

UBDC Technical Note 2022/3 (November 2022)

Building a high-quality annotated image library to improve object detection

Authors: Luís Serra (University of Glasgow)

Maralbek Zeinullin (University of Glasgow)

UBDC Technical Notes and Working Papers Series DOI:
10.5281/zenodo.7271768

Building a high-quality annotated image library to improve object detection

Computer Vision

Luís Serra
and
Maralbek Zeinullin

Project Report



October 21, 2022

Preface

This work was developed at the **Urban Big Data Centre**(UBDC) from the University of Glasgow, and funded by **Economic and Social Research Council** with the grant number ES/L011921/1. Any questions regarding this report should be directed to Luís Serra with the email `luis.serra@glasgow.ac.uk` or Maralbek Zeinullin with the email `maralbek.zeinullin@glasgow.ac.uk`.

1 Summary

Modern cities worldwide are facing considerable pressures to improve city attractiveness to dwellers and businesses. A key factor to raise the liveability of cities is the existence of a good network of walkways and cycleways. To implement such infrastructures under growing resource constraints, cities need to know beforehand how people are using public spaces. One of the best ways to study people behaviour in cities is to make use of CCTV systems with the help of computer vision models. This project aims to collect a set of CCTV-like images from four city centres and annotate persons, cyclists and vehicles with the purpose of: developing an object detection model to be deployed on CCTV cameras; and building a shareable repository of annotated images for use by other developers. The collection of images were captured between January and February 2022 with a high definition camera similar to the type used in CCTV systems. The annotation work started in January 2022 and ended in March 2022. The annotation comprised of a team of five annotators and two reviewers working with specialized software. The project annotated 99,246 unique objects in 10,446 images. The most annotated object class was “Pedestrian” with 81.9% of the total number of annotated objects.

2 Introduction

Cities around the world are turning to the Internet of Things (IoT) and smart city technology to better understand how people are using public spaces. This is especially true for city planners wishing to improve city attractiveness to dwellers and businesses. However, city planners are increasingly facing a huge pressure: on the one hand they need to raise the livability of a city, but on the other hand they must achieve this aim under growing resource constraints.

CCTV cameras are probably one of the widest known - and used - IoT devices. CCTV infrastructure in cities is usually used for community safety and crime prevention. These cameras are typically not active for long periods, thereby creating a window of resource availability where cameras can be used to gather data about the way people use public spaces.

Transport researchers from the Urban Big Data Centre (UBDC) have been carrying out research into active forms of travel and the possible benefits available to cities. Reduced congestion, reduced air pollution, healthier and happier populations, reduced demands on public spending and the creation of attractive environments for private investment [4] are all associated with increased active travel and reduced reliance on car usage. CCTV cameras offer opportunities to measure the volumes of different forms of travel, which can inform evaluations of investments designed to promote active travel, by providing imagery that can be processed to yield data about use of public space. This is achieved by using machine learning object detection technology. Unfortunately, Tensorflow [8] object detection models that we have evaluated, are unable of detecting cyclists. While some may be able to detect a bicycle or a person separately, they are incapable of distinguishing the specific case of a person riding a bike. Furthermore, off-the-shelf models were usually trained with photographs captured by pedestrians - rather than by mounted CCTV camera - with a lower point of view and inconsistent image quality. As a result, when faced with relatively unfamiliar CCTV imagery, these models perform less effectively. A new object detection model needed to be created to detect cyclists on images captured from CCTV cameras. A related and important consideration to achieve suitably high levels of performance is the volume of training data. Generally speaking, more labelled examples of imagery of interest will yield better quality models, but ultimately, it all depends on the complexity of the elements being detected.

Motivated by these circumstances, the objective of this project was to annotate persons, cyclists, and vehicles on at least 10,000 CCTV-like images, collected from the city centres of Glasgow, Newcastle, Manchester, and Sheffield. Those cities were selected as a result of their relative familiarity for UBDC researchers and the likelihood of being able to capture suitably high volumes of imagery of objects of interest. These included buses, cars, cyclists, crowds, lorries, motorcycles, pedestrians, taxis and vans - examples of each would be explicitly labelled as such on resulting imagery. The annotation dataset produced would provide a means to train and evaluate computer vision models that could ultimately be deployed within CCTV camera networks to capture data about active - and other - forms of travel. A second but equally important objective was to produce and make available to the wider academic research community a repository of annotated images of public space usage. These can in turn be used in the development and evaluation of other machine learning models. It is also anticipated that this resource can be expanded and enhanced with further imagery and/or annotated to support additional use cases. This latter objective is particularly important given the challenges of acquiring such data - particularly as a result of personal data sensitivities associated with imagery captured within public spaces.

Capturing imagery from several cities provides a broad spectrum of environments which will facilitate the development of more generalisable computer vision models, especially across cities in the UK. A rule of thumb when collecting machine learning (ML) data is to gather data that is representative of the entire range of inputs for which one aims to develop the predictive model. Models trained exclusively on Glasgow imagery would probably be biased towards specific characteristics of Glasgow.

3 Methods and Procedures

This section consists of three parts: Data capture, annotation software and annotation procedures.

The annotation work started on the 5th of January 2022 and ended on the 31st of March 2022. The Annotation Team was composed of five annotators and two coordinators. The five annotators were all MSc students, most studying at the University of Glasgow while the two coordinators are Data Scientists

at the Urban Big Data Centre, also at the University of Glasgow - and the authors of this technical summary.

3.1 Data capture

Videos for annotations were collected by [CTS Traffic and Transportation Ltd.](#), following their success within a University of Glasgow tender process. The collection occurred between 01/02/2022 and 19/03/2022 in Glasgow, Manchester, Sheffield and Newcastle. In each city, efforts were taken to locate sites and environments expected to yield high volumes of objects of interest: Cyclists on a cycle path, cyclists on the road, pedestrianised streets, streets with pedestrians and vehicles, streets mostly with vehicles. [Strava](#) data was referenced to identify specific locations associated with high volumes of active cyclists. A general convention for image capture was established whereby multiple cameras were positioned at each site at varying locations in order to capture objects of interest from different perspectives. These included views designed to capture objects moving towards, away from and across the field of view of cameras. Following deployment, each camera recorded 14 consecutive hours of video, between 6am and 8pm. The videos were produced in RGB colour mode, with a spatial resolution of 1280 x 720 pixels and a frame rate of 25 frames per second. The video format used was *MP4* with the codec *MPEG-4 AVC/H.264*. All cameras were pole-mounted at a height between 3 and 3.5 metres.

In Glasgow the Annotation Team labelled images from five cameras located in four sites, as shown in figure 1.

In Manchester, the Annotation Team labelled images from four cameras located in four sites, as shown in figure 2.

In Newcastle the Annotation Team labelled images from one camera located in one site, as shown in figure 3.

In Sheffield, the Annotation Team labelled images from one camera located in one site, as shown in figure 4.



Figure 1: Location of sites surveyed in Glasgow: 1. junction of Lancefield Quay with Finnieston Street, 2. junction of Broomielaw and King George V Bridge, 3. junction of Argyle Street with Queen Street, 4. junction of Argyle Street with Glassford Street and 5. junction of Wilson Street and Candleriggs.

3.2 Annotation software

In complex machine learning projects, images are generally annotated using dedicated annotation platforms. After comparing several annotation tools available in the market, we decided to use the **CVAT** tool (**C**omputer **V**ision **A**nnotation **T**ool). CVAT is an open-source browser-based application for annotating digital images and videos. It is especially designed to support collaborative work, providing a set of convenient tools to organize, annotate and review images.

Regarding the internal structure, CVAT is a **Docker** microservice application. The system includes several Docker containers¹ running different tasks such as storing data, creating annotation tasks and managing users.

¹A Docker container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another[1]

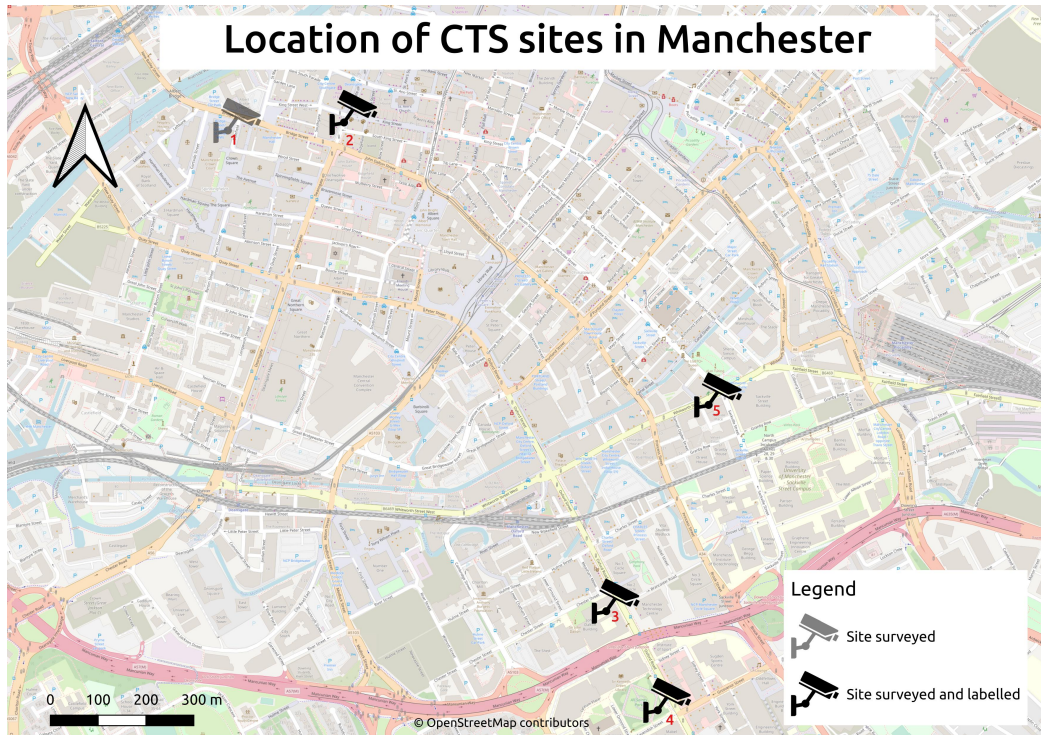


Figure 2: Location of sites surveyed in Manchester: 1. Junction of Bridge Street with St Mary’s Parsonage, 2. junction of King Street West with Deansgate, 3. junction of Oxford road with Chester Street, 4. Oxford road near Grosvenor Street and 5. junction of Whitworth Street with Sackville Street.

3.3 Annotation procedures

Although the original imagery was captured in a video format, the annotation procedure involves annotation of still image data. Therefore, we were required to extract frames from the videos recorded to a still image format. To ensure suitable diversity of imagery and objects, the frames were randomly selected during varying hours and weather conditions in jpeg image format.

Although time-consuming, image annotation may seem to be a straightforward and routine process. However, early in the project, several situations arose where the choice of how or where to apply labels was not clear or involved an element of subjective interpretation. How should persons that are not pedestrians be labelled? How should a motorbike rider or pillion passenger be handled? This was particularly problematic where labelling was being done by several individuals, each with their own assumptions about

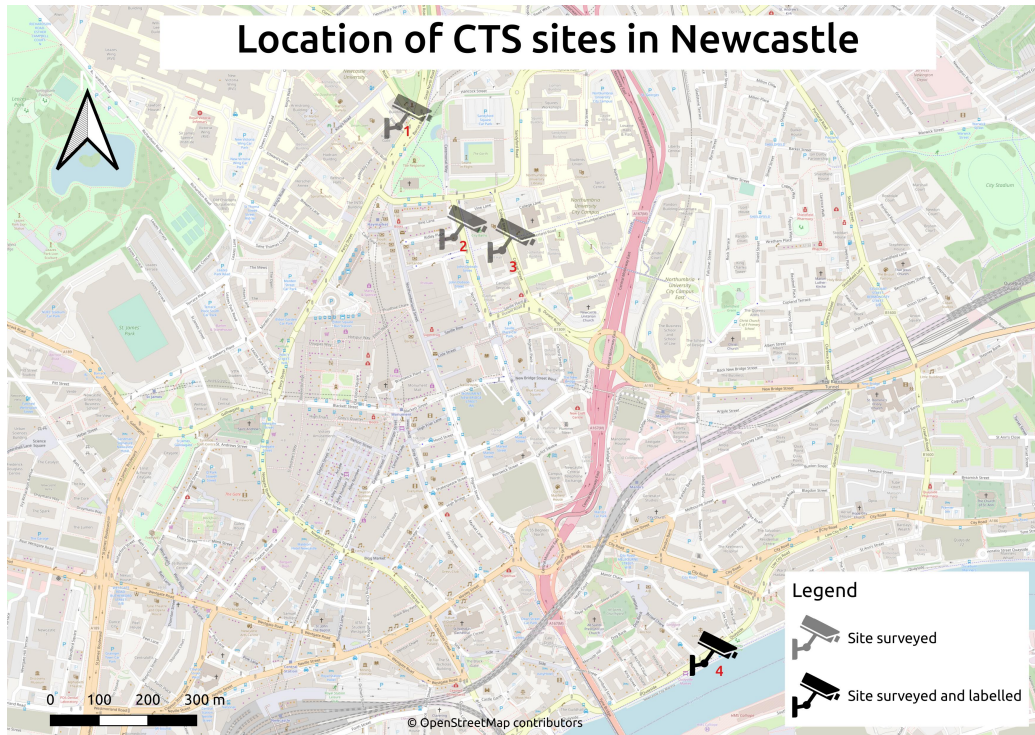


Figure 3: Location of sites surveyed in Newcastle: 1. Junction of Great North Road with Claremont Road, 2. John Dobson Street, between St Mary’s PI and Northumberland Road, 3. junction between Northumberland Road and College Street, 4. junction between Quayside and Broad Chare.

how to approach such situations, and therefore introduce inconsistency into the training dataset. To resolve these uncertainties, the team agreed on a set of conventions to be applied uniformly:

- A person pushing a bike by hand should be considered a cyclist and labelled as such.
- All visible persons labelled as “pedestrians”, with the exception of cyclists.
- The label “crowd” should be used sparingly and only in cases of large and compact gatherings of people where it was extremely difficult to distinguish individuals.
- A vehicle should be labelled as a “taxi” if the typical curved shape of a UK hackney-style taxi was clearly recognisable, even when the yellow sign on top (or the word “taxi”) was not visible.

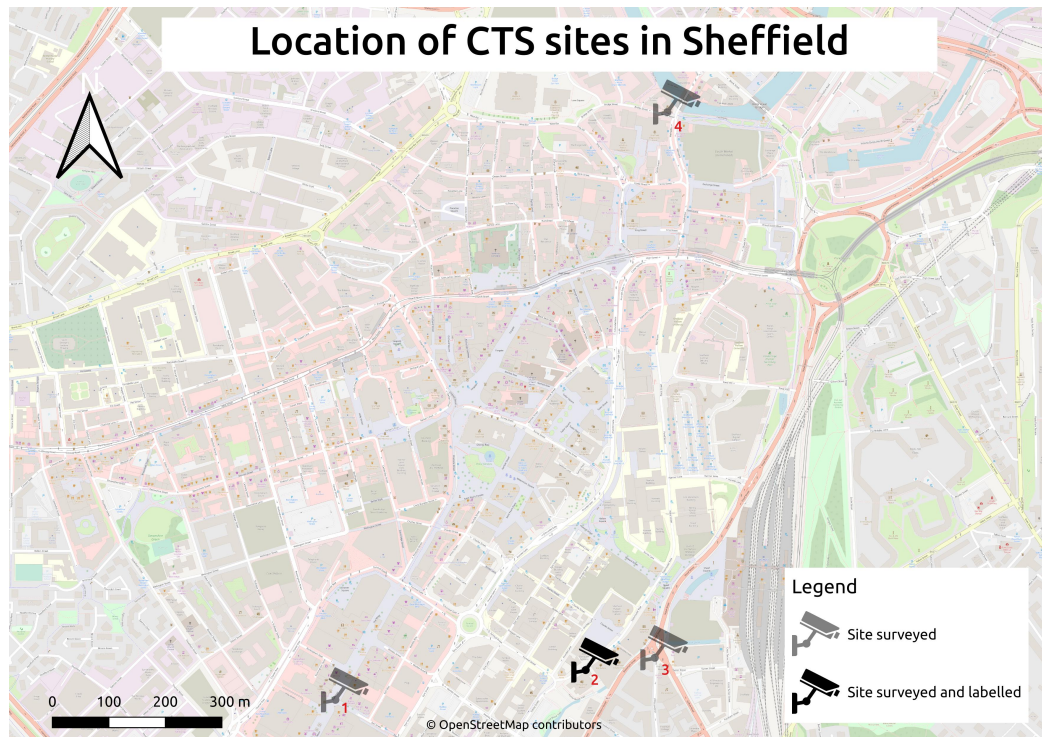


Figure 4: Location of sites surveyed in Sheffield: 1. The Moor, between Rockingham Gate and Earl Street, 2. Brown Street, near Arundel Lane, 3. junction between Shoreham Street, Suffolk Road and Sheaf Street.

Although labelling conventions were important to improve consistency, labelling variations continued to exist among annotators due to factors that include differences in visual acuity; differences in labelling decisions (especially regarding faraway objects not clearly visible); and user errors as a result of tiredness and the inherent repetitiveness of the annotation task. To further improve consistency and minimise errors, a series of procedures and good practices were established:

- Annotation training to take place during the initial stage of the project
- Following the training, labellers would annotate a common set of images and reflect on choices made during subsequent group discussion.
- Any meaningful issues encountered to be discussed and resolved, ensuring a common understanding of how to address future cases.
- Creation of a dedicated Microsoft Teams page to discuss issues and good practices as well as to share diverse documents (*e.g.* classification

of vehicles in appendix B.)

- Maximum 6h of work per day
- Enforced 10 min break for each working hour
- The whole team (labellers and reviewers) would meetup once per week to work in the same room together in order to more easily discuss labelling issues and reflect on decisions made
- Each finished annotation task - or every single label applied by the annotators - to be reviewed by the review team, resulting in one of two outcomes: **Accept**, when no issues were identified, and **Reject**, when at least one issue was present. The latter would require the annotation task to be returned to the annotators to correct the mislabelling(s).

In addition to image representativeness and labelling consistency, two further key factors were considered to achieving a high-quality standard on the annotated dataset:

- **Completeness** - meaning all objects of interest within an image were detected and localised, thus minimising the number of false positives and false negatives.
- **Positional accuracy** - meaning the drawn bounding boxes enclosed the entirety of the object, but without being too tight to the extent of cutting off a portion of the object, nor too large leaving large amounts of unrelated pixels around the object of interest.

Consistent with the project's Data Protection Impact Assessment strategy, the images were only available on the UBDC intranet and only to the Annotation Team members.

4 Results

The project labelled 99,246 unique objects in 10,446 images. An example of a labelled image can be found in Appendix A.

Table 1: Number of labelled objects per class.

Object	Count
Bus	3,241
Car	9,697
Cyclist	1,937
Crowd	645
Lorry	318
Motorcycle	52
Pedestrian	81,336
Taxi	387
Van	1,633
Total	99,246

The label “Pedestrian” comprises almost 82% of the total number of labelled objects, according to the pie chart in figure 5. Despite a great effort to find city locations and hours of the day where the likelihood to find cyclists was high, this label contributes to less than 2% of the total number of labels.

5 Discussion

We gathered image data from a diverse range of every day life in cities using CCTV-like cameras. The diversity of city environments cover most of the situations we wished to detect city activity, especially active forms of travel.

The fact that we collected imagery from CCTV-like cameras with similar attributes to CCTV cameras operating in cities, will certainly contribute to develop better models to be deployed in these environments.

The goal to reach 10,000 labelled images was achieved and even surpassed, but at the expenses of a great class imbalance, largely dominated by the label *pedestrian* with 81.9% of the total number of labels in the annotated dataset (figure 5). Although class imbalance is not problematic when less represented classes occupy a large area of the images, such as the classes of lorries and vans, it can be problematic with objects which tend to occupy smaller areas, such as cyclists and motorcyclists, because object detectors perform better at detecting large objects[3]. More annotated data is usually needed to resolve performance issues when detecting physically smaller object types. Despite class imbalance and with the exception of motorcyclists, we believe that we

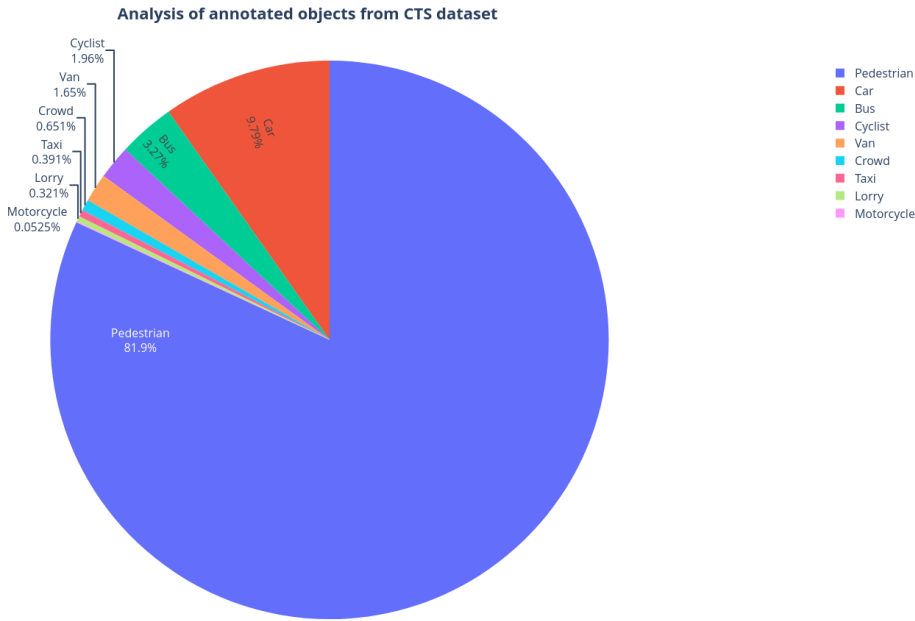


Figure 5: Proportion of each object class in the annotated dataset.

achieved a sufficient number of labelled examples for the remaining classes to successfully train an object detection model.

Another reason to have labelled 10,000 images was to have a robust number of correctly annotated examples that could minimise the negative impact caused by random mislabelling². Some studies suggest a linear correlation between class noise and test accuracy, where an additional 10% of noise leads to 4% reduction in accuracy [2]. To reduce class noise, all annotated images were reviewed.

Great effort was made to gather sufficient examples of cyclists in a diverse range of situations such as riding a bike, pushing a bike by hand, and holding bikes in a stationary manner. Our convention requires a person to be at least holding a bike to be considered a cyclist and labelled as such. However, when training a model, it may be challenging to train a model to effectively distinguish between a pedestrian passing near a parked bike and a “true” cyclist.

²Random mislabelling, also known as unbiased mislabelling, happens when an object class is accidentally replaced by another class.

A decision was made not to add an additional label of *occlusion* to partially visible objects - primarily to ensure the quicker completion of the labelling task . Occluded object reduce the performance of detectors [5], especially one-stage detectors³, such as Yolo [6]. In order to improve detection of partially occluded objects, it is important that occlusion information is available in the annotation dataset [7]. The addition of the label *occlusion* to occluded objects, can be considered possible future work, and is a potential opportunity for other researchers accessing the labelled dataset via the UBDC Data Service.

Finally, regarding the quality of the labelling, we believe that we took all the necessary steps to minimize mislabelled objects and labelling errors in general, but ultimately we are conscious that errors still exist simply because the annotation work was performed by humans, which are inherently prone to errors. Furthermore, there were several situations of uncertainty - particularly when the objects to label were not clearly visible. One of those situations, later identified, was related with the class label “crowd”. This label was mostly used in the faraway sections of pedestrianised streets where the individuals gathered together were not easily distinguishable. The problem arose with the difficulty of the annotators to establish similar bounding box boundaries, when faced with similar gathering situations. This problem of inconsistency, was later reflected in the difficulty of object detection models - trained with this data - to detect large gatherings.

6 Conclusions and Recommendations

Collecting and annotating a diverse range of images that represent most city environments in the UK, while minimising the likelihood of mislabelled data, has been a huge effort.

Apart from motorcyclists, all the other classes of objects are well represented in the annotated dataset, allowing us - and other users of the dataset - to develop and evaluate object detection models within CCTV infrastructure.

Annotation is a very repetitive task and the more that can be done to minimize repetitiveness, the more interesting the work becomes and the fewer labelling errors are made. One procedure to minimize repetitiveness is to

³Region proposal and object classification is done in one step.

allocate a sequence of tasks⁴ to label within different environments and with differing image volumes. The size of the task is dependent on the average number of objects to label per image: the task should be smaller when there are many objects per image. On the contrary, the size of the task can be higher if there are fewer objects per image. For instance, pedestrianised streets at rush hours are usually very crowded and the size of a task made up with such images should not exceed 100.

Other types of annotations can be explored in future, such as *Semantic Segmentation* to get the most of the data in the images. *Semantic Segmentation* aims to classify each pixel on the image that belongs to an object of interest, if not to classify all pixels. An image with segmentation masks allows the training and experimenting with all types of machine learning models, e.g., *foreground/background separation*. Furthermore, segmentation masks are more precise than bounding boxes as they cover only the location of the actual object. Bounding boxes - as relatively coarse rectangle shaped polygons - often include neighbouring regions or intersect with other bounding boxes. However, some of the drawbacks are that segmentation masks are time consuming to annotate by hand or to correct annotation errors, e.g. pixels which are mislabelled.

In addition to use within its own CCTV automation projects, UBDC will make these annotated images available to other academic developers wanting to develop or validate their own models. We hope the labelled dataset produced constitutes an invaluable resource to develop better models for detection of objects within imagery produced within CCTV systems.

⁴In this project, a “task” is a set of images collected from the same camera which are allocated to a labeller.

References

- [1] Docker. Use containers to build, share and run your applications. <https://www.docker.com/resources/what-container/>, 2022. (accessed on August 2022).
- [2] David Flatow and Daniel Penner. On the robustness of convnets to training on noisy labels. *Technical report, Stanford University*, 2017.
- [3] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.
- [4] Philip Insall. Active travel in the city of the future. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/367499/future-cities-active-travel.pdf, 2014. (accessed on June 2022).
- [5] Adam Kortylewski, Qing Liu, Angtian Wang, Yihong Sun, and Alan Yuille. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *International Journal of Computer Vision*, 129(3):736–760, 2021.
- [6] Alberto Rizzoli. The ultimate guide to object detection. <https://www.v7labs.com/blog/object-detection-guide>, 2022. (accessed on June 2022).
- [7] Kaziwa Saleh, Sándor Szénási, and Zoltán Vámosy. Occlusion handling in generic object detection: A review. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000477–000484. IEEE, 2021.
- [8] Tensorflow. Tensorflow 2 detection model zoo. https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md, 2021. (accessed on June 2022).

A Example of annotated image




Figure 6: Example of annotated image captured in Manchester. Notice bounding boxes with different colours. Each object class has a unique colour to help minimize mislabelling between classes. Sketch of image to protect privacy.

B Differences between different types of vehicles




Difference between van, lorry and bus

Van - a smaller boxlike vehicle that resembles a panel truck, often has double doors both at the rear and along the curb side, fitted with rows of seats, or equipped with living quarters for traveling and camping.

Source: <https://commercialvehiclecontracts.co.uk/faq/choosing-your-vehicle/van-size-guide>



VAN SIZE GUIDE

		LOAD LENGTH	LOAD WIDTH	LOAD HEIGHT	PAYLOAD
	MICRO VAN	1.3m / 4.2ft	1.2m / 3.9ft	0.8m / 2.6ft	500-530kg
	SMALL VAN	1.2m / 3.9ft	1.5m / 4.9ft	1.2m / 3.9ft	600-1000kg
	PICKUP	1.5m / 4.9ft	1.5m / 4.9ft	0.8m / 2.6ft	1000-1200kg
	MEDIUM VAN	2.4m / 7.8ft	1.7m / 5.5ft	1.4m / 4.5ft	900-1200kg
	CREW CAB / MINIBUS	1.5m / 4.9ft	1.6m / 5.2ft	1.4m / 4.5ft	800-900kg
	DROPSIDE / TIPPER	4m / 13.1ft	2m / 6.5ft	N/A	1000-1300kg
	LARGE VAN	3.4m / 11.1ft	1.7m / 5.5ft	1.7m / 5.5ft	1200-1500kg
	LUTON VAN	4m / 13.1ft	2m / 6.5ft	2.2m / 7.2ft	1200-1600kg

Source: <https://www.vandemon.co.uk/blog/article/van-makes-models-in-different-industries/>

POPULAR Industries and the VANS they use...



Most common van used by Ambulances:
Mercedes Sprinter



Most common van used by British Gas:
Volkswagen Caddy



Most common van used by Royal Mail:
Peugeot Partner



Most common van used by Ice Cream vans:
Mercedes Sprinter (chassis)

Lorry (US: Truck) is the largest and may also be called an articulated lorry or a heavy goods vehicle (HGV). These normally only travel on major roads and carry the largest goods. Usually, they have flat fronts.

Source: <https://www.returnloads.net/how-to-price-haulage-work/>



Buses

While they vary in size and shape, all pictured below are buses and should be labelled as such.

Source: https://en.wikipedia.org/wiki/First_Glasgow#/media/File:First_Glasgow_bus_SF07_FDP.jpg



Source: <https://www.heraldsotland.com/news/14534239.glasgow-bus-company-city-sprinter-operated-city-streets-without-insurance-two-months/>



Source: https://en.wikipedia.org/wiki/Glasgow_Citybus#/media/File:16-11-16-Glasgow_street_scene-RR2_7267.jpg



Source: <https://www.dailyrecord.co.uk/ayrshire/stagecoach-cancels-more-100-services-24958078>



Source: <https://www.flickr.com/photos/77000628@N02/33070448658>

