Evaluating the Performance of Ensemble Methods and Voting Strategies for Dense 2D Pedestrian Detection in the Wild

Aboli Marathe
Symbiosis Centre for Applied
Artificial Intelligence (SCAAI),
Symbiosis International Deemed
University (SIU), India
SCTR's Pune Institute of
Computer Technology, India
aboli.rajan.marathe@gmail.com

Rahee Walambe
Symbiosis Centre for Applied
Artificial Intelligence (SCAAI),
Symbiosis International Deemed
University (SIU), India
Symbiosis Institute of Technology,
SIU, India
rahee.walambe@sitpune.edu.in

Ketan Kotecha
Symbiosis Centre for Applied
Artificial Intelligence (SCAAI),
Symbiosis International Deemed
University (SIU), India
Symbiosis Institute of Technology,
SIU, India
director@sitpune.edu.in

Abstract

As vehicles experience a wide variety of driving settings in the wild, 2D pedestrian detection offers a substantial barrier to autonomous vehicle navigation systems. In this work, we demonstrate the effectiveness of a lightweight ensemble architecture for pedestrian detection in the wild, which combines detectors and data augmentation techniques to improve the performance of well-established detectors. The framework uses voting strategies to increase the explainability of object detection in navigation systems while also improving the precision of bounding box predictions on the dataset. The ensemble of the best model and augmentation technique achieved 41.41 % AP in detecting pedestrians in the wild using the consensus voting strategy on the WiderPerson dataset.

1. Introduction

The rapid development of self-driving vehicles over the last decade has increased the demand for reliable object detection systems. Vehicles encounter a wide range of challenging environments to navigate through on a daily basis, so the systems guiding the vehicle's navigation must be robust, precise, lightweight, and fast. In these environments, the systems must detect and interpret social cues, follow traffic laws, and make critical decisions in emergency situations. Furthermore, these decisions must be explainable and ethically responsible in order to ensure the safety and trust of passengers and pedestrians.

Vehicles in the wild face unpredictable driving conditions such as poor weather and unstructured road conditions. However, one of the most difficult challenges is navigating through crowds and around unpredictable pedestrians. Pedestrians are more difficult to detect against backgrounds in image processing than other items such as automobiles and buildings. They are frequently found in clusters, in various positions, and in low resolution when captured by sensors on autonomous vehicles. These challenges must be overcome by the deployed detection models, which must identify individual pedestrians, provide precise detections, and be unaffected by diverse surroundings. Because the navigation decisions are made by assessing the vehicles' state in its surroundings, the reliability of navigation systems is highly dependent on the quality of detections provided by the models. Thus pedestrian detection is a significant challenge in computer vision and is an important component of object recognition for autonomous systems. In

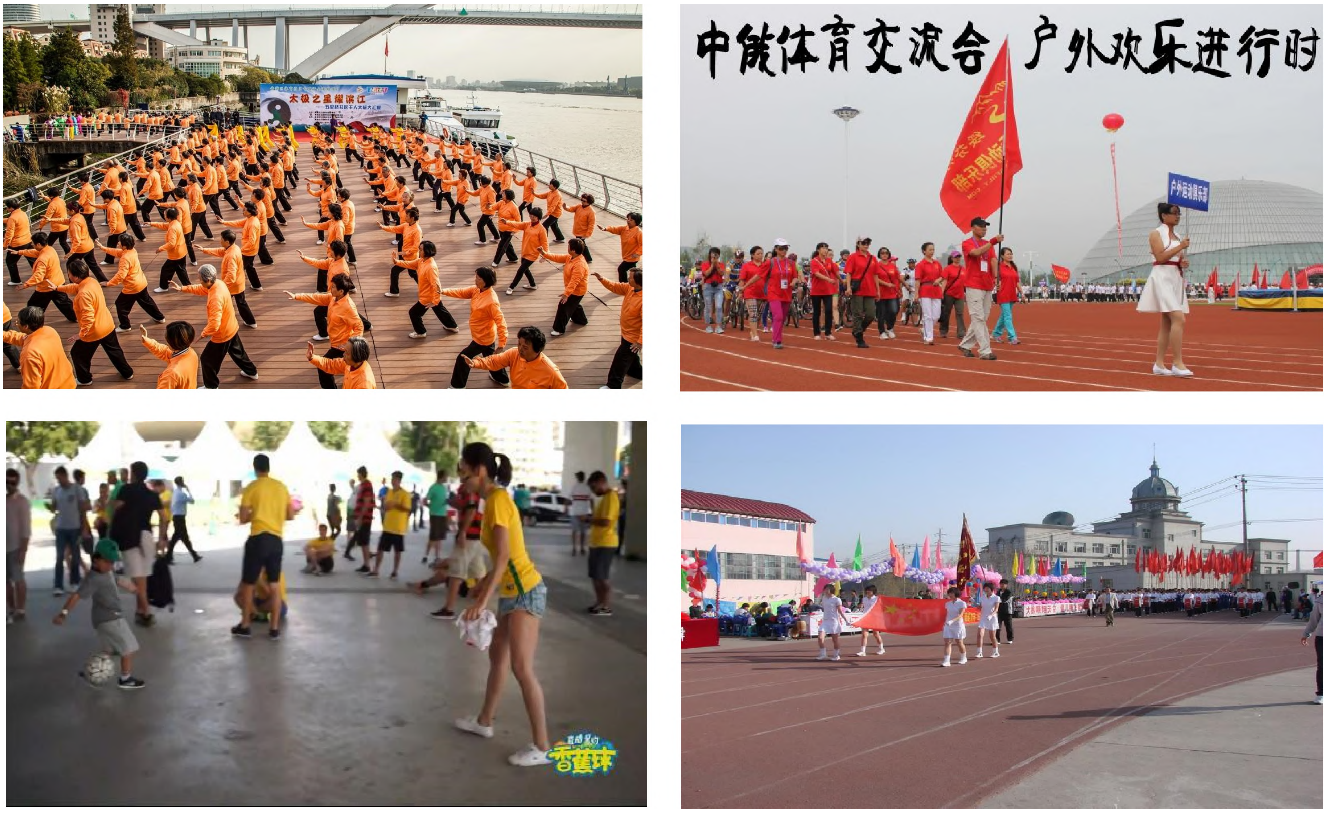


Figure 1. WiderPerson Dataset Sample

this work, we are tackling the challenge of pedestrian detection in the wild by applying a lightweight ensemble framework that combines detectors and augmentation techniques using voting strategies. Our main contributions include:

- Experimenting with data augmentation techniques to determine which augmentation is most effective for detecting pedestrians in the wild.
- Combining detections from multiple object detectors with distinct architectures to detect pedestrians in a variety of environments.
- Using the ensemble framework to study the effects of consensus, affirmative, and unanimous voting strategies on overall prediction precision.

2. Related Work

As companies race towards deploying self-driving cars in the near future, the demand for better detectors and algorithms is increasing. Building robust systems for detecting and tracking pedestrians with image processing techniques and deep learning models is rapidly gaining attention and many works propose innovative methods of achieving this. [8,10,12,15,19,21,23] Because of its applications in safety, surveillance, and robotics, tracking pedestrians has been a part of multiple studies and surveys that introduce innovative perspectives as well as introduce the ethical concerns surrounding this technology. [2, 3, 26] Studying the bias in datasets is an important concern in pedestrian detection, especially when we deal with crowd surveillance and large scale real world applications. A common bias seen in pedestrian detection tasks is higher miss rates for female pedestrians and poor detection results for children. For applications like autonomous vehicles, biased datasets will ultimately lower the reliability of such vehicles and threatens to affect the safety and trust between these cars and humans in the future.

2.1. Datasets

The need for diverse pedestrian datasets for training models inspired the creation of several datasets including INRIA [6] , ETH [10], TUD-Brussels [32], and KITTI [14]. INRIA [6] contains pedestrians in bias-free poses in diverse settings, ETH [10] and TUD-Brussels [32] are mid-sized video datasets and ETH [10] and KITTI [14] provide stereo information. The focus of this study, pedestrian detection in the wild was implemented using WiderPerson,

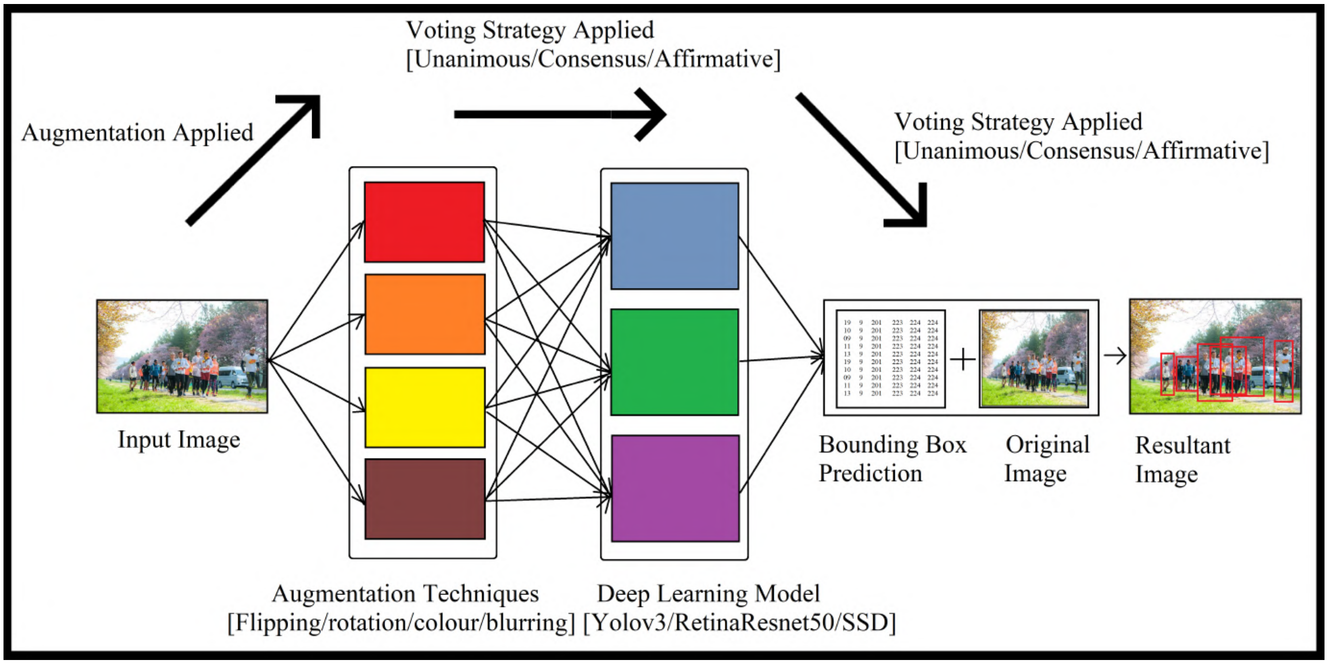


Figure 2. Ensemble Framework

a dataset that was introduced in 2019 for dense pedestrian detection in the wild, from scenarios not limited to traffic scenes and has been used in this study. [37]

2.2. Algorithms

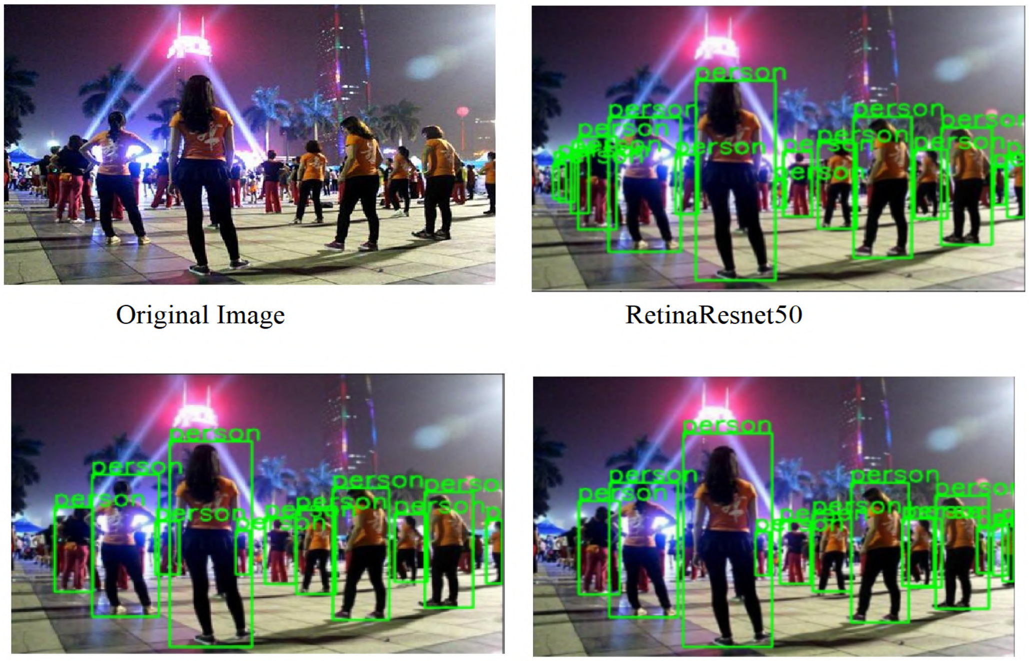
Using computer vision for pedestrian detection was first seen around 2003, when Viola and Jones introduced the VJ detector. [30] Subsequently, several traditional detectors were introduced including the HOG detector, ACF and Checkerboards. [7, 9, 36] The use of Convolutional Neural Networks improved the performance on this task drastically with accurate object detection using CNN-based models [4,16,35] . Other works included interesting approaches like convolutional sparse coding for pretraining the CNN, scale-dependent pooling and layer-wise cascaded rejection classifiers, multiscale detection and multi-task network architecture. [1, 4, 22, 29, 34] . The application of ensemble learning for pedestrian detection has covered approaches including structured ensemble learning method with spatially pooled features [25], Random Dropout and Ensemble Inference Network [13] , tree classifier ensemble [33] and ensemble of classifiers using different feature representation schemes of the pedestrian images [24] performed well on the datasets. The work in this paper has been based on the ensemble technique proposed in [5, 31] for combining detectors for object detection.

3. Data Description

The WiderPerson dataset [37] is a large compilation of 13, 382 images scraped from the internet using 50 keywords including pedestrian, cyclist, walking, running, marathon that display human beings in a variety of scenarios, environments and performing different activities in the wild which can be seen in Figure 1. The dataset has an average density of 28.87 persons per image which is relatively higher compared with all previous datasets. It also covers a wide range of scales and the data distribution at multiple scales is relatively uniform. As shown in Figure 1, the dataset spans a wide variety of locations, seasons and contains pedestrians spread across the image except the sky area in the upper portion. For this study the entire original validation subset of 1000 images was considered, which was provided in the WiderPerson dataset [37], for testing the performance of the ensemble framework and formed a total of 4000 images after the augmentation was applied.

4. Methodology

The ensemble framework for detecting pedestrians as shown in Figure 2 using the ensemble technique is based on 4 components that will be discussed in this section; the ensembling strategy, data augmentation, deep learning models used and the voting strategies.



Original Image

RetinaResnet50

Yolov3

SSD

Figure 3. Performance of Baseline Models

Model	**Class[Pedestrian] AP (%)**
RetinaResnet50	**40.81**
SSD	30.46
Yolov3	35.27

Table 1. Results of Baseline Models on WiderPerson Dataset

4.1. Ensemble Strategy

The ensemble algorithm uses the IoU metric with a threshold of 0.5 to group predictions and acknowledge the presence of objects in the images. [28] For two bounding boxes b1 and b2, the overlapped region is calculated by Equation 1.

$$IoU(b1, b2) = \frac{area(b1 \cap b2)}{area(b1 \cup b2)} \tag{1}$$

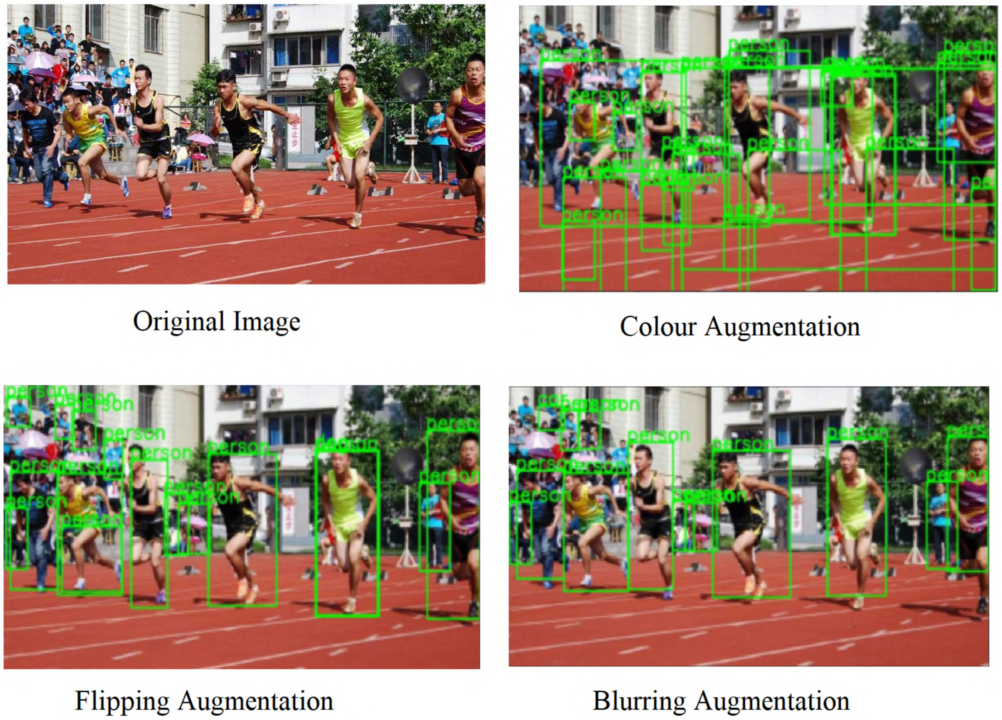
4.2. Data Augmentation

Multiple data augmentation techniques were carried out to improve the predictions of the framework and the best augmentation technique was chosen for the final experiments using the score on the baseline models.

- Colour Augmentation: By raising different RGB colour channels, the images in the dataset were augmented.
- Flipping Augmentation: By flipping the images along different axes; horizontal and vertical, the images in the dataset were augmented.
- Blurring Augmentation: Four blurring augmentation techniques, average blurring, bilateral blurring, basic blurring and Gaussian blurring were used for the experiments.

4.3. Models

For this study, 3 deep learning models were selected for the ensemble framework which showed good performance on object detection on other datasets in the past. The training sets used for training the 3 models were PASCAL VOC [11] and COCO [18]. These training datasets have one important similarity with the task of this study, they both contain humans as one of main classes in realistic scenes in the



Original Image

Colour Augmentation

Flipping Augmentation

Blurring Augmentation

Figure 4. Comparison of Data Augmentation Techniques using SSD Model and Affirmative Strategy

Model	Class[Pedestrian] AP (%)
Flipping	27.31
Blurring	26.99
Colour	**31.10**

Table 2. Results of Data Augmentation Techniques on WiderPerson dataset using SSD Model and Affirmative Strategy

images. For this property and the volume of data present in them, they were selected for training the models.

- Yolov3 [27] - A modification of the original Darknet model, with 53 layers stacked onto the architecture, the Yolov3 is a real time detection algorithm which makes detections at 3 different scales. This model was trained on VOC dataset. [11]

- RetinaResnet50 [17] - RetinaNet uses Feature Pyramid Networks which improves multi-scale predictions and focal loss, which addresses class imbalance. This model was trained on the COCO dataset. [18]

- SSD Resnet [20] - This is a Single Shot MultiBox Detector network with the inside VGG16 replaced with a ResNet50 network. This model was trained on VOC dataset. [11]

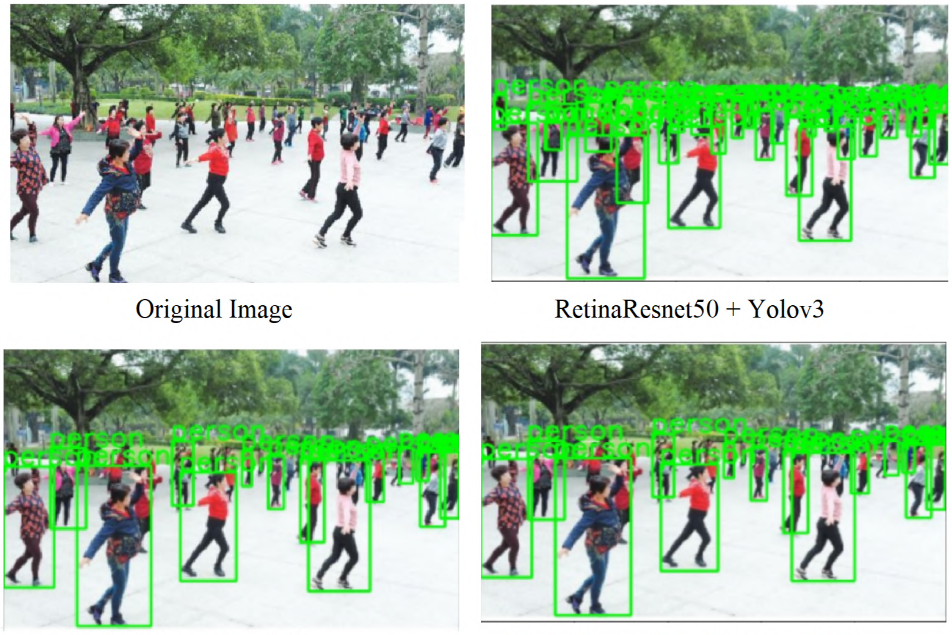
4.4. Voting Strategies

To determine how the predictions should be combined to create the final predictions, we applied voting strategies to obtain the results:

- Affirmative: When one of the detectors predicts that a region contains an object initially, such a detection is considered as valid.

- Consensus: The majority of the initial detectors must agree to consider that a region contains an object.

- Unanimous: All the methods must unanimously agree to consider that a region contains an object.



Original Image

RetinaResnet50 + Yolov3

RetinaResnet50 + SSD

RetinaResnet50 + Yolov3 + SSD

Figure 5. Comparison of Ensemble Models using Affirmative Strategy

Model	**Unanimous Class[Pedestrian] AP (%)**	**Affirmative Class[Pedestrian] AP (%)**	**Consensus Class[Pedestrian] AP (%)**
Yolov3	34.81	35.65	35.43
SSD	30.21	31.10	30.73
RetinaResnet50	**39.41**	**41.31**	**41.41**
Yolov3 + SSD	29.71	35.64	35.63
Yolov3 + RetinaResnet50	33.84	39.75	40.01
RetinaResnet50 + SSD	29.41	39.17	39.55
Yolov3 + RetinaResnet50 + SSD	28.32	38.71	36.84

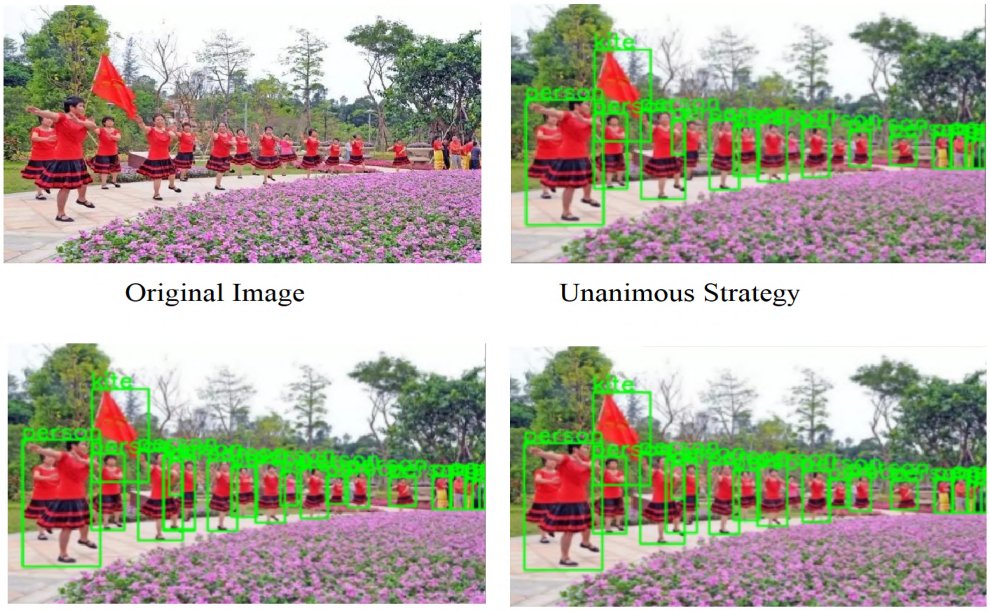
Table 3. Results of Final Ensemble Models on WiderPerson Dataset Using All 3 Voting Strategies

5. Experiments

In this study several experiments were conducted for determining the performance of the ensemble methodology for pedestrian detection. The metric selected for the experiments was Average Precision(AP) for the Pedestrian Class detection. First all 3 baseline object detection models were tested on the original dataset, without any augmentation and the results are presented in Table 1. Then one baseline model (SSD) and one voting strategy (affirmative) was chosen for running the detectors on the three different types of augmentation to find the best data augmentation technique of them as shown in Table 2. Finally the best augmentation was applied to the 3 baseline models and 4 ensemble models with all 3 strategies: consensus, affirmative and unanimous. These results are presented in Table 3.

6. Results

The results of the experiments demonstrate the effectiveness of baseline models as presented in Figure 3, augmentation techniques shown in Figure 4, ensemble models shown in Figure 5 and voting strategies as shown in Figure 6 in pedestrian detection. The performance of the baseline RetinaResnet50 outperforms all the other baseline models with 40.81% AP. Selecting the SSD model with affirmative strategy, the performance of different data augmentation



Original Image

Unanimous Strategy

Affirmative Strategy

Consensus Strategy

Figure 6. Comparison of Voting Strategies using RetinaResnet50 model and Colour Augmentation

techniques was tested and it was found that colour augmentation generated the best results, outperforming the second best technique flipping by 3.79% AP. Using colour augmentation, then the performance of 7 models were compared using AP. RetinaResnet50 performed the best in all 3 strategies, unanimous, consensus and affirmative with 39.41%, 41.41% and 41.31% respectively as visualized in Figure 7. The unanimous strategy worked worse than the baseline models after data augmentation. On different ensemble models, different voting strategies performed differently but affirmative and consensus strategies worked best overall. This result indicates that having a higher number of pedestrian detections is providing better results when combined than a smaller number of detections confirmed by all three models. One potential cause for this could be the high density of pedestrians which is accurately determined only by some models due to the model architectures but cannot be captured by the rest, leading to poor unanimous decisions.

7. Conclusion and Future Work

For autonomous vehicles, surveillance systems, traffic safety and many other applications, detecting pedestrians in densely packed clusters, challenging scenarios, and diverse environments is a significant challenge. In this work we have demonstrated the application of a lightweight ensemble framework for pedestrian detection in diverse environments. On the WiderPerson dataset, we were able to achieve good results using the ensemble algorithm, voting procedures, and data augmentation for pedestrian recognition, and we were able to draw numerous interesting inferences from the results. Colour based data augmentation worked best for the dataset and improved the performance of the baseline models significantly.

As seen by the good performance of RetinaResnet50 with consensus strategy in recognising pedestrians in the wild, combining more models did not always result in better results. The reason for this is that the use of voting strategies, particularly unanimous strategy combines predictions from all the models and if some models perform worse than the others, they do not aid the performance of the overall ensemble but keep it constant, or even worse than the single best detector. The use of affirmative and consensus techniques produced the best results, implying that the base model detections of some detectors were mainly correct but that even if the other models missed these objects, the unanimous strategy forced them to eliminate the correct bounding box predictions. As expected, unanimous strategy which expects all collaborating models to agree upon all predictions, has the minimum number of detections and an affirmative strategy that accepts any one set of predictions

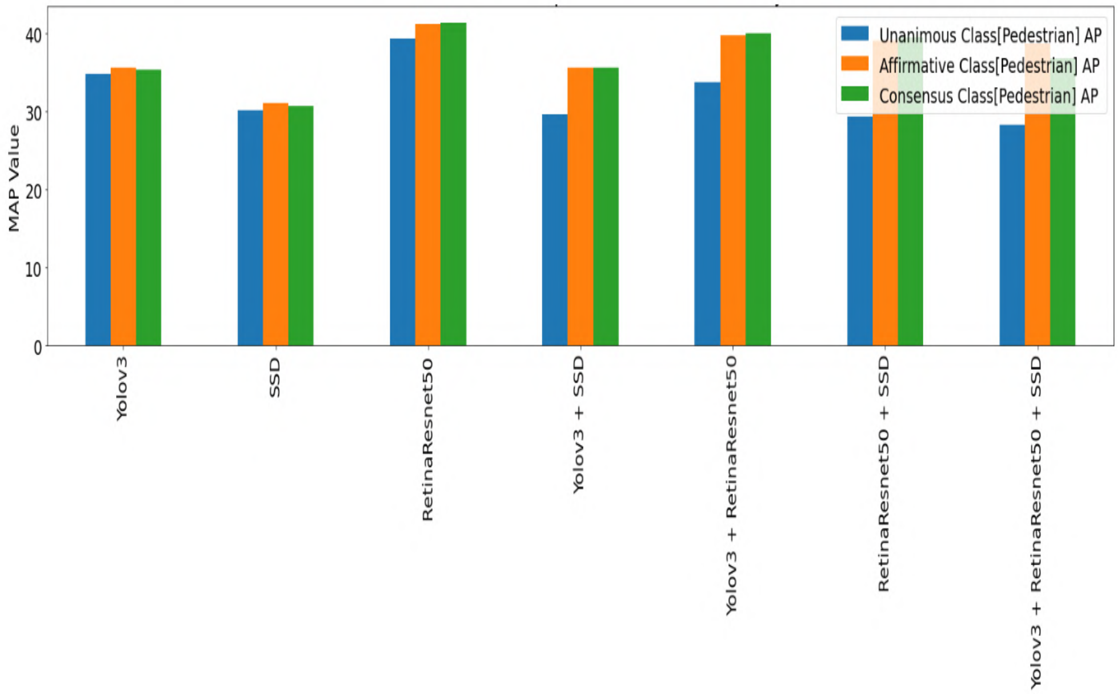


Figure 7. Visualization of Ensemble Framework Results

has the maximum number of detections. The use of voting strategies thus provides explainability and allows the users to understand how the ensemble framework makes the final detections. This voting strategy framework can also be used between models of different parameters, or for users to use for different tasks. For example, in self-driving cars where we require high certainty and explainability for pedestrian detection, the consensus strategy can be leveraged. We intend to investigate the efficacy of the ensemble framework for pedestrian recognition in difficult conditions like lower resolution imagery and to detect special classes such as traffic policemen in future research. The experiments can be broadened to include a larger range of models and can be conducted on a variety of 2D and 3D datasets. In addition to enhancing the performance of models, the bias in existing pedestrian recognition datasets must be investigated in order to ensure the ethical deployment of future AI-based navigation systems and to acquire end-user trust. Pedestrian detection algorithms can be leveraged responsibly for the benefit of society in applications such as traffic safety and management, senior assistance, and security surveillance in the future.

References

[1] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. Real-time pedestrian detection with deep network cascades. 2015. 3

[2] Martim Brandao. Age and gender bias in pedestrian detection algorithms. *arXiv preprint arXiv:1906.10490*, 2019. 2

[3] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300:17–33, 2018. 2

[4] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016. 3

[5] Angela Casado-Garcıa and Jónathan Heras. Ensemble methods for object detection. 2020. 3

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 2

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 3

[8] Piotr Dollár, Boris Babenko, Serge Belongie, Pietro Perona, and Zhuowen Tu. Multiple component learning for object

detection. In *European conference on computer vision*, pages 211–224. Springer, 2008. 2

[9] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014. 3

[10] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 4, 5

[12] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2

[13] Hiroshi Fukui, Takayoshi Yamashita, Yuji Yamauchi, Hironobu Fujiyoshi, and Hiroshi Murase. Pedestrian detection based on deep convolutional neural network with ensemble inference network. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 223–228. IEEE, 2015. 3

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2

[15] Bastian Leibe, Nico Cornelis, Kurt Cornelis, and Luc Van Gool. Dynamic 3d scene analysis from a moving vehicle. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[16] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE transactions on Multimedia*, 20(4):985–996, 2017. 3

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4, 5

[19] Zhe Lin and Larry S Davis. A pose-invariant descriptor for human detection and segmentation. In *European conference on computer vision*, pages 423–436. Springer, 2008. 2

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 5

[21] Subhransu Maji, Alexander C Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2

[22] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3136, 2017. 3

[23] Stefan Munder, Christoph Schnorr, and Dariu M Gavrila. Pedestrian detection and tracking using a mixture of view-based shape–texture models. *IEEE Transactions on intelligent transportation systems*, 9(2):333–343, 2008. 2

[24] L. Nanni and A. Lumini. Ensemble of multiple pedestrian representations. *IEEE Transactions on Intelligent Transportation Systems*, 9(2):365–369, 2008. 3

[25] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Pedestrian detection with spatially pooled features and structured ensemble learning. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1243–1257, 2015. 3

[26] IAD Reading, CL Wan, and KW Dickinson. Developments in pedestrian detection. *Traffic engineering and control*, 36(10):538–542, 1995. 2

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 5

[28] Adrian Rosebrock. Intersection over union (iou) for object detection. *Online] http://www. pyimagesearch. com/2016/11/07/intersection-overunion-iou-for-objectdetection*, 2016. 4

[29] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3626–3633, 2013. 3

[30] Paul Viola, Michael J Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005. 3

[31] Rahee Walambe, Aboli Marathe, and Ketan Kotecha. Multiscale object detection from drone imagery using ensemble transfer learning. *Drones*, 5(3):66, Jul 2021. 3

[32] Christian Wojek, Stefan Walk, and Bernt Schiele. Multi-cue onboard pedestrian detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 794–801. IEEE, 2009. 2

[33] Yanwu Xu, Xianbin Cao, and Hong Qiao. An efficient tree classifier ensemble-based approach for pedestrian detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(1):107–117, 2010. 3

[34] Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2137, 2016. 3

[35] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European conference on computer vision*, pages 443–457. Springer, 2016. 3

[36] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1751–1760, 2015. 3

[37] Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z Li, and Guodong Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE Transactions on Multimedia*, 22(2):380–393, 2019. 3