Visual Pollution Detection Using Google Street View and YOLO

Md. Yearat Hossain, Ifran Rahman Nijhum, Abu Adnan Sadi, Md. Tazin Morshed Shad, Rashedur M. Rahman
Department of Electrical and Computer Engineering, North South University
Plot-15, Block-B, Bashundhara, Dhaka 1229, Bangladesh.
yearat.hossain@northsouth.edu, ifran.nijhum@northsouth.edu, abu.sadi05@northsouth.edu, tazin.shad@northsouth.edu, rashedur.rahman@northsouth.edu

***Abstract*—In recent years, visual pollution has become a major concern in rapidly rising cities. This research deals with detecting visual pollutants from the street images collected using Google Street View. For this experiment, we chose the streets of Dhaka, the capital city of Bangladesh, to build our image dataset, mainly because Dhaka was ranked recently as one the most polluted cities in the world. However, the methods shown in this study can be applied to images of any city around the world and would produce close to a similar output. Throughout this study, we tried to portray the possible utilisation of Google Street View in building datasets and how this data can be used to solve environmental pollution with the help of deep learning. The image dataset was created manually by taking screenshots from various angles of every street view with visual pollutants in the frame. The images were then manually annotated using CVAT and were fed into the model for training. For the detection, we have used the object detection model YOLOv5 to detect all the visual pollutants present in the image. Finally, we evaluated the results achieved from this study and gave direction of using the outcome from this study in different domains.**

***Keywords*—Visual Pollution, Deep Learning, Object Detection, YOLO, Google Street View, CVAT**

I. Introduction

As we strive on building a more modernised world each day, the impact of this modernisation leaves a negative footprint on the environment around us. Fast-growing cities worldwide are getting filled with unwanted visual objects. ”Visual Pollution” is a term frequently used to describe the pollution created by these visual objects, often termed ”Visual Pollutants” present in the environment around us. The domain of visual pollutants can include any objects unpleasant to our eyes, which may consist of ”Billboards”, ”Tangles of Electric Wires”, ”Street-Litters”, ”Construction-Materials”, ”Graffiti”, ”Cellphone Towers”, and ”Worn Out Buildings” [5].

As a consequence of these visual pollutants, historic cities are being harmed by the uncontrolled display of billboards and signage blocking views from all slides. This phenomenon may be seen in current metropolitan areas. The studies in [5] addresses this issue as visual pollution, which is a well-established term for the degradation of visual quality in areas with unwanted objects like billboards, signage, graffiti, etc. Although a common concern is growing regarding visual pollution, this remains a problem unsolved for many years due to the lack of tools and understanding of the causes. Understanding the devastating effect caused by visual pollution, we made an effort to show an approach to detect these visual pollutants from the street images collected from Google Street View, which in the long run will help organisations tackle this issue of identifying the areas with the highest visual pollutants and minimize visual pollution by taking appropriate actions.

To detect visual pollutants, we built an adequate sized image dataset that had visual pollutants in them. However, assembling a vast collection of images using a camera seemed unfeasible. Hence Google Street View [2] was used to build this moderate-sized image dataset. Google Street View is a visual representation of places on google maps consisting of millions of panoramic pictures [12]. The image dataset was constructed by taking screenshots from numerous angles of every street view that had visual pollutants in the frame. These images were then preprocessed and annotated using CVAT for training our object detection model.

In our experiment, we made an attempt to detect six types of visual pollutants: “Billboards”, “Bricks”, “Construction-Materials”, “Street-Litters”, “Communication Towers”, “Tangles of Electric Wires”. Given an image consisting of billboards, the object detection model can detect and locate the object in the image, drawing a box around the billboard. Object detection has gained a lot of research attention associated with video analysis and image interpretation. Even though there have been various studies on the introduction, management and classification of visual pollution [1], [15]-[19], our work introduces the concept of detecting visual pollution using object detection and Google Street View.

II. Related Works

A. Visual Pollution

There has not been much research work done before related to the classification of visual pollution using computer vision. However, in the paper [1], the authors introduced a method of classifying visual pollutants using a deep learning model. In this paper, four types of visual pollution classes were considered, and they used the google image search engine to collect their dataset. Their final dataset contained a total of 800 images. The paper proposed a Convolutional Neural Network (CNN) architecture as their deep learning model. In the end, they managed to achieve a decent training accuracy of 95% and a testing accuracy of 85% for visual pollution classification.

978-1-6654-0690-1/21/$31.00 ©2021 IEEE

In the paper [9], Visual pollution is described as a broad concept that encompasses limitations on the capacity to see distant elements, subjective issues of visual clutter, intrusive structures on stunning views, and other visual vandalism.

A wide range of literature on "Visual Pollution" was examined in the book [5] by Andriana Protella. In his book 'Life Between Buildings, Jan Gehl[3] talks about the relationship between outdoor visual quality and outdoor activity. Visual quality influences how people use city centres, how long particular activities persist, and which activity types grow. In his work 'Cities are Good for Us' (1990), Harley Sherlock [4] relates how high visual quality outdoors can promote activities outside. High visual quality encourages safe, responsible behaviour from those around us and can foster connection between citizens and local governments, resulting in a stronger feeling of community.

B. Google Street View

In paper [7], the authors showcased the usefulness of Google Street View for dataset collection and used that dataset to audit neighbourhood environments. This paper compared neighbourhood measures recorded from Street View images with the measurements recorded on field observation in a previous study. Data were collected for 143 items associated with seven neighbourhood environment constructions. In the final results, they found high levels of concordance for about 54.3% of the items. In the end, they came to the conclusion that Google Street View can be utilised to evaluate the neighbourhood environments.

C. Object Detection

A convolutional neural network (CNN) is a deep learning neural network that can process assembled data and extract core features out of that. CNN is used widely in deep neural networks, but it's primarily prominent for its usage in image classification, detection, etc. As we have previously mentioned in the paper, "A new nexus in environmental management" [1],introduced a method of classifying visual pollutants using a deep learning model. In the classification algorithm, we only classify images. The model takes the whole picture and predicts if it belongs to a particular class. For instance, in figure 1, we can see an example neural network which can classify if the given image is a "Construction Material" image or not.

However, object detection models can classify the object and detect the object's position in the picture by drawing a bounding box around the object. Some object detectors like Mask RCNN [14] can also generate segmentation masks that perfectly reflect the object for each instance. Modern object detection models can detect multiple instances of an object in an image.

Object detection is not a new concept. In 2001 Viola-Jones Detector real-time face detection framework was created. HOG detector, which was a famous person detector, came out in 2005. As an expansion of the HOG detector, P. Felzenzwalb introduced the DPM (Deformable Part-based Model) in 2008.

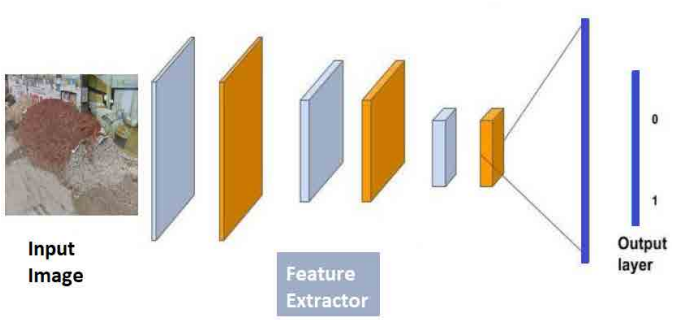

Fig. 1. An example neural network.

Then R. Girshick made various improvements. The majority of the early object detection algorithms used handcrafted features. After 2010 the performance of handcrafted features became saturated. [10]

In 2012, deep convolutional neural networks learned stable and high-level feature representations of an image for the first time. R.Girshick proposed the Regions with CNN features (RCNN) for object detection. R-CNN performed 50% better than the DPM algorithm in 2014. However, detecting a single image takes approximately 40 seconds. In 2014, K. He et al. proposed Spatial Pyramid Pooling Networks (SPPNet). SPPNet's essential addition sparked the rise of the Spatial Pyramid Pooling (SPP) layer. The SPP layer enables a convolutional neural network to develop a fixed-length representation independent of the image/region of interest size and does not require rescaling. The detection accuracy is the same as RCNN. However, it is 20% faster than RCNN.

R. Girshick proposed a Fast R-CNN detector the following year. It demonstrates ROI pooling which might be used to train a detector and a bounding box regressor simultaneously. On the VOC07 dataset, R-CNN's accuracy was 58.5%. Faster R-CNN, on the other hand, has a 70% accuracy rate. It was also 200 times faster than traditional R-CNN. Faster RCNN was proposed by S. Ren et al. in 2015, a bit later than Fast RCNN. It's the first end-to-end deep learning object detector and, as well as the first near-real-time deep learning object detector. Early detectors used selective search to find region proposals, which is a slow process. Faster R-CNN introduced the Region Proposal Network, and this broke the speed bottleneck of Fast R-CNN. R-FCN was later introduced, which included position-sensitive score maps, allowing R-FCN to detect faster [10]. R. Joseph et al. proposed YOLO [6], a new object detection model, in 2015. YOLO (You only look once) is an extremely fast model. As the name suggests, it only looks once. Previous detection algorithms used regions to localise objects in the image. In contrast, YOLO looks at the complete picture and detects, localise the object. YOLO can see multiple classes in a photo simultaneously, making it swift and broadly used for object detection. Yolo has multiple versions. In our

work, we used YOLO v5 [10]. Researchers around the world use YOLO for object detection. For example, In the paper [8], they collected and labelled 400 images of pavement cracking. After that, the dataset is used to train a YOLOv5 model. The results show that this network can successfully detect cracks, with an mAP of over 70% and a detection time of 152 milliseconds.

III. Dataset

A. Dataset Classes

There can be many objects which can cause a negative effect on a person's view, and all these objects are considered visual pollutants. These can be because of excessive use of billboards and signage, hanging wires, waste material on the sides of the streets, communication, grid towers, and many more. In this study, we prioritised the visual pollutants that are regularly seen on the streets of Dhaka. The visual pollutants focused in this research are: "Billboards", "Street Litters", "Construction Materials", " Bricks", "Wires", and "Towers".

1) Billboards: Billboards are large structures that are typically used for advertising purposes. Billboards are primarily found in urban areas with high traffic to be visible to a large number of people. This is also the case in Bangladesh, especially in Dhaka city. A considerable number of billboards can be seen on the side of roads, buildings, rooftops, and so on. Since there are no proper management laws, companies and organisations advertise their products by putting up as many billboards as possible.

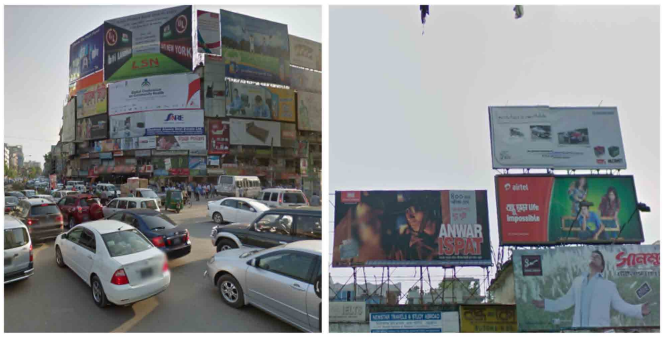

Fig. 2. Images of Billboards Collected from Street View

2) Street Litters: As Dhaka is not a well-developed city, it lacks a proper waste management system, and the people are not conscious about waste management either. As a result, piles of garbage lying on the sides of the streets is a prevalent scene in Dhaka. Since street litter can be seen almost everywhere around the city, it can be challenging to train a detection model to detect all kinds of street litters. Street litters can be of different shapes, sizes, and colours. So, in this study, we only tried to focus on medium to large piles of litter that can be seen on the sides of the streets.

Fig. 3. Images of Street Litters Collected from Street View

3) Construction Materials: Construction works of new buildings and roads are a very common sight in Dhaka. But because of the lack of space and proper management, these construction materials are placed on the sides of the roads almost all of the time. Some of these construction materials include sand, cement, broken bricks and stones etc. These materials sometimes even block half of the roads and pedestrian lanes. These things can cause traffic jams or even sometimes accidents because they are blocking the streets.

Fig. 4. Images of Construction Materials Collected from Street View

4) Wires: Dhaka does not have an underground electrical system where all the electrical wires are placed underground. Instead, overhead electrical wires connecting from one electric pole to another can be seen all around the country. It is mainly because it is very expensive to move all the electrical wires underground. Also, maintaining these cables can be really expensive too. Not only do these hanging wires cause visual pollution, but they are also a great risk to human lives as well. This problem has become more severe in the past couple of years as broadband internet is becoming more and more available than before. Wires of internet lines can be seen coiled up on top of electric poles everywhere.

5) Bricks: Despite the fact that bricks fall under the construction material category, we firmly believe that their unique shape, size and colour are easily differentiable from other construction materials. As a result, they were kept in a separate class.

6) Towers: There are mainly two kinds of towers that were focused on in this study: cell towers and electrical grid towers.

Fig. 5. Images of Wires Collected from Street View

Fig. 6. Images of Bricks Collected from Street View

To improve their communication service, big telecommunication companies in Dhaka installed a huge number of cell towers all over the country. As a result, cell towers can be seen very frequently on the side of the roads and on the rooftops of tall buildings. Electrical Grid towers are also a very common sight in Bangladesh. They can be generally seen on the sides of roads, riverbanks, open fields, and many other places all over the country.

Fig. 7. Images of Towers Collected from Street View

B. Dataset Collection

We were unable to locate a dataset on Dhaka's visual pollution. As a result, we had to make our dataset. Collecting image data by physically walking around Dhaka's streets and taking photographs seemed infeasible so, we explored "Google Street View" [2] as an alternative data collection tool for this study. Google Street View has a massive database of panoramic photographs of streets all over the world.

In order to collect the data from Google Street View, we tracked down different streets of Dhaka in Google Street View containing visual pollutants and collected screenshots from multiple angles. While collecting screenshots, many objects that were not visual pollutants were kept alongside our targeted objects because the model used for this research was a detection model and it needed to be able to differentiate between our targeted object and objects that were not causing visual pollution. Also, Google Street View provides a 360° view of the streets. So, we were able to take screenshots of the visual pollutants from multiple different angles and positions. This also helped the detection model to recognise visual pollutants from different angles and positions. This was one of the biggest advantages of using Google Street View for data collection.

C. Dataset Pre-processing

After collecting images using Google Street View, first, we needed to review our images and remove any images that were not relevant. Images that were of poor quality or were outliers were manually removed. Finally, a total of 1400 images were selected for the final dataset. Our training and testing set ratio were 80:20. So, a training set of 1120 images and a testing set of 280 images were selected. Table-1 shows the number of images that were selected for each class in our dataset. Here we can see that the "Street Litters" and the "Construction Materials" class have more images. We decided to include more images for these two classes so that we could improve the accuracy of the model. Because the objects of these classes do not have a definite shape, size and colour. As a result, the object detection model might underperform when given these types of data. For these reasons, the data of these classes were kept a bit higher than the rest. The images were resized to a fixed size in the final step because it's a common approach to feed detection model images of the same size. As a result, all the images in the dataset were resized into 500x500 pixels.

TABLE I
IMAGES COLLECTED PER CLASS

Class	***Image Count***
Billboards	200
Street Litters	300
Construction Materials	300
Bricks	200
Wires	200
Towers	200
total-	1400

D. Annotation

Image annotation is the process of labelling or classifying objects within an image which then can be used to train and test detection models. Images are needed to be carefully and precisely annotated because the accuracy of the detection

model depends on it. Thus more accurate annotation results in a more accurate model. One of the most common methods of image annotation is drawing rectangular bounding boxes around the target objects within the image. A bounding box tells the model exactly where within the image the object is located. But there are also other techniques of image annotation such as segmentation, landmarking, lines and splines etc.

We used the Computer Vision Annotation Tool (CVAT) [13] developed by Intel to annotate the images in our dataset. We used the rectangular bounding box method to annotate our images. If multiple visual pollutant classes were present within a single image, multiple bounding boxes were used to annotate those classes.

After annotating all the images, we exported the annotation data in YOLO format. In YOLO format, the annotations of an image are saved as a '.txt' file with the same name as the original image. Each line within the text file represents a bounding box annotation within the image. Each line contains the following information about a bounding box - class number, x-coordinate, y-coordinate, width and height.

Here, the x and y-coordinate, width and height values are normalised between 0 to 1. An example of YOLO annotation format is given below -

2 0.738400 0.676400 0.523200 0.202800

3 0.237300 0.740020 0.474600 0.230760

Annotation was one of the most critical parts of our research. Because the performance of the detection model depended directly on how accurately we annotated the images in our dataset.

Fig. 8. Images Annotated using CVAT

IV. Methodology

A. YOLO

YOLO is an object detection model introduced by Redmon J et al. [6]. Since its introduction, it has gained popularity due to its capability of fast real-time object detection. Unlike other object detection models, YOLO handles the image in a single passage over the model so it looks at the entire image only once. In the case of another popular object detection model Faster RCNN, an image is passed through various layers where the very first part is made of a CNN model, which would extract features maps from an input image. Then another network would propose and verify bounding boxes, and the last one would classify the object with bounding boxes. Still, in YOLO, the entire process happens once as YOLO handles the entire process in a single CNN. Hence YOLO achieves a much faster detection speed than any other model.

After its publication, YOLO has been widely used in various applications, and the same author published two other improved versions named YOLOv2 and YOLOv3. Later, different authors published their improved versions of YOLO named YOLOv4 and v5. YOLOv5 is the most advanced version of YOLO, and it achieves the fastest and most accurate results compared to its predecessors. YOLOv5 overcomes the limitations of the previous models in various ways. It comes with a feature called mosaic augmentation, which combines four images into four tiles which eventually helps to model to detect smaller objects. YOLOv5 comes in various sizes like small, medium, large,5x etc. The larger the model, the longer it takes to train and evaluate, but the accuracy will be higher on larger models. In our experiment, we only used YOLOv5 small and large variations.

B. Non-Max Suppression

In object detection tasks, a model performs classification and localisation at the same time. To localise an object in an image, a model can generate multiple bounding boxes of different dimensions. But naturally, we expect a single bounding box for each object which has the highest probability score for that object. In this case, the object detection model uses non-max suppression techniques to suppress the bounding boxes except the best one. This is actually performed using two things, the confidence score of the bounding box and the value of Intersection Over Union (IoU) of the bounding boxes. IoU is an evaluation metric used to measure the overlap between bounding boxes and calculated by comparing the ground truth label and predicted coordinates of the bounding boxes. Usually, IoU scores of more than 0.5 are considered a good prediction for bounding boxes.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area over Union}}$$

Fig. 9. Intersection Over Union

C. Evaluation Metrics

We evaluated our model with three evaluation metrics, which are mAP, precision and recall.

Mean Average Precision or mAP is used for measuring the accuracy of an object detection model. The greater the value of mAP, the better the model's performance. It is calculated by taking the mean value of average precision over all classes based on the IoU thresholds. In YOLOv5 average precision is calculated based on the threshold value of IoU as 0.5.

Precision demonstrates the number of correct positive predictions. On the other hand, Recall shows the number of correctly detected images out of all predicted detections. For example, we have a class in our study called "Construction Materials". Precision is out of all construction materials images how many of them our model got right. In contrast, recall means out of all images that were predicted as construction materials images, how many of them the model got right.

True Positive (TP) is the group of positive attributes that are accurately detected as positive attributes are known as true positive. On the other hand, True Negatives are the negative attributes that are correctly detected as negative attributes. For false positives and negatives, it's the opposite case. False-positive is the group of negative classes detected as positive attributes, and False negatives are the group of positive attributes detected as negative attributes.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

D. Transfer Learning

Transfer learning is a widespread and useful technique used in training machine learning and deep learning models. Training a model can be costly for the hardware because it requires a significant amount of computational resources. But using transfer learning, this issue can be solved. In transfer learning, a model is used which is previously trained using a different or similar type of dataset to perform similar kinds of tasks. In our research, we used YOLOv5 models that are pre-trained on the MS COCO dataset, which has 80 different classes of objects containing persons, bicycles, cars etc. As the models are pre-trained on the COCO dataset, they are already good at detecting various objects from an image. Using transfer learning, we trained our model to detect some newer objects as we defined in our dataset.

E. Experimental Setup

Our experimental setup consisted of a machine running on Intel Core i7 8700K, 16 GB of DDR4 memory, 6GB Nvidia RTX 2060, and Kubuntu as the operating system. In the case of software, Python and Pytorch were used for the pre-processing, training and testing process, CVAT was used for data annotation.

V. Results and Analysis

In our training, we kept the batch size 16 and ran the training session for 100 epochs. The dataset was split into an 80:20 ratio meaning 80% of the entire dataset went to the training part, and the rest 20% went to validation. We trained our dataset using transfer learning on small and large variations of the model where the models were pre-trained on the COCO dataset. We used SGD optimiser with a learning rate of 0.01 and a weight decay of 0.0005. After the 100th epoch, the small model achieved the mAP of 0.80, where the large version of the model gained 0.82(Table-2) as the best score. It took 15 minutes to train the small model and 39 minutes to train the large one using our setup. As the training finishes, the model saves both the last and best weights. Here, The mAP the model shows after each epoch is calculated based on the validation set.

TABLE II
YOLOv5 Small vs Large Model

Model	*Training Time*	*Best mAP*
YOLOv5s	15 minutes	0.80
YOLOv5l	39 minutes	0.82

We will show the precision, recall and mAP scores and graphs for the YOLOv5 large model for further evaluation. Table-3 showcases the precision, recall and mAP scores that we achieved using the YOLOv5 Large model. From table-3 we can see that the model achieved an mAP score of 0.59 on the first ten epochs. Then, from epoch 10-19, the mAP score slightly decreased to 0.53. After epoch 20-29, the mAP score jumped to 0.63 and right after epoch 30-39, the score significantly improved to 0.78. After epoch 40-49, the model achieved the mAP score of 0.82, which is the highest score the model achieved during the training. After that, the mAP decreased to 0.77 over the next 10 epochs. After epoch 60-69, the mAP again went up to 0.79. From epoch 70-79 and 80-89, the model gained the mAP score of 0.78. Finally, after the last 10 epochs, the model achieved a final mAP score of 0.79.

TABLE III
YOLOv5 Large Metrics

Epoch	*Precision*	*Recall*	*mAP*
0-9	0.50	0.66	0.59
10-19	0.58	0.57	0.53
20-29	0.64	0.62	0.62
30-39	0.74	0.75	0.78
40-49	0.78	0.78	0.82
50-59	0.77	0.77	0.77
60-69	0.80	0.77	0.79
70-79	0.79	0.76	0.78
80-89	0.79	0.78	0.78
90-99	0.83	0.78	0.79

The confusion matrix of our model is illustrated in figure 10. A confusion matrix helps to determine how well our model can predict the test images. This matrix gives us a deeper understanding of how our model is performing in each class.

The X-axis depicts the real values, and the Y-axis depicts "Predicted" values. For example, in the case of billboards, we can see the cell illustrates 0.82. This means our model could correctly predict 82% of the images, which were billboard images. In the same way, we can see for bricks, construction materials, street litter, towers and wires, the predictions are 0.89, 0.83, 0.72, 0.99 and 0.73, respectively.

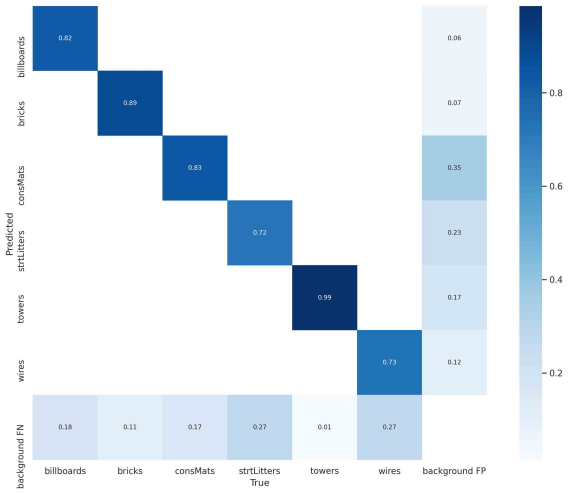

Fig. 10. Confusion Matrix

Figure 11 demonstrates the confidence vs precision graph. We can see the graph is upward sloping. This means the average precision level is increasing against confidence. In contrast, the recall curve, which is illustrated in figure 12, is downward sloping against confidence.

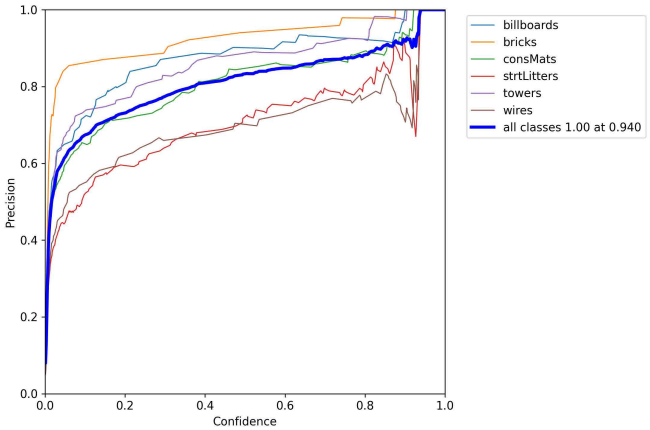

Fig. 11. Confidence vs Precision

Among all the classes of our dataset, bricks and towers performed the best, and street litters and wires performed average compared to other classes. This happened due to the shape, colour, and texture of the objects. The shape of towers and the texture and colour of bricks were significantly distinguishable from the other objects, which led the model to detect them with higher accuracy. On the other hand, the

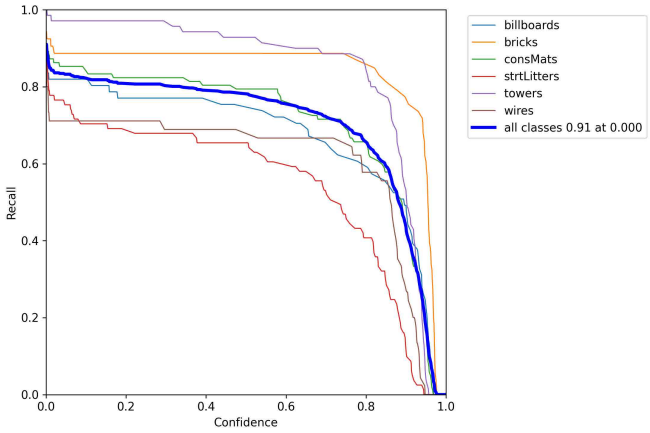

Fig. 12. Confidence vs Recall

model struggled a bit to detect street litters and wires due to the confusing shapes and textures of the objects. Fig-13 shows some of the predictions made by the model on the validation set.

Fig. 13. Validation Set Prediction Examples

VI. Conclusion and Future Work

With Dhaka being one of the world's most polluted cities, utilising Google Street View to hunt down visual pollutants from the city's streets was a relatively simple operation. However, we are confident that any city in the world can be utilised to create the image dataset and that the findings will be somewhat similar if the methodologies and strategies demonstrated throughout this study are meticulously followed. Furthermore, as visual pollution directly correlates with modernisation, we

attempted to present a modern technique by which this visual pollution problem can be minimised substantially.

Constructing the image dataset from scratch using only screenshots from Google Street View of Dhaka city proved yet again the potential of Google Street View and how it can be used as a tool in building image datasets. Throughout this experiment, we tried to emphasize the existence of visual pollution around us and also made an effort to show how these unwanted objects around us termed as "Visual Pollutants" can be detected using deep learning, which we believe in the long run help humanity minimize visual pollution from the environment around us. Future work may extend the volume of images allocated to each class in the image dataset. In addition, the variety of classes may also be enlarged, allowing for a more in-depth study of the subject. Moreover, with a more enhanced and broader dataset, the performance of the model is likely to improve. The approaches demonstrated in this study may be used by the government as well as any other entity interested in improving the visual quality of the city, resulting in a higher quality of life for the residents. The government or organizations may detect, collect, and store information on visual pollution from the streets automatically in real time using the methods shown in this research. This data may then be used to conduct additional analysis and identify areas of the city where visual pollution is notably significant. Additionally, a map with locations with high levels of visual pollution may be created, which can be used to alert citizens about affected regions, and these measures can lower the danger of visual pollution in the near future.

REFERENCES

[1] Ahmed, N., Islam, M., Tuba, A., Mahdy, M. and Sujauddin, M., 2019. Solving visual pollution with deep learning: A new nexus in environmental management. Journal of Environmental Management, 248, p.109253.

[2] Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L. and Weaver, J., 2010. Google Street View: Capturing the World at Street Level. Computer, 43(6), pp.32-38.

[3] Gehl, J. and Koch, J., 2011. Life between buildings. Washington: Island Press.

[4] Sherlock, H., 1990. Cities are good for us. London: Transport 2000.

[5] Portella, A., 2016. Visual Pollution: Advertising, Signage and Environmental Quality. London: Routledge.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[7] Rundle, A., Bader, M., Richards, C., Neckerman, K. and Teitler, J., 2011. Using Google Street View to Audit Neighborhood Environments. American Journal of Preventive Medicine, 40(1), pp.94-100.

[8] Shu, Z., Yan, Z. and Xu, X., 2021. Pavement Crack Detection Method of Street View Images Based on Deep Learning. Journal of Physics: Conference Series, 1952(2), p.022043.

[9] Sumartono, S., 2009. Visual Pollution in the Context of Conflicting Design Requirements. ITB Journal of Visual Art and Design, 3(2), pp.187-196.

[10] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," arXiv.org, 2019. https://arxiv.org/abs/1905.05055.

[11] V. Vikas Gupta Anastasia Murzova and A. Murzova, "Image Classification using CNNs in Keras," LearnOpenCV, 05-May-2021. [Online]. Available: https://learnopencv.com/image-classification-using-convolutional-neural-networks-in-keras/. [Accessed: 28-Oct-2021].

[12] Google Maps Street View. 2021. Discover Street View and contribute your own imagery to Google Maps.. [online] Available at: ¡https://www.google.com/streetview/¿ [Accessed 28 October 2021].

[13] Cvat.org. 2021. Computer Vision Annotation Tool. [online] Available at: ¡https://cvat.org/¿ [Accessed 28 October 2021].

[14] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN", 2018.

[15] Z. Kucharikova and J. Simko, "Visual pollution localization through crowdsourcing and visual similarity clustering," 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), 2017, pp. 26-31, doi: 10.1109/SMAP.2017.8022662.

[16] S. Chmielewski, "Chaos in motion: Measuring visual pollution with Tangential View Landscape Metrics," Land, vol. 9, no. 12, p. 515, 2020.

[17] K. Wakil, A. Tahir, M. Q. Hussnain, A. Waheed, and R. Nawaz, "Mitigating urban visual pollution through a multistakeholder spatial decision support system to optimize locational potential of billboards," ISPRS International Journal of Geo-Information, vol. 10, no. 2, p. 60, 2021.

[18] S. Chmielewski, "Towards managing visual pollution: A 3D ISOVIST and voxel approach to advertisement Billboard Visual Impact Assessment," ISPRS International Journal of Geo-Information, vol. 10, no. 10, p. 656, 2021.

[19] S. Chmielewski, D. J. Lee, P. Tompalski, T. J. Chmielewski, and P. Weżyk, "Measuring visual pollution by outdoor advertisements in an urban street using INTERVISIBILTY analysis and public surveys," International Journal of Geographical Information Science, vol. 30, no. 4, pp. 801–818, 2015.

[20] S. Sumartono, "Visual pollution in the context of conflicting design requirements," ITB Journal of Visual Art and Design, vol. 3, no. 2, pp. 187–196, 2009.