildings, trees etc.



(a) (b) (c)

(d) (e) (f)

**Figure 1.  Selection of different images representing the different billboard classes and imaging conditions from the dataset: (a)-(c) Roadside Billboards; (d)-(f) Street Furniture Billboards**

The billboards can be approximately grouped into two classes based on their overall size and location: Roadside Billboards (see Figure 1(a)-(c)) are approximately 325x250 cm  and can be placed on the pavements , or major intersections , mounted on one or two posts; Street Furniture billboards (see Figure 1(d)-(f)) are approximately 169 x 111  cm  and often placed alongside road central reservations or pedestrian areas and mounted using a single post or a Base . The only obvious distinguishing feature between the billboard classes is that the Roadside Billboards are square shaped, and Street Furniture Billboards are rectangular shaped**.** All billboards have a frame and include the advertising company logo.

Billboards have been manually labelled using the LabelImage software to annotate a bounding box around the billboard or billboards within each image. In all cases the bounding box includes the entirety of the billboard frame as well as partial representation of the mounting posts when present. All bounding boxes retain a small proportion of background information. In total 1052 images were annotated, 532 images containing Roadside  Billboards and 520 images containing Street Furniture billboards, and labelled with the appropriate class label. This provided an annotated PASCAL VOC format dataset.

## 3.2 SSD Convolutional Neural Network

The Single Shot Detector (SSD) [Liu, W. et al. 2016] is a feed-forward convolutional network that predicts bounding boxes and the classes directly from feature maps in one single pass, hence why it is known as the Single Shot Detector. The SSD detector is composed of 2 parts as illustrated in Figure 2: extraction of feature maps, and application of convolution filters to detect objects. In most cases the early network layers are a standard VGG-16 network [Simonyan, K., & Zisserman, A. 2014] which has been truncated prior to any classification layers used to extract the feature maps; the remaining network structure is composed of six additional convolutional feature layers that are appended to the end of the truncated VGG-16 network.

After passing through the VGG-16 layers we obtain a feature layer with a number of bounding boxes corresponding to object region locations. These bounding boxes may be different sizes and aspect ratios as a vertical rectangle is more fit for Street Furniture billboard, and a square is more fit for a Roadside billboard. For each bounding box the class score is computed along with offsets which correspond to the distance from the original bounding box shape.

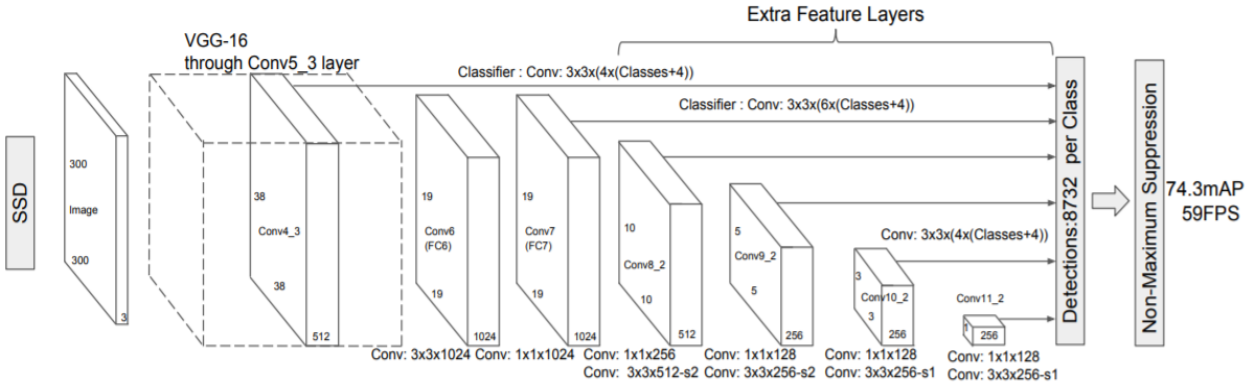


**Figure 2. Architecture of SSD [Liu, W. et al. 2016]**

To further improve detection, the feature maps go through the remaining network convolution feature layers where the location and class scores are computed using small convolution filters. SSD applies 3 × 3 convolution filters for each location to make predictions. The 3 × 3 convolution filters compute the results just like the regular CNN filters. SDD uses multiple layers to detect objects at different scales because as the spatial dimension is reduced the feature map resolution also reduces. In SSD, the lower resolution layers are used to detect objects at a larger scale and the higher resolution layers are used to detect objects at a smaller scale. Multi-scale feature maps have been shown to improve accuracy significantly [Liu, W. et al. 2016].

During SSD model training, the loss function is calculated using values obtained from the labelled, predicted and offset categories. The loss function is the sum of a classification loss and localization loss controlled by cross validation further comparing with matched bounding boxes as shown in equation 1:

$$L = \frac{1}{N}(Lc + \aleph Ll) \tag{1}$$

where, $L$ is the loss, $Lc$ is the classification loss, $Ll$ is the localisation loss, $N$ is the number of matched values of the bounding box, and $\aleph$ is the cross validation calculated balanced weight between two losses.

## 4. Experimental Setup and Results

We used the annotated image dataset described in Section 3.1 with the Tensorflow object detection API [Abadi, M. *et al.* 2016] and is used with pre-trained SSD mobileNet [Howard, A. G. *et al.* 2017]. The dataset consist of 1052 images is divided into two parts: 926 images (88.02%) of the images used as the training set and 126 images (11.98%) which are unseen during training and used to test the resulting network. In order to train the system, we tuned a number of hyperparameters, including batch-size, step-size and learning rate. Batch-size was varied from 24 to 1, step-size was varied from 25,000 to 75,000, and learning rate was varied between 0.001 and 0.004. Overall, the optimal parameters were found to be batch-size = 1, step-size = 75,000 and learning rate = 0.004.

Performance evaluation was conducted in parallel with the hyperparameter tuning to determine the optimal parameters. To do so, we use metrics such as precision, recall, average recall (AR), mean average precision (mAP), intersection over union (IoU) and loss each described as follows:

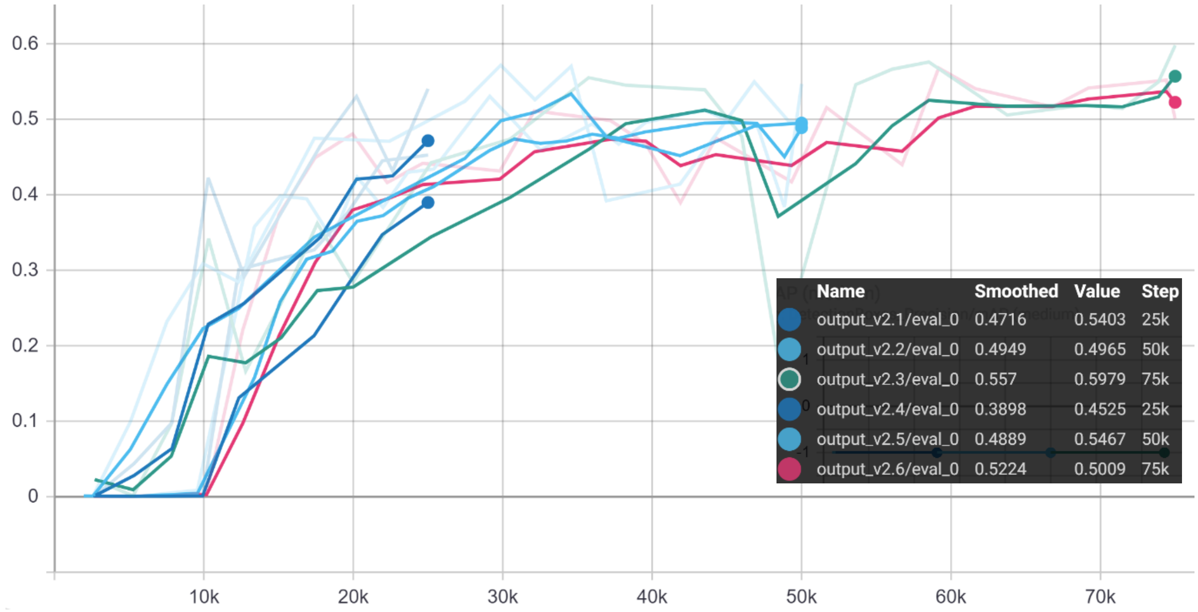
$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad , Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- Precision quantifies the number of positive class predictions that actually belong to the positive class
- Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.
- AR is the average of correctness of category prediction over IoU's.
- mAP calculates the average precision for each class based on the network's predicted bounding values and is combined with IoU, defined as the Area of the overlap divided by the area of the union of a predicted bounding box to adjust the accuracy if the match.
- Loss is the localisation loss calculated as the difference between the ground truth bounding box value and the predicted bounding box.

| **Step size** | 25,000 | 50,000 | **75,000** |
|---|---|---|---|
| **Loss** | 6.156 | 6.774 | **6.022** |
| **AR** | 62,71 | 53.88 | **64.90** |
| **mAP** | 54.03 | 49.65 | **59.79** |
| **mAP@.50IOU** | 80.01 | 71.19 | **83.55** |
| **mAP@.75IOU** | 63.21 | 65.37 | **73.04** |
| **(a) Learning rate = 0.004** | | | |

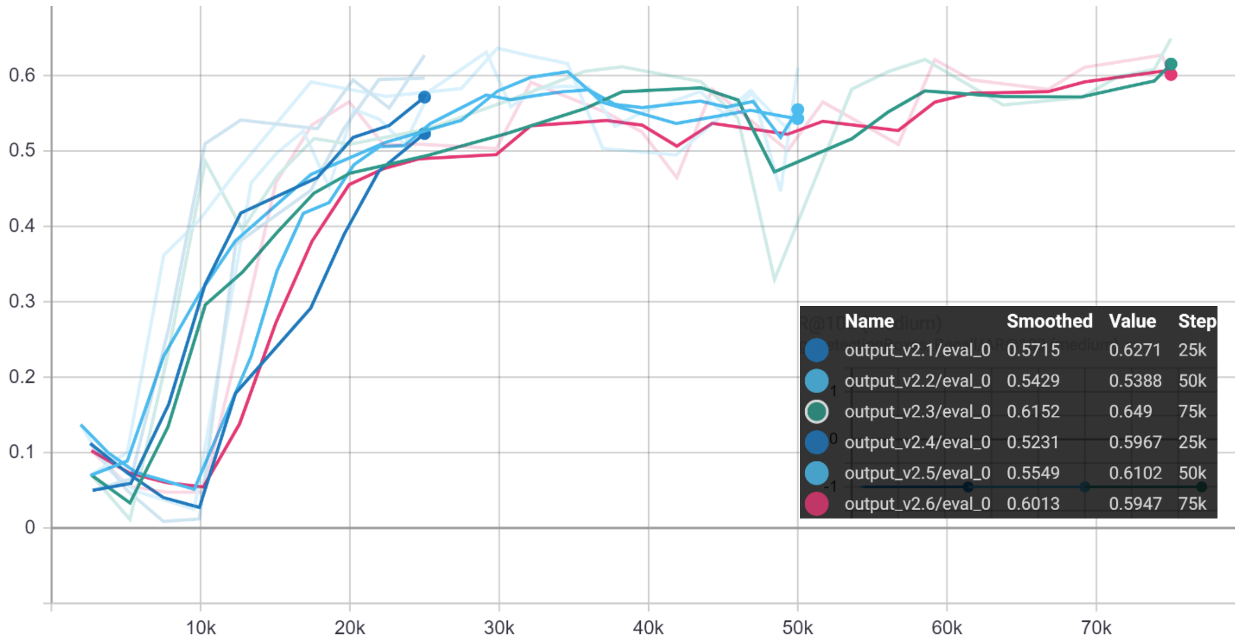
| **Step size** | 25,000 | 50,000 | 75,000 |
|---|---|---|---|
| **Loss** | 6.406 | 6.253 | 6.642 |
| **AR** | 59.67 | 61.02 | 59.47 |
| **mAP** | 45.25 | 54.67 | 50.09 |
| **mAP@.50IOU** | 78.88 | 80.55 | 75.41 |
| **mAP@.75IOU** | 48.37 | 67.71 | 62.51 |
| **(b) Learning Rate = 0.001** | | | |

**Table 1: SSD output values with respect to hyperparameter changes with (a) learning rate of 0.004 and (b) learning rate of 0.001**



| Name | Smoothed | Value | Step |
|---|---|---|---|
| output_v2.1/eval_0 | 0.4716 | 0.5403 | 25k |
| output_v2.2/eval_0 | 0.4949 | 0.4965 | 50k |
| output_v2.3/eval_0 | 0.557 | 0.5979 | 75k |
| output_v2.4/eval_0 | 0.3898 | 0.4525 | 25k |
| output_v2.5/eval_0 | 0.4889 | 0.5467 | 50k |
| output_v2.6/eval_0 | 0.5224 | 0.5009 | 75k |

**Figure 3: mAP values with respect to step size**



| Name | Smoothed | Value | Step |
|---|---|---|---|
| output_v2.1/eval_0 | 0.5715 | 0.6271 | 25k |
| output_v2.2/eval_0 | 0.5429 | 0.5388 | 50k |
| output_v2.3/eval_0 | 0.6152 | 0.649 | 75k |
| output_v2.4/eval_0 | 0.5231 | 0.5967 | 25k |
| output_v2.5/eval_0 | 0.5549 | 0.6102 | 50k |
| output_v2.6/eval_0 | 0.6013 | 0.5947 | 75k |

**Figure 4: AR values with respect to step size**

Observing the highest AR values in Figure 3 and Figure 4 alongside the mAP in Table 1 we can conclude that the highest accuracy is obtained when the learning rate is set to 0.004 and step size is set to 75,000 resulting in the lowest loss value in Table 1(a). Once training is completed the test set is used to validate performance and visual examples of the network in detecting the two classes of billboards Roadside and Street Furniture are shown in the Figure 5.

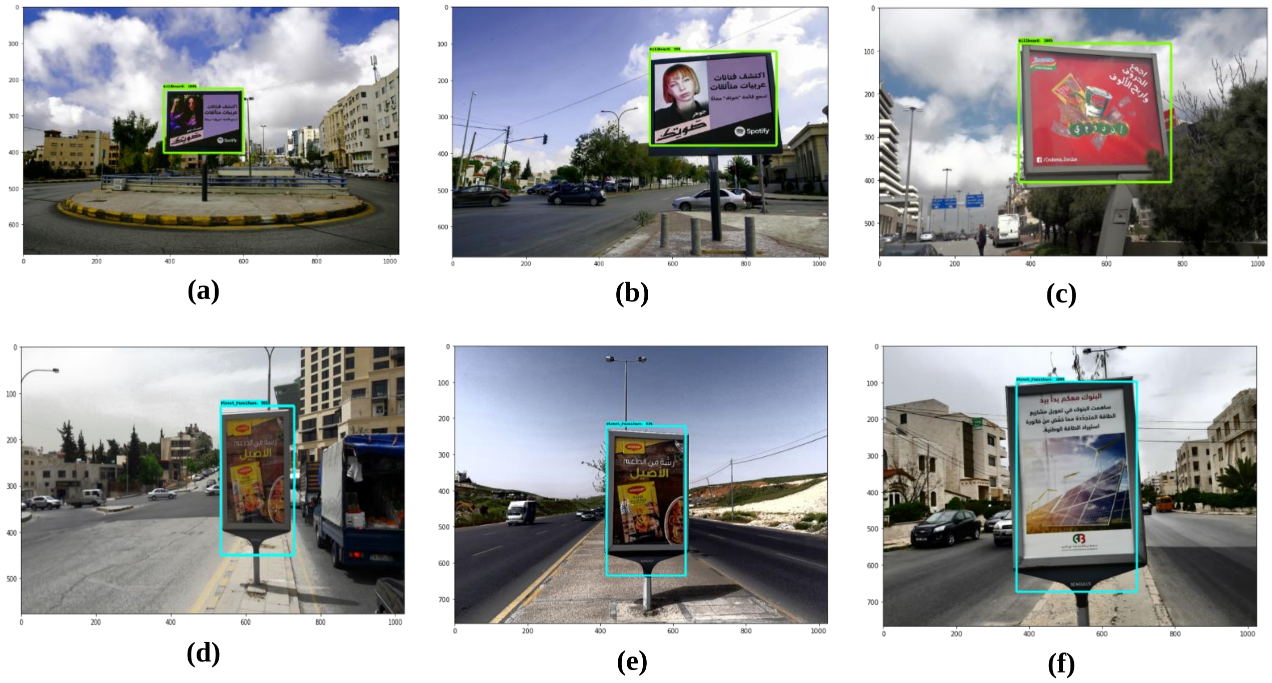


**Figure 5. Examples of the final network performance in detecting the two classes of billboards using unseen testing images: (a)-(c) Roadside Billboards; (d)-(f) Street Furniture Billboards**

## 5. Conclusion

This paper presents an approach to detecting advertising billboards in outdoor environments. Using transfer learning with SSD the outdoor billboards were successfully detected with 60% training accuracy. However, in testing there are some cases when billboards were not detected as either of the two classes resulting in missed detection. We are currently exploring increasing the training dataset size and augmenting the dataset with additional variations of billboards to improve the detection rate. We are also exploring the use of different segmentation masks other than bounding boxes to determine if they can improve detection performance. Additionally, we will consider the use of other deep learning architectures such as RCNN, RFCN and YOLO.

## Acknowledgements

We would like to express gratitude towards Digital Natives for providing the dataset and funding the PhD Scholarship at Ulster University.

## References

[Abadi, M. *et al.* 2016] Abadi, M. et al. (2016) ‘TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems’. http://arxiv.org/abs/1603.04467.

[Alom, M. Z. et al. 2018] Alom, M. Z. et al. (2018) ‘The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches’. http://arxiv.org/abs/1803.01164.

[Bochkarev, K. and Smirnov, E. 2019] Bochkarev, K. and Smirnov, E. (2019) ‘Detecting advertising on building façades with computer vision’, *Procedia Computer Science*, 156, pp. 338–346. doi: 10.1016/j.procs.2019.08.210.

[Borisova, O. and Martynova, A. 2017] Borisova, O. and Martynova, A. (2017) ‘Comparing the Effectiveness of Outdoor Advertising with Internet Advertising’, Jamk, (September), p. 85.

[Cai, G., Chen, L. and Li, J. 2003] Cai, G., Chen, L. and Li, J. (2003) ‘Billboard advertising detection in sport TV’, Proceedings - 7th International Symposium on Signal Processing and Its Applications, ISSPA 2003, 1, pp. 537–540. doi: 10.1109/ISSPA.2003.1224759.

[Hossari, M. et al. 2018] Hossari, M. et al. (2018) ‘ADNet: A deep network for detecting adverts’, CEUR Workshop Proceedings, 2259, pp. 45–53.

[Howard, A. G. *et al.* 2017] Howard, A. G. *et al.* (2017) ‘MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications’. http://arxiv.org/abs/1704.04861.

[Krishna, O. and Aizawa, K. 2018] Krishna, O. and Aizawa, K. (2018) ‘Billboard, Saliency Detection in Street Videos for Adults and Elderly|”, *2018 25th IEEE International Conference on Image Processing (ICIP),* pp. 2326–2330.

[Li, Y. et al. 2016] Li, Y. et al. (2016) ‘Nighttime lane markings recognition based on Canny detection and Hough transform’, 2016 IEEE International Conference on Real-Time Computing and Robotics, RCAR 2016, pp. 411–415. doi: 10.1109/RCAR.2016.7784064.

[Liang, P. et al. 2018] Liang, P. et al. (2018) ‘Planar object tracking in the wild: A benchmark’, Proceedings - IEEE International Conference on Robotics and Automation, pp. 651–658. doi: 10.1109/ICRA.2018.8461037.

[Liu, W. et al. 2016] Liu, W. et al. (2016) ‘SSD: Single shot multibox detector’, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9905 LNCS, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[Morera, Á. et al. 2020] Morera, Á. et al. (2020) ‘SSD vs. Yolo for detection of outdoor urban advertising panels under multiple variabilities’, Sensors (Switzerland), 20(16), pp. 1–23. doi: 10.3390/s20164587.

[Morera, Á. et al. (2019)] Morera, Á. et al. (2019) ‘Robust detection of outdoor urban advertising panels in static images’, Communications in Computer and Information Science, 1047(June), pp. 246–256. doi: 10.1007/978-3-030-24299-2_21.

[Rahmat, R. F. et al. 2019] Rahmat, R. F. et al. (2019) ‘Advertisement billboard detection and geotagging system with inductive transfer learning in deep convolutional neural network’, Telkomnika (Telecommunication Computing Electronics and Control), 17(5), pp. 2659–2666. doi: 10.12928/TELKOMNIKA.v17i5.11276.

[Rahmat, R. F. et al. 2019] Rahmat, R. F. et al. (2019) ‘Android-based automatic detection and measurement system of highway billboard for tax calculation in Indonesia’, Indonesian Journal of Electrical Engineering and Computer Science, 14(2), pp. 877–886. doi: 10.11591/ijeecs.v14.i2.pp877-886.

[Shi, W., Bao, S. and Tan, D. 2019] Shi, W., Bao, S. and Tan, D. (2019) ‘FFESSD: An accurate and efficient single-shot detector for target detection’, Applied Sciences (Switzerland), 9(20). doi: 10.3390/app9204276.

[Simonyan, K., & Zisserman, A. 2014] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[Watve, A. K. and Sural, S. 2007] Watve, A. K. and Sural, S. (2007) ‘Detection of on-field advertisement billboards from soccer telecasts’, pp. 12–17. doi: 10.1049/cp:20060494.

[Wilson, R. T. and Casper, J. 2016] Wilson, R. T. and Casper, J. (2016) ‘The role of location and visual saliency in capturing attention to outdoor advertising: How location attributes increase the likelihood for a driver to notice a billboard ad’, Journal of Advertising Research, 56(3), pp. 259–273. doi: 10.2501/JAR-2016-020.