Weakly Supervised Training of Monocular 3D Object Detectors Using Wide Baseline Multi-view Traffic Camera Data

Matthew Howe[12]
matthew.howe@adelaide.edu.au
Ian Reid[1]
ian.reid@adelaide.edu.au
Jamie Mackenzie[2]
jamie.mackenzie@adelaide.edu.au

[1] Australian Institute for Machine Learning
University of Adelaide
Adelaide, Australia

[2] Centre for Automotive Safety Research
University of Adelaide
Adelaide, Australia

Abstract

Accurate 7DoF prediction of vehicles at an intersection is an important task for assessing potential conflicts between road users. In principle, this could be achieved by a single camera system that is capable of detecting the pose of each vehicle but this would require a large, accurately labelled dataset from which to train the detector. Although large vehicle pose datasets exist (ostensibly developed for autonomous vehicles), we find training on these datasets inadequate. These datasets contain images from a ground level viewpoint, whereas an ideal view for intersection observation would be elevated higher above the road surface. We develop an alternative approach using a weakly supervised method of fine tuning 3D object detectors for traffic observation cameras; showing in the process that large existing autonomous vehicle datasets can be leveraged for pre-training. To fine-tune the monocular 3D object detector, our method utilises multiple 2D detections from overlapping, wide-baseline views and a loss that encodes the subjacent geometric consistency. Our method achieves vehicle 7DoF pose prediction accuracy on our dataset comparable to the top performing monocular 3D object detectors on autonomous vehicle datasets. We present our training methodology, multi-view reprojection loss, and dataset.

1 Introduction

The majority of collisions between road users in urban areas occur at intersections. Visual surveillance could have a significant impact on providing tools to understand conflicts between road users. This work presents a model to infer location, size, and yaw rotation (7DoF pose) of vehicles from monocular images. This will enable road safety researchers to build 3D models of traffic flow and understand the severity, likelihood, and frequency of conflicts to enable better design and flow-control of intersections.

Current methods of training models to collect road users' 7DoF pose rely on expensive sensors such as LiDAR and large amounts of hand labelled data. For these reasons, existing methods are not cost effective when performing multi-intersection studies; a common method for intersection safety evaluations. It is becoming increasingly frequent for

road safety researchers to utilise monocular cameras to collect road user behaviour data [13, 17, 28, 36, 39]. Since 3D object detectors for road safety are not readily available methods utilise ground plane calibration and off the shelf 2D object detectors to approximately localise vehicles passing through intersections. The ability to easily collect precise location, rotation, and space occupancy data of road users is invaluable for accurate measurements of safety and will open the door to a broader application of 3D vision in road safety.

To utilise the simplicity and low-cost of monocular video collection, an object detector must be able to infer the 7DoF pose of vehicles from a frame. Typically, a monocular 3D object detector (mono3DOD) for vehicles is trained on autonomous vehicle data which contains thousands of hand annotated 3D objects that utilise LiDAR to accurately label 7DoF poses. There is a domain gap between ego vehicle data and intersection observation cameras. This is due to the perspective difference between cameras mounted at vehicle roof height versus elevated above the road surface. While datasets from traffic observation viewpoints exist, they do not contain LiDAR data or 3D object annotations that would be needed to train a mono3DODs conventionally.

In this paper we develop a method which is able to utilise inexpensive sensors and requires no additional manual labelling to train a mono3DOD. We leverage an existing (pretrained) mono3DOD designed for ground-level detection, and show how this can be fine-tuned in a weakly supervised manner using multi-view camera data. Using our methodology, road safety researchers can fine-tune a model for their camera setup through the temporary installation of additional overlapping cameras for 10 minutes and achieve 7DoF pose prediction accuracy that is comparable to fully supervised models.

Summary of contributions

1. We develop a multi-view reprojection loss to train 3D object detectors in a weakly supervised fashion to an accuracy comparable to top performing mono3DODs.

2. We provide the wide baseline multi-view (WIBAM) dataset. Captured using monocular cameras surrounding an intersection, with automated annotations and a hand labelled test set.

2 Related Work

In this section we discuss mono3DODs and how the common datasets used to train them limit their application in road safety. We then review research how multi-view geometric constraints can be used in lieu of ground truth labels, namely in 3D pose estimation and depth prediction.

Monocular 3D Object Detectors: Monocular 3D object detection aims to infer the location of objects in the scene from a single image. For vehicle 3D object detectors, roll and pitch are set to zero with a flat ground assumption. Mono3DODs commonly predict an object's location in the image frame, yaw rotation, size, and depth. Differing methods utilise key points and CAD models [2, 5, 23, 25, 30], others utilise a tight fit constraint of a between a 2D and 3D bounding box [8, 22, 24, 27, 29], and some make direct predictions of 7DOF pose [3, 7, 12, 19, 35, 40, 41]. These monocular 3D object detection methods are trained with full supervision on autonomous vehicle datasets which are expensive to collect and hand label.

Autonomous vehicle datasets: Training of mono3DODs commonly use the autonomous vehicle datasets NuScenes [4], KITTI [15], Waymo [1], and Lyft [20]. These datasets were

collected using cameras on top of an ego vehicle's roof. The angle between a line from the camera centre to the centre of another vehicle with reference to the ground plane, reffered to as elevation angle, for autonomous vehicles is around 5 degrees. Traffic cameras observe vehicles at elevation angles as large as 50°. Mono3DODs have not been exposed to vehicles observed at these elevation angles during training. This results in noisy and inaccurate predictions of 7DoF pose. To train mono3DODs for high elevation angles in the conventional manner road safety researchers would need access to LiDAR data and extensive hand labelling which is often unobtainable.

Multi-view datasets: Existing publicly available multi-view traffic datasets with overlapping field of view (FoV) cameras either have too little data to train a mono3DOD [33, 37] or images from elevations that were too low [10, 11].

Beyond traffic based datasets, there have been several multi-view pedestrian datasets that contain overlapping FoVs, such as WILDTRACK and Toulouse campus dataset [6, 26]. However, these datasets do not contain 3D annotations or traffic scenes that would be required for our use case.

Geometric weak-supervision of monocular object detectors: In the context of object detection, geometric multi-view constraints were used to train pose and depth estimators without ground truth data. Simon *et al*. [34] annotate 3D hand poses using 2D hand pose detector and 3D reconstruction. A monocular hand key point detector is then trained on these annotations. Predictions of pose detectors can be compared over multiple views to calculate a consistency loss that can train a monocular model in a weakly supervised manner [18, 32]. Additionally, a multi-view geometric constraint can be used as a supervision signal for training depth prediction models [14]. Reprojection consistency losses have also been used for 3D shape prediction of vehicles in the case of Occlusion-Net [11]. Occlusion-Net utilises a multi-view reprojection loss where predicted object-centric 3D shapes are reprojected onto other views and compared with 2D key point detections to train models in a weakly supervised manner.

These weakly-supervised methods all aim to incorporate the additional information from multiple views to improve the predictions of object-centric pose but do not attempt to exploit the multi-view data to train a mono3DOD for 7DoF pose. We introduce a method to utilise weak geometric supervision to train a mono3DOD.

3 WIBAM dataset

The wide baseline multi-view (WIBAM) dataset collected for the purposes of training our weakly supervised model contains sequential image data from multiple cameras of an inner-city four-leg intersection. This intersection has dedicated right turn lanes, large divides between counter flowing traffic, and a steady flow of traffic. Cyclists, pedestrians, cars, and utility vehicles all use this intersection as a corridor through the city. A satellite image of the intersection and samples of the dataset can be found in Section **??** of the supplementary material.

Four GoPro Hero 7 cameras were set to use "linear field of view mode" (i.e. GoPro's radial distortion correction), and mounted at roughly the same height on the traffic poles in the four corners of the intersection. We maximise the inter-camera FoV overlap by pointing them towards the middle of the intersection. This allows constant observation of vehicles from multiple views. Unsynchronised video data was recorded at 50FPS by each camera at a resolution of 2560×1440 pixels, for roughly 15 minutes over the same period, in daylight

conditions on a clear day.

Traffic light signals, which were observable from all the camera views, were used to synchronise the recordings. Synchronisation was done semi-automatically with knowledge of the locations of the traffic lights in each view. When a change in traffic light is observed in a leading video the frames in the trailing videos are dropped until they have observed the light change. Figure **??** in the supplementary material shows a graph of the before and after delay between videos across the dataset. The maximum delay was reduced from 16 frames (0.3s) down to 3 frames (0.06s).

After synchronisation videos are down sampled to 12.5 FPS at a resolution of 1920×1080 resulting in $8,273$ images per camera; $33,092$ images in total. The images were split into three subsets for training, validation, and testing each containing 75%, 15%, and 10% of the dataset, respectively. Each subset is separate in time such that the test and validation set contain none of the same vehicles. We hand annotate 12% of the test split, as outlined in Section 3.3. Intrinsic and extrinsic camera calibration as well as a homography matrix for the ground plane are included with the dataset. These were found using measured correlated points between the camera planes and ground plane. We use the perspective-n-point algorithm to solve extrinsic calibration and the linear least means squares to solve the homography matrix.

3.1 Automatic multi-view annotations

We automatically annotate vehicles using an off the shelf 2D object detector [16]. In order to train a mono3DOD with multi-view geometric supervision, associations of a vehicle's 2D bounding boxes between the views must be known. Clustering is used to associate detections corresponding to the same vehicle across cameras. Specifically, rays are cast through each camera's 2D detections to find where the ray intersections the ground plane. These points of intersections are then assigned to vehicles via DP-means clustering [21]. Examples of the output of the clustering algorithm can be found in Figure **??** of the supplementary material.

3.2 Automatically annotated dataset information

The WIBAM dataset contains automatic annotations for vehicles detected in all images. A total of $45,762$ automatically annotated multi-view vehicles, leading to $116,702$ 2D bounding boxes of vehicles are contained in the dataset. There are $31,333$ vehicles viewed from two cameras, $3,680$ from three, and $10,749$ from four. Note that the high number of two camera observations are due to many predictions occurring as vehicles are approaching and leaving the intersection. When vehicles are within the intersection we observe that, in the majority of cases, all four cameras can observe the vehicle. In the case where a vehicle is occluded, by a bus for example, results in three camera detections. We estimate that each vehicle is viewed an average of 45 times resulting in a total of $1,000$ unique vehicles captured at different times, viewing angles, and locations throughout the intersection.

3.3 Hand labelled test set

In order to quantify the performance of our weakly supervised method we developed a test set using our multi-view annotation tool. The tool is used to annotate vehicle 7DoF pose and visibility by viewing the reprojections of annotations onto each camera view. A screenshot of the labelling tool GUI is included in Section **??** of the supplementary material.

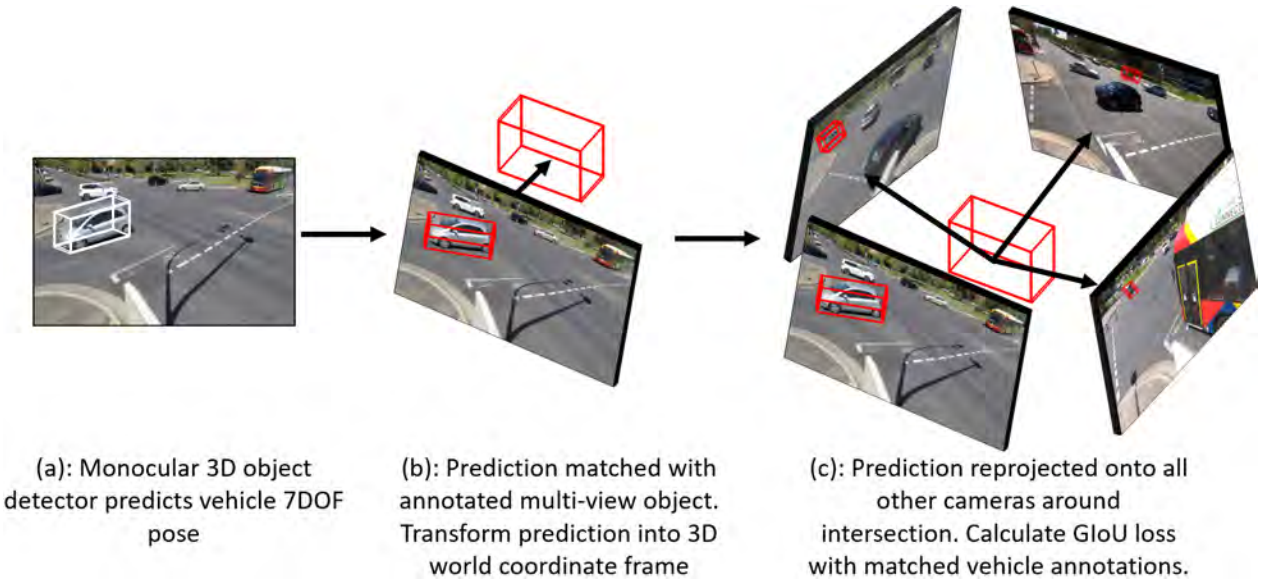


Figure 1: Method for calculating the reprojection loss at train time

The WIBAM test set which contains 400 hand annotated images resulting in 1,651 hand labelled 7DoF vehicle annotations, of which 400 are partly occluded. It is important to note that when vehicles are visible in fewer than three views it is challenging to accurately manually label them. However, in the more frequent case where three or four cameras observe the vehicle manual annotations can be produced using our tool. Distributions of detected vehicles' distances, elevation angles, and visibilities be found in Section **??** of the supplementary material.

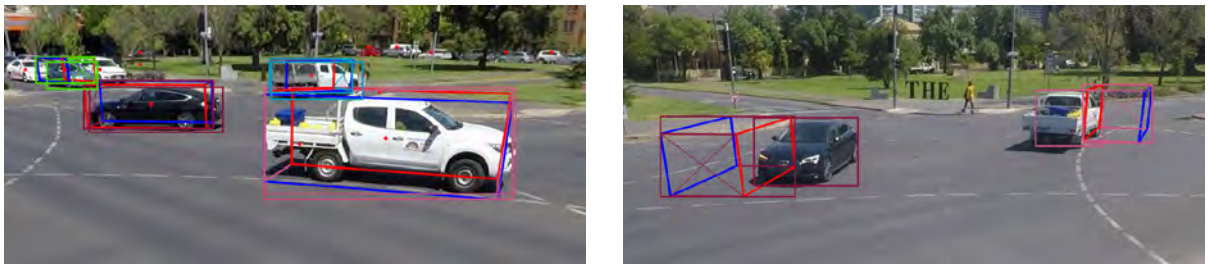
4 Multi-view loss

$$L_{MV} = \frac{1}{n}\sum_{n=1}^{n}(1 - GIoU_n) + L_{focal} \tag{1}$$

The multi-view reprojection loss 1 is compromised of two functions which are summed together. A focal loss for measuring the discrepancy between the predicted vehicle centre points and the centre points predicted by the 2D object detector, as in CenterNet [40]. This component is responsible for fine tuning the predicted 2D object centre detections from the mono3DOD so that it learns to identify vehicles in the new domain. The second function is our geometric consistency loss which utilises 3D object detections' reprojections onto other cameras to calculate the Generalised Intersection over Union (GIoU) [31]; encouraging consistency of the 3D predictions between views.

Our multi-view loss is calculated using the monoc3DOD to make predictions on one of the images from a single time frame, see Figure 1(a). Each vehicle prediction is then matched with the automatic multi-view annotations described in Section 3.1. Using this match, the 3D object is transformed from a camera-centric to a world coordinate frame, see Figure 1(b). The vehicle in world coordinate frame is then projected onto all other camera views of the scene as shown in Figure 1(c). A tight fitting 2D bounding box around the reprojection can then be compared with automatic multi-view 2D object detection boxes using GIoU.

Figure 2 highlights that simply training with a single camera is insufficient to improve



(a) (b)

Figure 2: Training images with 3D bounding box reprojections (a) 3D detections reprojected on the input image (b) 3D detections reprojected onto the second camera

Method	3D IoU ↑	BEV IoU ↑	ATE (m) ↓	ASE ↓	AOE (deg) ↓
Baseline	0.25 (s=0.18)	0.33 (s=0.25)	1.51 (s=0.97)	**0.12 (s=0.06)**	5.67 (s=5.47)
WIBAM model	**0.48 (s=0.09)**	**0.64 (s=0.13)**	**0.43 (s=0.43)**	0.23 (s=0.07)	**4.69 (s=2.91)**

Table 1: Localisation performance comparison between baseline and the fine-tuned WIBAM model

7DoF pose predictions. Explicitly, a prediction using the image from (a) as input results in what appears to be a good prediction. However, once this pose is projected onto a camera from a separate viewpoint, shown by (b), we visually inspect a large error.

5 Fine-tuning a 3D object detector

We use a DLA-34 [38] model from CenterNet [41] pre-trained on NuScenes [4] as our mono3DOD 'baseline model'. We fine tune the baseline model on the WIBAM dataset with a batch size of 80 images on two Nvidia V100 GPUs with an initial learning rate of $3.125e^{-5}$. We drop the learning rate by a factor of 10 when validation loss plateaus for 4 epochs with a tolerance of ± 0.001. The fine-tuned model is referred to as the 'WIBAM model'.

5.1 Quantitative results

We evaluate the performance of the mono3DOD by comparing the predictions to the manually labelled test set data; see Section 3.3. We consider errors in the different degrees of freedom separately: average translation error (ATE), average scale error (ASE), and average orientation error (AOE). Translation error is measured using the euclidean distance between the centres of the predicted and ground truth cuboids. Scale error is reported as $1-IoU$ of the ground plane bounding boxes after aligning the rotation and translation. This is equivalent to the absolute proportional difference in ground plane area. Orientation error is quantified using the smallest yaw angle between the ground truth and predicted orientations. Additionally, We calculate the 3D IoU and birds eye view (BEV) IoU to give a holistic measurement of 7DoF predictions.

Table 1 reports the performance of the baseline and WIBAM models on the WIBAM test set. Our experiments show that the 3D and BEV IoU are higher by 92% and 94%, respectively. This can be attributed to the lowering of ATE by 71% and AOE by 17%.

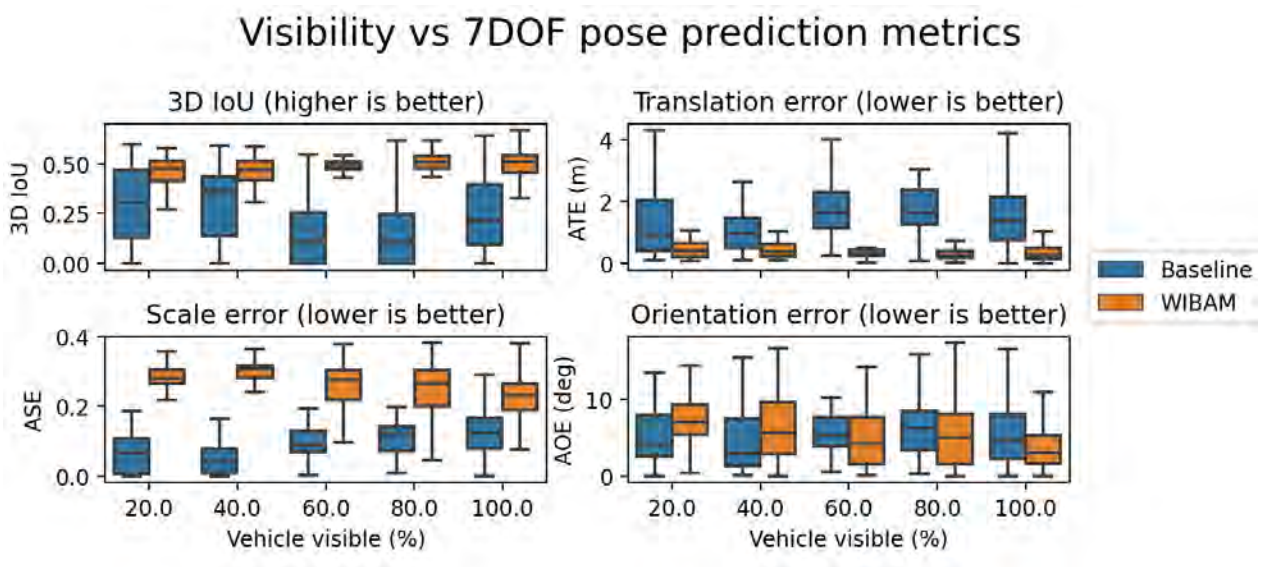


Figure 3: Effect on 7DoF pose prediction with varying amounts of vehicle occluded/truncated to camera.

Predictions produced by the WIBAM model induce less variance in error and IoU metrics when compared to the baseline. This demonstrates that our method of training produces predictions which are more consistent. We perform Z-tests to ensure our performance is significantly better and found that all errors in Table 1, except scale, are significant to the 0.01 level.

We observe a higher scale error in the WIBAM model's predictions, which is believed to be caused by partially occluded vehicles. The 2D object detector used will predict a bounding box encapsulating the visible part of the vehicle. However, the full 3D bounding box predicted will be reprojected for calculation of the GIoU. This creates a signal, via the GIoU loss, encouraging the model to underestimate the size of vehicles.

We evaluated the performance of the mono3DODs when objects are occluded and truncated. Figure 3 shows how performance metrics vary as a function of vehicle visibility. Note that 3D IoU is not significantly impacted by visibility, and is consistently higher for the WIBAM model versus baseline. This can be attributed to a large reduction in translation error; see the 7DoF prediction component errors in Figure 3.

The effect of occlusion on scale error is particularly apparent in Figure 3 where less visible vehicles have higher scale errors; a relationship not observed in baseline model predictions. We can handle these situations for truncated vehicles by clamping 3D bounding box reprojections at the image boundaries but it is not as trivial in the case of occlusions.

We trained models using varying amounts of the training set to evaluate what a suitable amount of training data would be. The data selected is from the beginning of the video and is treated sequentially to simulate model fine-tuning on a given sample collected while cameras are installed. Different amounts of time, from half a minute (5%) up to eight minutes (100%), are tested to see how much footage is required for our method to be effective

Training with 40% (3 minutes 18 seconds) of the data produces the largest performance improvement on the test set; see Figure 4. Large orientation errors are observe when training on small amounts of data. This indicates that insufficient training data leads to poor generalisation on the test set. ATE is observed to immediately improve with any amount of data but

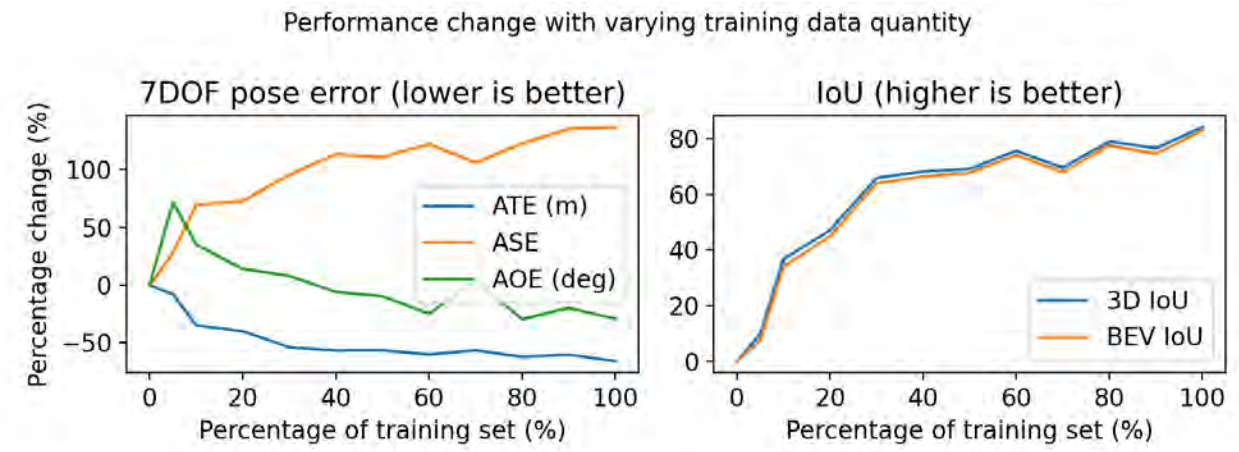


Figure 4: Training with varying levels of data. Data is shown as a percentage improvement over the baseline trained model; i.e. 0% training set. (a) 7DoF pose components (b) Intersection over Union.

Method	2D ATE (m) ↓	ASE ↓	AOE (deg) ↓
WIBAM model on WIBAM test set	0.41	0.22	4.69
CenterNet [40] on NuScenes cars test set	0.47	0.14	5.38
FCOS3D [9] on NuScenes cars test set	0.56	0.15	5.15

Table 2: Comparison of our 7DoF pose errors with NuScenes SOTA models

has diminishing returns beyond 40% of training data. A model trained with any amount of multi-view data is able to improve the predicted 3D and BEV IoU. This finding is helpful to show that for a particular distribution in time of day and vehicle flow, around 10 minutes of HD video collected at 12.5FPS is capable of returning significantly better 3D IoU.

Table 2 shows absolute 7DoF pose errors of the WIBAM model on the WIBAM test set. We also show the current state of the art for supervised mono3DODs trained and tested on NuScences. Note that the error for each is comparable. Although this comparison is not strictly "apples to apples", we see that our method is able to use weak supervision to improve the performance so that it is comparable to the best fully supervised models.

5.2 Qualitative results

To demonstrate our improvement in 7DoF pose accuracy, we show before and after visualisations of the baseline and WIBAM model predictions. Figure 5 contains a comparison between the 3D bounding boxes produced by the baseline model and WIBAM model. Each row is a specific vehicle cropped from different viewpoints at the same time. Detection camera predictions are shown in the first columns. Subsequent columns are crops of these predictions reprojected onto the other three camera views. Row one is an example of our method's ability to refine size predictions, row two shows an example of improvements to localisation, and rows three and four contain examples of predictions on truncated and occluded vehicles.

Reprojections of 3D bounding boxes onto a camera frame often do not display the subtle errors present in model predictions. Figure 6 is a BEV illustration of WIBAM and baseline model predictions projected onto the intersection. Initially estimations made for vehicle one

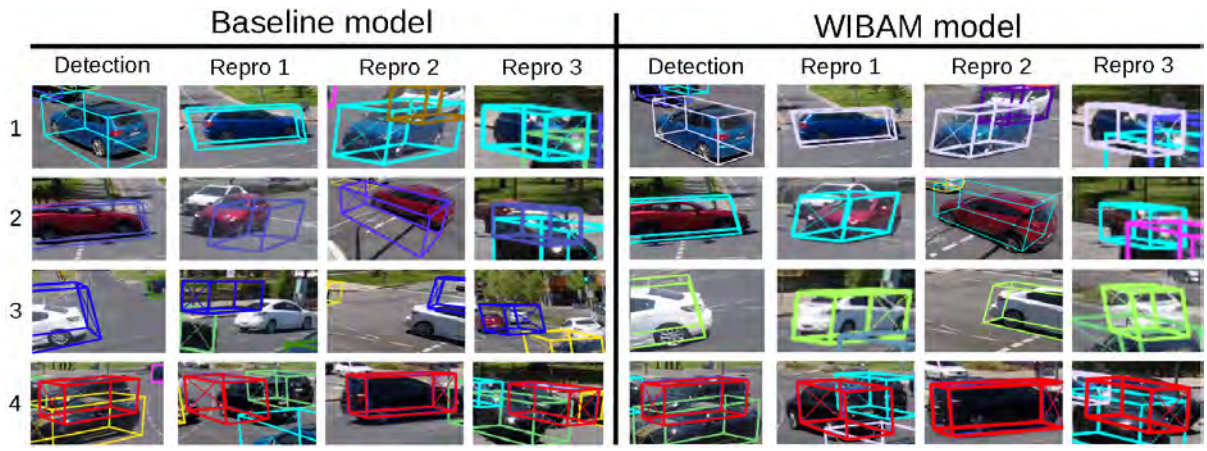


Figure 5: Detections of the same vehicles with the baseline model and the WIBAM model

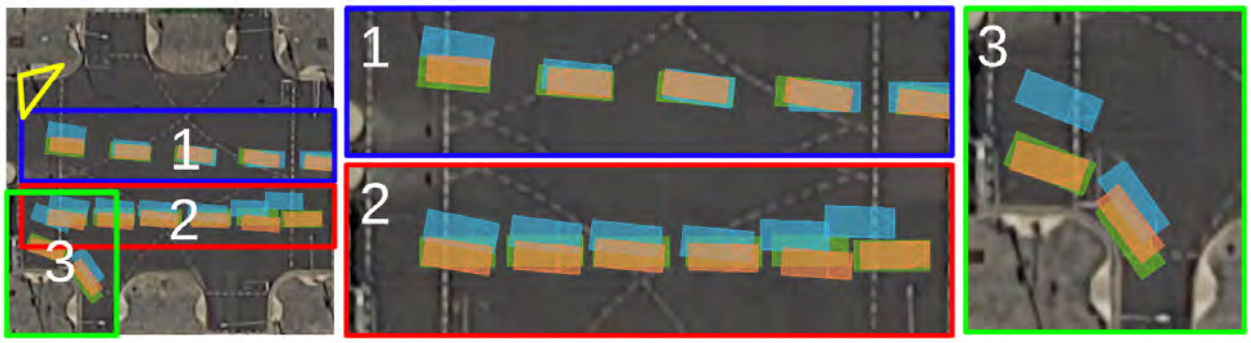


Figure 6: Multiple time instances of predicted vehicle locations in the intersection all projected onto a BEV image. Baseline predictions are shown in blue, WIBAM model predictions shown in orange, and hand annotated ground truth shown in green. Vehicles reprojections are separated, labelled, and colour coded for clarity of comparison. The yellow triangle shows the camera location. Imagery ©2021 Aerometrex Pty Ltd, Map data ©2021 Google. We provide videos of these sequences in the supplementary material.

are consistent with ground truth but a jump is observed at the end of the trajectory. Vehicle two shows that the consistency of predictions is improved with our fine tuning. Vehicle three shows a drastic improvement in localisation of a vehicle which is far away and was partially occluded by vehicle two.

Figure 5 and 6 illustrate the improved vehicle 7DoF pose predictions of the WIBAM model over the baseline. Our method is able to make 7DoF pose predictions of vehicles in the scene more accurately than the baseline and do so more consistently over time.

6 Conclusion

In this paper we have developed a weak supervision method for fine-tuning monocular 3D object detection models on high mounted, static intersection observation cameras. Specifically, our method allows a road safety researcher to fine-tune a model for a specific camera set up with 10 minutes of additional video data from cameras with overlapping fields of view and no hand labelling. Unlike long term multi-view data collection, short term multi-view

data collection can be easily setup once with inexpensive, consumer grade cameras.

Our experiments show that 3D object detection models trained on ego vehicle data suffer in performance when used on intersection observation cameras. Our loss uses weak labels from a pretrained 2D object detector to ensure consistency between views of a model's reprojected 3D object detections. We show that our method of fine-tuning a monocular 3D object detector achieves vehicle 7DoF pose prediction accuracy on the WIBAM test set comparable to SOTA models on the NuScenes test set.

We also release the WIBAM dataset with 45k automatically annotated multi-view vehicles (116k 2D bounding boxes), 1,651 7DoF pose annotations of vehicles labelled with our multi-view annotation tool, and the WIBAM model. We would like to extend our method for vulnerable road users such as cyclists and pedestrians.

Acknowledgements

This research has been supported through the Australian Government Research Training Program Scholarship. High performance compute resources used in this work were funded by the Australian Research Council via LE190100080.

References

[1] Waymo open dataset: An autonomous driving dataset, 2019.

[2] Ivan Barabanau, Alexey Artemov, Evgeny Burnaev, and Vyacheslav Murashkin. Monocular 3d object detection via geometric reasoning on keypoints. *CoRR*, abs/1905.05618, 2019. URL http://arxiv.org/abs/1905.05618.

[3] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9286–9295, 2019. doi: 10.1109/ICCV.2019.00938.

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[5] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teulière, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1827–1836, 2017. doi: 10.1109/CVPR.2017.198.

[6] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018. doi: 10.1109/CVPR.2018.00528.

[7] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Hee Min Choi, Hyoa Kang, and Yoonsuk Hyun. Multi-view reprojection architecture for orientation estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2357–2366, 2019. doi: 10.1109/ICCVW.2019.00289.

[9] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020.

[10] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[12] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11669–11678, 2020. doi: 10.1109/CVPR42600.2020.01169.

[13] Ting Fu, Weichao Hu, Luis Miranda-Moreno, and Nicolas Saunier. Investigating secondary pedestrian-vehicle interactions at non-signalized intersections using vision-based trajectory data. *Transportation Research Part C: Emerging Technologies*, 105: 222–240, 2019. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2019.06.001.

[14] Ravi Garg, Vijay Kumar B.G., Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 740–756, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46484-8.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[16] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018.

[17] Thomas Götschi, Alberto Castro, Manja Deforth, Luis Miranda-Moreno, and Sohail Zangenehpour. Towards a comprehensive safety evaluation of cycling infrastructure including objective and subjective measures. *Journal of Transport and Health*, 8:44–54, 2018. ISSN 2214-1405. doi: https://doi.org/10.1016/j.jth.2017.12.003.

[18] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5242–5251, 2020. doi: 10.1109/CVPR42600.2020.00529.

[19] Eskil Jörgensen, Christopher Zach, and Fredrik Kahl. Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss. *CoRR*, abs/1906.08070, 2019. URL http://arxiv.org/abs/1906.08070.

[20] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Lyft level 5 perception dataset 2020. https://level5.lyft.com/dataset/, 2019.

[21] Brian Kulis and Michael I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics, 2012.

[22] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1019–1028, 2019. doi: 10.1109/CVPR.2019.00111.

[23] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. RTM3D: real-time monocular 3d detection from object keypoints for autonomous driving. *CoRR*, abs/2001.03343, 2020. URL https://arxiv.org/abs/2001.03343.

[24] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1057–1066, 2019. doi: 10.1109/CVPR.2019.00115.

[25] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4289–4298, 2020. doi: 10.1109/CVPRW50498.2020.00506.

[26] Thierry Malon, Christine Senac, G. Roman-Jimenez, Patrice Guyot, Sylvie Chambon, Vincent Charvillat, Alain Crouzil, André Péninou, Julien Pinquier, and Florence Sedes. Toulouse campus surveillance dataset: scenarios, soundtracks, synchronized videos with overlapping and disjoint views. pages 393–398, 06 2018. doi: 10.1145/3204949.3208133.

[27] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Košecká. 3d bounding box estimation using deep learning and geometry. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5632–5640, 2017. doi: 10.1109/CVPR.2017.597.

[28] Matin S. Nabavi Niaki, Nicolas Saunier, and Luis F. Miranda-Moreno. Is that move safe? case study of cyclist movements at intersections with cycling discontinuities. *Accident Analysis and Prevention*, 131:239–247, 2019. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2019.07.006. URL https://www.sciencedirect.com/science/article/pii/S0001457518306092.

[29] Andretti Naiden, Vlad Paunescu, Gyeongmo Kim, ByeongMoon Jeon, and Marius Leordeanu. Shift r-cnn: Deep monocular 3d object detection with closed-form geometric constraints. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 61–65, 2019. doi: 10.1109/ICIP.2019.8803397.

[30] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A general framework for monocular 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. doi: 10.1109/TPAMI.2021.3074363.

[31] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. June 2019.

[32] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images, 2018.

[33] G. Roig, X. Boix, H. Shitrit, and P. Fua. Conditional random fields for multi-camera object detection. https://www.epfl.ch/labs/cvlab/data/data-multiclass/, 2011.

[34] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. pages 4645–4653, 07 2017. doi: 10.1109/CVPR.2017.494.

[35] Andrea Simonelli, Samuel Rota Bulò, Lorenzo Porzi, Manuel Lopez-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1991–1999, 2019. doi: 10.1109/ICCV.2019.00208.

[36] Paul St-Aubin, Luis Miranda-Moreno, and Nicolas Saunier. An automated surrogate safety analysis at protected highway ramps using cross-sectional and before–after video data. *Transportation Research Part C: Emerging Technologies*, 36:284–295, 2013. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2013.08.015.

[37] Elias Strigel, Daniel Meissner, Florian Seeliger, Benjamin Wilking, and Klaus Dietmayer. The ko-per intersection laserscanner and video dataset. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1900–1901, 2014. doi: 10.1109/ITSC.2014.6957976.

[38] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018. doi: 10.1109/CVPR.2018.00255.

[39] Sohail Zangenehpour, Luis F. Miranda-Moreno, and Nicolas Saunier. Automated classification based on video data at intersections with heavy pedestrian and bicycle traffic: Methodology and application. *Transportation Research Part C: Emerging Technologies*, 56:161–176, 2015. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2015.04.003.

[40] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019. URL http://arxiv.org/abs/1904.07850.

[41] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ECCV*, 2020.

1 Supplementary material

1.1 WIBAM dataset samples

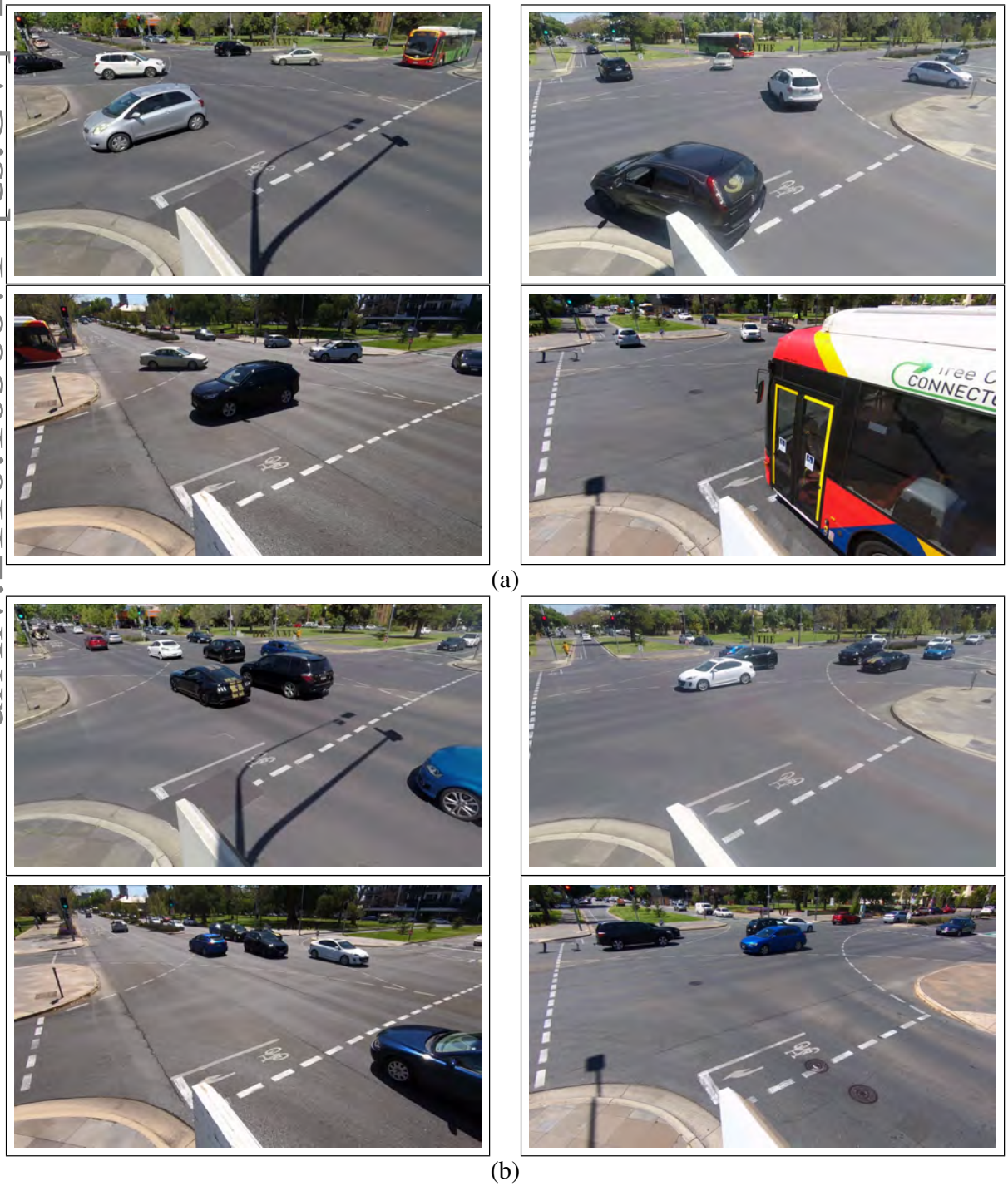


Figure 1: Example images of the dataset showing the view from each camera. (a) Sample from the training set (b) Sample from the validation set

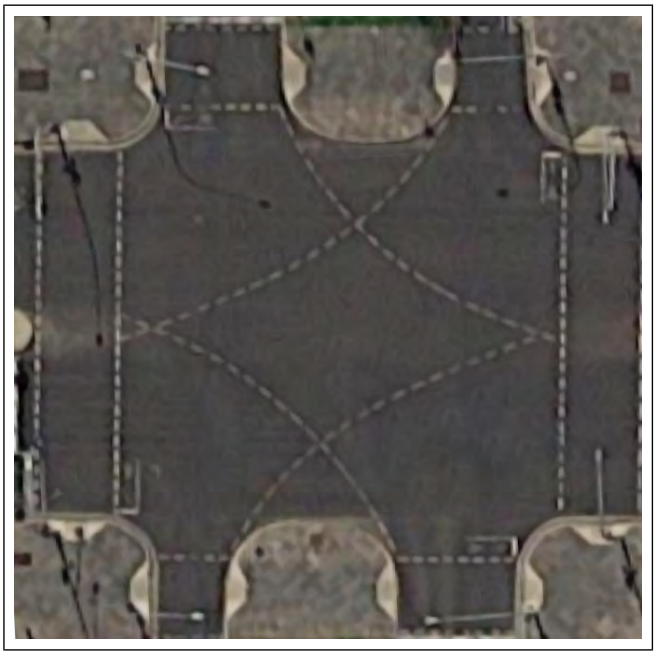


Figure 2: Satellite image of the intersection being observed. Imagery ©2021 Aerometrex Pty Ltd, Map data ©2021 Google

1.2 Camera synchronisation

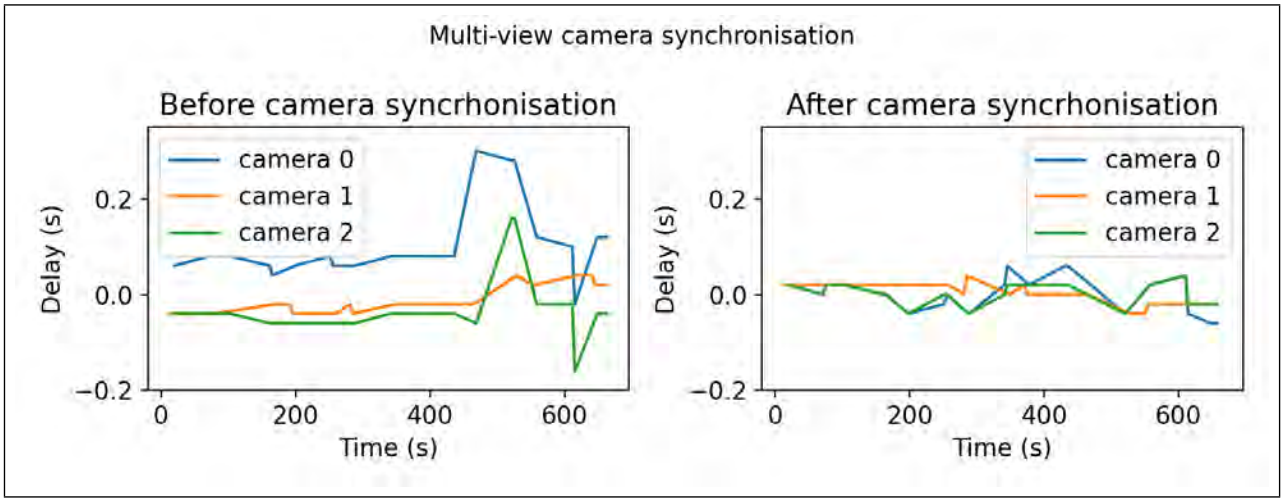


Figure 3: Before and after synchronising the dataset using the traffic lights changing. Delay with reference to camera 3 in the dataset.

Figure 3 shows the improvement in synchronisation between the multi-view cameras using our semi-automatic method. Before synchronisation of cameras, it is clear that there is a large error up to 16 frames towards the end of data collection. This large error can result in cars having zero overlap on the road surface between views. To measure the light changes we can use the change from green to yellow as the measurement signal and yellow to red light change as synchronisation signal. This measures the delay between cameras at the end of the light cycle and leaves a gap between red lights to measure if the cameras go out of sync again.

1.3 Clustering output

(a)

(b)

Figure 4: (a) Example of the clusters formed with vehicles detected from multiple views (b) Largest number (16) of clusters in the dataset. The colours correspond to a matched cluster. Black dots represent the four cameras with lines drawn out from the camera to the detection centre on the ground plane. Best viewed in colour and zoomed in.

Figure 4 shows the output of the clustering algorithm used for the multi-view vehicle associations. The diagram on the left shows the BEV clusters where the black dots are the camera centres with lines to the detection ground points. Objects which are clustered have a unique colour between the detection images on the right and the BEV diagram on the left. Figure 4(b) shows the highest number of matched vehicles the system found with 16 unique objects viewed at this time instance.

1.4 Automatically annotated multi-view vehicle distributions

Varying the positions of vehicles within the intersection give a wide range of depths and elevations vehicles are viewed from. There is also a different numbers of vehicles associated and matched across views at any given time, with a maximum of 16 shown in Figure 4(b). Figure 5 shows the distributions of these in the entire dataset. Our dataset views vehicles at a wide range of elevations, where autonomous vehicle data is typically around 5°.

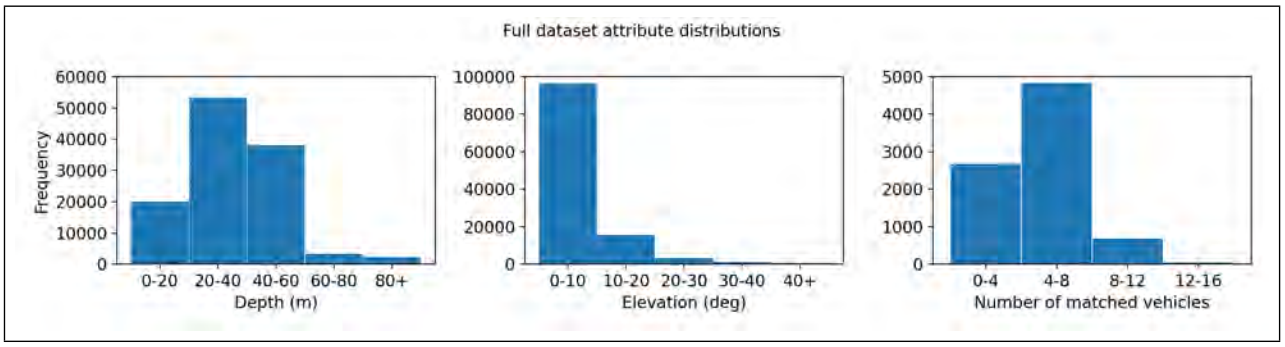


Figure 5: Distributions of key data elements within the WIBAM dataset

1.5 Test set distributions

Figure 6 shows the distributions of depth and elevation angle in the hand labelled test set of the WIBAM dataset. The distributions of elevations and distances are representative of the whole dataset however vehicles at larger distances outside of view for 4 cameras are difficult to label accurately hence there is a drop off above 40m+. The high mounted cameras result in less occlusions of vehicles therefore the majority are not occluded at all.

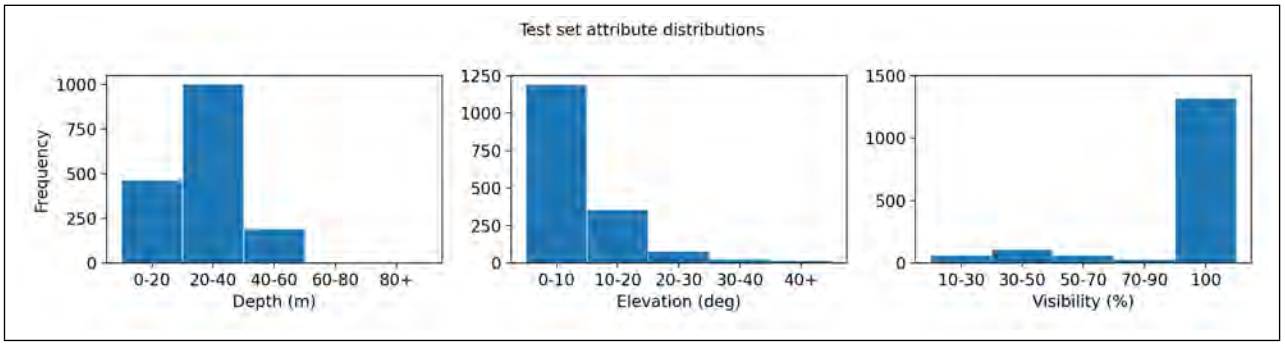
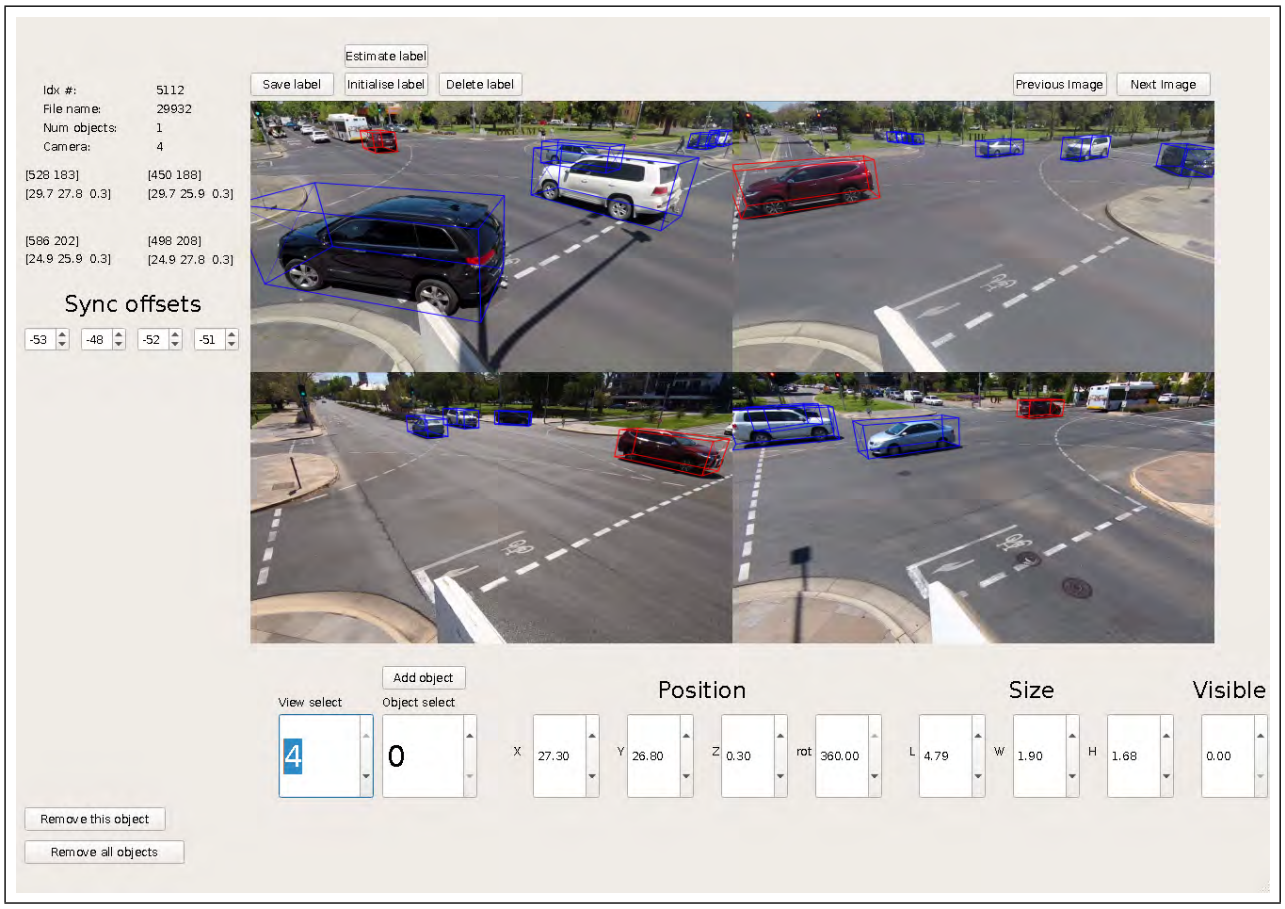


Figure 6: Hand labelled annotation distributions

1.6 Labelling tool GUI

The labelling tool allows the user to annotate vehicles with 7DoF pose. We found that annotating vehicles using this tool takes around 2.5 minutes per time instance or just over half a minute per image labelled.



(a)

Figure 7: The labelling tool created for labelling monocular multi-view data