Gender-based Detection and Tracking of Child Pedestrians using Machine Learning

Conference Paper · July 2024

DOI: 10.1109/ICECET61485.2024.10698292

1 author:

Tyler Ward
University of Kentucky

IV. International Conference on Electrical, Computer and Energy Technologies (ICECET 2024)
25-27 July 2024, Sydney-Australia

Gender-based Detection and Tracking of Child Pedestrians using Machine Learning

Tyler Ward
Department of Engineering Sciences
Morehead State University
Morehead, United States
tbward@moreheadstate.edu

***Abstract*—As the prevalence of fully autonomous vehicles (AVs) on public roads grows in the coming decades, it becomes increasingly important to ensure the safe operation of these vehicles, especially in the presence of pedestrians. A robust perception system in particular is crucial to ensuring safe interactions of these vehicles with humans. It is a noted issue in AV research that these vehicles can sometimes struggle to accurately detect child pedestrians. While there have been advancements made in this area in recent years, there remains a gap in the literature surrounding considerations of a child pedestrian's gender in the detection process. Past research in the field of transportation has shown that the behavior of child pedestrians differs depending on the child's gender, making gender a potentially important consideration in the perception layer of an AV. This paper presents a gender-based pedestrian detection framework based on the YOLOv8 object detection method that is capable of detecting not only whether a child pedestrian is male or female, but also distinguishing them from an adult pedestrian. The system also employs the ByteTrack multiple object tracking technique to track the detected pedestrians through a scene. Performance metrics for this system are included, and the results demonstrate impressive detection and tracking capabilities.**

Keywords*— *Autonomous vehicles, Computer vision, Machine learning, Object detection, Object recognition

I. Introduction and Literature Review

In recent years, there has been a notable increase of autonomous vehicles (AVs) being deployed on public roads around the world. It is estimated that 18.43 million new cars will be sold in 2024 that have a level of automation built in to allow drivers to take their hands off the wheel, and by 2030 close to 95% of all new vehicles on the market could offer a high or full level of automation [1]. Because of this, ensuring that these vehicles can safely interact with human drivers and pedestrians is of paramount importance. Studies show that third parties like pedestrians are the leading cause of accidents involving AVs [2].

An AV's architecture consists of three layers: the perception layer, the planning layer, and the trajectory control layer [3]. The perception layer estimates the environment surrounding an AV by combining real-time information from multiple sensors collects data from exterior sensors [3]. The planning layer determines optimal driving behavior and generations a collision free local trajectory for the vehicle to follow [3]. The trajectory control layer is responsible for the generation of appropriate control commands to follow the planned trajectory in the presence of a dynamic traffic environment [3]. Errors in any of these layers can lead to accidents, although the perception and planning layers are most likely to be affected by pedestrian behavior [2].

Much research has been conducted surrounding the topic of pedestrian detection and tracking for AVs, yet there remain gaps regarding the detection and tracking of child pedestrians specifically [4]. This distinction is important because children are less likely to pay attention when crossing a road [5], are less capable of judging the speed of oncoming traffic [6], and are more likely to behave unpredictably compared to adults [7]. Studies have also shown that child pedestrians behave differently based on their gender [8].

Gender differences in children's pedestrian behaviors were examined in [9]. It was found that boys played on the sidewalk and crossed the road accompanied by an adult more often than girls, while girls walked side-by-side with partners more than boys. It was also found that in middle grades, boys watched traffic more often than girls when crossing the road and ran or hopped on the sidewalk more often than girls. Finally, it was concluded that as the age of the children increased, girls performed more safely in pedestrian tasks, but similar developmental trend was not observed in boys. This finding that male children generally report higher rates of risk-taking behavior than females was validated in [10]. Examples of the riskier behavior exhibited by boys are given in [11], which reported observations that girls waited longer before crossing the road than boys and paid more attention to traffic, while boys missed fewer opportunities to cross the road than girls, and engaged in more anticipations. According to [12], which measured the speed of child pedestrians and found that male children have significantly higher average walking speeds compared to females, the influence of gender becomes more significant in older ages.

To aid the perception capabilities of AVs when encountering child pedestrians, this paper reports the development of a detection and tracking system to distinguish between male and female child pedestrians, as well as to distinguish child pedestrians from their adult counterparts. This system is based on the YOLOv8 object detection (OD) method and the ByteTrack multiple object tracking (MOT) scheme. Previous methods of computationally determining the gender of a child are discussed below.

3D body metrics such as height, shoulder and hip breadth, and head length were used in [13] to estimate age and determine gender. For gender detection, the ratio of shoulder breadth to height and the ratio of hip breadth to shoulder breadth were used, while age estimation was performed based on the proportion of body height to the head height. The authors found that their age estimation model performed better than their gender detection method, with a correlation coefficient of 0.84 and a mean absolute error of 0.9287.

Several of the same authors as [13] performed a later study [14] on a similar topic involving child age and gender determination. This time, two hierarchically arranged self-organizing neural networks were used to process vectors containing 3D body information. This method achieved an average accuracy of 95.2% for age estimation and 90.3% for gender estimation, demonstrating an improvement on their previous study.

Texture descriptors, including histogram of oriented gradients (HOG), uniform local binary patterns (ULBP), and Gabor wavelet transform (GWT), are used in [15] to identify discriminative features that are effective in gender classification. The authors then employ a support vector machine (SVM) to classify the gender of a child, and compare three different kernels for the SVM model to determine which configuration yields the best results. They found that SVM with a cubic kernel function offers the highest accuracy of 89% when classifying based on features extracted from the HOG technique.

A parallel CNN was used in [16] to discriminate gender, classify age groups, and recognize carried objects from gait energy images (GEIs). The authors' methodology for the determination of age and gender involves the used of customized filters in the first convolutional layer of the CNN to extract relevant features from the GEIs. For gender discrimination, the extracted features included postural sway, waist-hip ratio, and shoulder-hip ratio. For age classification, the stability and regularity of the gait was considered, as well as stride length, and dynamic changes in posture.

Real-time gender recognition is explored in [17]. A progressive calibration network (PCN) is used to detect faces regardless of their orientation, and a Gabor filter is used to extract edge and texture features from the detected faces. A mean discrete wavelet transform (meanDWT) is used to overcome some limitations of the Gabor filter, such as redundancy and large dimensions. Different machine learning classifiers including Naïve Bayes, logistic regression, and SVM with linear and radial basis function (RBF) kernels are then compared to determine which offers the best gender classification performance. It was determined that the SVM approach with a linear kernel offered the best performance.

The YOLOv8 method used in this research has the potential to improve on the methods previously explored in the literature in various ways. For example, YOLOv8 is capable of automatically extracting features from the raw images, a capability that surpasses traditional methods like HOG, ULBP, and GWP which were used in [15]. YOLOv8 could also be trained on a multi-task learning framework to simultaneously predict characteristics like age, gender, and other attributes like carried objects which were discussed in [16]. It is also capable of operating with compact yet effective representations of the input data, allowing it to avoid challenges related to redundancy and the handling of large dimensions which were concerns raised in [17]. Overall, compared to the methods explored in the existing literature, the YOLOv8 model will provide a robust and accurate method of identifying child pedestrians based on their gender.

II. METHODOLOGY

A. Data Source

To facilitate the goals of this research, the dataset compiled in [18] was used. The original purpose of this dataset was to enable distinction between adult and child pedestrians for AVs, so the images contained within were also suitable for male vs. female classification of child pedestrians as well. The original dataset contained 150 images of adult and child pedestrians, evenly split between images focused on men, women, boys, and girls, with two class labels for the pedestrians: 'adult' or 'child'. However, because the goal of this research was to focus on the child pedestrians, less images of adults were needed, so images containing only adult pedestrians were removed from the dataset, as to not create a class imbalance skewed towards adults when the existing 'child' class was split.

Once the adults only images were removed from the dataset, the images containing a pedestrian classified as a child were re-annotated so instead of being labeled as simply 'child', the pedestrian was labeled either 'boy' or 'girl'. Once this re-annotation was complete, there were 167 adults, 93 boys, and 86 girls depicted in the images in the dataset. There were 73 total images with an image size of 0.17 mp and 346 annotations with an average of 4.6 annotations per image. When categorized by pixel size, there were 48 medium sizes images in the dataset, 11 large images, and 14 jumbo images. Fig. 1 shows the number of objects in each class in the dataset, Fig. 2 shows the size distribution of the images in the dataset, and Fig. 3 shows the heatmaps for the objects in each of the three classes.

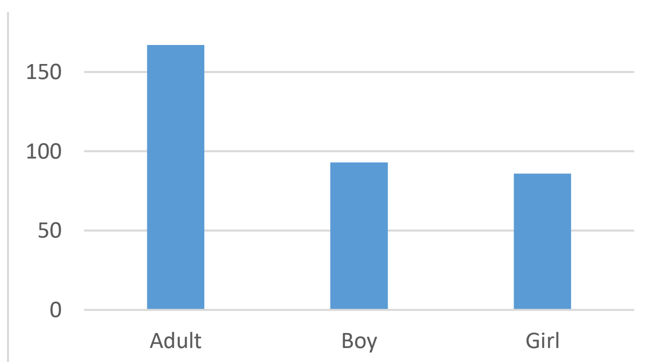

Fig. 1. Number of objects per class in the dataset

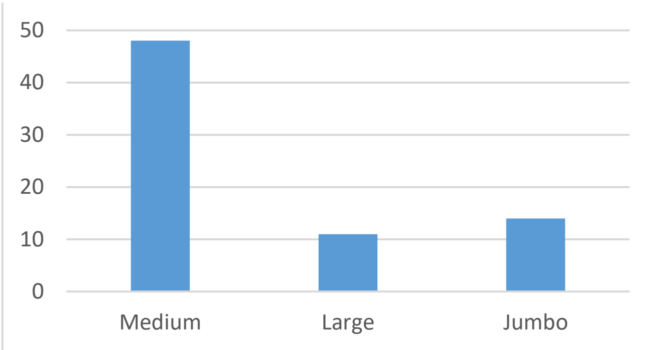

Fig. 2. Size distribution of the images in the dataset

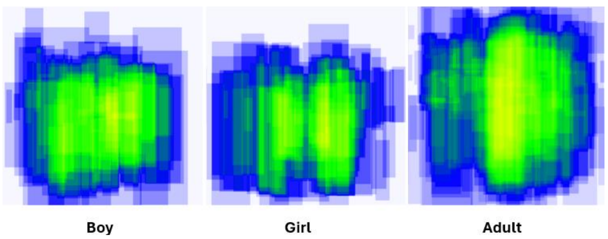

Fig. 3. Heatmaps for each of the three classes

B. Preprocessing

Studies show that data preprocessing goes a long way towards fixing any problems and is an important step when preparing to work with data in any capacity [19]. The images in the dataset underwent an auto-orientation process during the preprocessing phase to ensure that the ordering of pixels was uniform across all of the images. Auto-orientation was selected as the preprocessing technique to be used based on findings from previous research into child pedestrian detection that found that this technique most improved the final results of the YOLOv8 model [4].

C. Augmentation

As with preprocessing, augmentation has emerged in recent years as a powerful method for computer vision and other deep learning tasks [19]. One augmentation technique to our dataset prior to model training: mosaic augmentation. Mosaic augmentation takes random sections of four different images, combines them into a single new image, and adds this new image to the dataset. This allows the model to be exposed to a variety of background and object arrangements within an image, allowing them to become more robust to different object scales, rotations, and occlusions.

Similar to the auto-orientation technique applied in the preprocessing phase, mosaic augmentation was selected based on comparative findings from previous research on the task of child pedestrian detection [4]. After applying the mosaic technique, we were left with 197 images total in our dataset. An example image after undergoing mosaic augmentation is shown in Fig. 4.

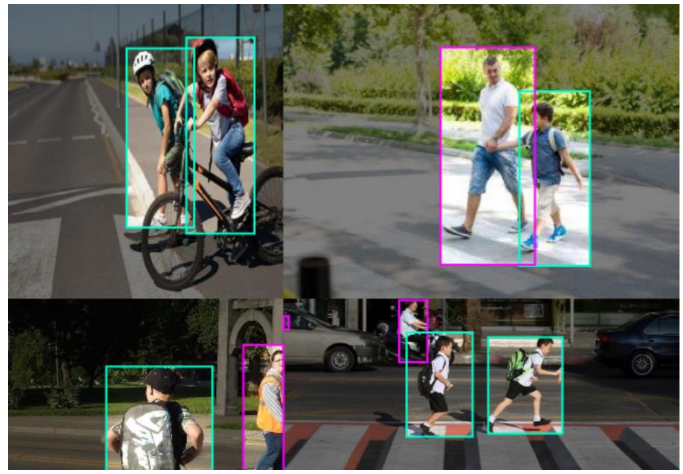

Fig. 4. An image in the dataset after mosaic augmentation

D. YOLOv8

This research employs the YOLOv8 OD method to detect child pedestrians and determine their gender. YOLOv8 is the latest in the You Only Look Once (YOLO) family of computer vision models, first proposed in [20]. YOLO is in iterative development, meaning multiple entities are working on updating and improving the model constantly. YOLOv8 was developed by [21].

YOLOv8 can perform detection, classification, segmentation, pose estimation, and tracking tasks, and is both smaller and faster than its predecessors while achieving better results across a variety of performance metrics. The architecture of YOLOv8 can deviate slightly depending on which task is being performed. As our goal was OD, the architecture of the YOLOv8 used in this research consisted of seven convolutional layers, eight fast cross stage partial bottleneck layers with two convolutions (C2f), one spatial pyramid pooling – fast (SPPF) layer, two upsampling layers, four concatenation layers, and one detection layer.

In the convolutional layers, a 2D convolution is performed which captures spatial hierarchies in the input data. Following this process, 2D batch normalization is applied, stabilizing the learning process by normalizing the output of the convolution. This normalization is crucial for training acceleration, as it addresses internal covariate shift. After the normalization is complete, the Sigmoid-weighted linear unit (SiLU) activation function introduces non-linearity to the model to assist with learning complex patterns.

The C2f layers are comprised of two convolutional layers and aim to enhance computational efficiency while maintaining or improving the learning capabilities of the network. The first convolutional layer expands the channel dimension, while the second aggregates the outputs from a series of bottleneck blocks and the input. These bottleneck blocks process the split outputs sequentially, facilitating efficient feature extraction and representation. A notable feature of the C2f layers is that they use both chunk and split methods in their forward passes, offering versatility in handling tensor operations.

The SPPF layer, which is a streamlined version of a spatial pyramid layer, is designed to enhance the receptive field and aggregate contextual information from different scales. The layer consists of two convolution blocks and a max-pooling layer. The first convolutional block reduces the channel dimensions, preparing the input for subsequent pooling. During the forward pass, the input undergoes a series of transformations. It is first processed through the initial convolution, followed by the max-pooling layer. This pooling operation is repeated, generating multiple pooled outputs that represent different spatial resolutions, which are then concatenated with the initial convolution output. This concatenated tensor is then passed through the second convolutional block.

The upsampling layers enable efficient and accurate resizing of input data. The concatenation layers concatenate a list of tensors along a specified dimension. The detection layer serves as the output layer for the network, processing and interpreting the features extracted by previous layers, converting them into OD predictions. The full architecture of the YOLOv8 model can be seen in Fig. 5.

E. ByteTrack

Traditional MOT approaches rely heavily on high confidence detections [22]. ByteTrack, a MOT method devised in [22], seeks to address the critical issue of missed detections and fragmented trajectories due to object occlusion or low detection scores. Satisfactorily addressing these issues contributes greatly to a robust and accurate MOT system.

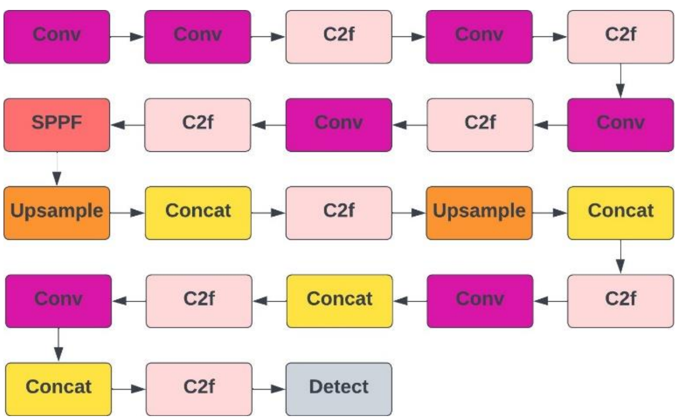

Fig. 5. The architecture of the YOLOv8 model

The novelty behind ByteTrack lies in an association metric that considers all detection boxes regardless of their confidence scores. By leveraging similarities with existing tracklets, ByteTrack effectively recovers objects from background detections, a capability that previous MOT methods struggled with. ByteTrack's inclusive approach to detection boxes ensures that more objects are correctly tracked, even in scenarios where objects are partially occluded or blurred.

The authors of [22] demonstrated that ByteTrack achieved significant performance improvements on standard benchmarks for MOT. ByteTrack's generic nature allows it to be seamlessly integrated into existing tracking frameworks, offering substantial improvements in identity preservation and the reduction of identity switches. It is because of these robust capabilities, as well as its seamless integration with YOLOv8 that ByteTrack was selected as the MOT system to be used for this research.

III. Results

It is important to evaluate the classification process and measure the performance of the YOLOv8 implementation. Methods of evaluating the performance of OD models involved the use of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. The evaluation metrics are defined as:

1) *Precision (P):* The percentage of correctly classified positive cases relative to the cases classified as positive.

$$P = \frac{TP}{TP + FP} \tag{1}$$

2) *Recall (R):* The percentage of positive cases that were successfully classified as positive.

$$R = \frac{TP}{TP + FN} \tag{2}$$

3) *mean Average Precision (mAP):* The average precision (AP) of each class over a number of classes (n).

$$mAP = \frac{1}{n}\sum_{i=1}^{N} AP_i \tag{3}$$

The YOLO algorithm has several parameters that can affect its performance. The image size, batch size, the number of epochs, and the weights used can all be modified, and different configurations of these parameters can affect the results. The YOLOv8 implementation used in this research was trained with an image size of 640x640 pixels, a batch size of 16, 300 epochs, and the YOLOv8x weights. The results of the training process are shown in Table 1.

Compared to the results of several of the previously discussed works, the methods described in this paper represent an improvement in accuracy on these previous models. Compared to [14], which presents a gender classification accuracy of 90.2%, the YOLOv8 model achieves an accuracy of 99.05% and 92.4% for the mAP50 and mAP50-95 metrics respectively when the rates for accurate classification of boys and girls are averaged. In [15], the highest accuracies achieved by the authors from the use of the SVM classification algorithm with a cubic kernel were 92% for identification of boys and 86% for girls. The results of the YOLOv8 model were 99.5% and 98.6% respectively for boys and girls when measured by mAP50, representing an improvement on the results of [15].

A bar graph representation of the accuracy values is shown in Fig. 6. A confusion matrix representing the breakdown in predicted vs. actual classes is shown in Table 2. A sample prediction from the proposed method is included in Fig. 7.

Speed metrics for the ByteTrack model are shown in Table 3. A sample series of frames from the ByteTrack model is shown in Fig. 8.

The confusion matrix, which is a simple method of visualizing the results of a classification problem, shows that of the 21 samples in the adult class, 19 were correctly classified as an adult, while two were incorrectly classified as the background. For samples in the boy class, 22 were correctly classified and one was misclassified as background. For the samples in the girl class, nine were correctly classified and one of was misclassified as an adult. Additionally, there was one object in the background that was misclassified as a girl.

The measured speed metrics include the speed of preprocessing, inference, and postprocessing. For OD methods like YOLOv8, the preprocessing speed is the rate at which the raw input data is prepared into a format that can be processed by the model, the inference speed is the speed at which the actual computation by the neural network takes place, and the postprocessing speed is how quickly the raw predictions from the model are interpreted into a usable format. The results show that the YOLOv8 method used in this research achieved a preprocessing speed of 4.1 ms, an inference speed of 44.2 ms, and a postprocessing speed of 6.5 ms.

The short preprocessing time is very beneficial for real-time applications such as pedestrian detection where the processing of the input data must keep up with live feeds. The achieved inference speed, while not exceptionally fast, is also reasonable for many real-world applications. The postprocessing speed signals the capability of the system to make timely and actionable predictions.

TABLE I. TRAINING RESULTS OF THE YOLOV8 MODEL

Class	Precision	Recall	mAP50	mAP50-95
Overall	96.8%	97.5%	99.2%	93.1%
Adult	100.0%	99.7%	99.5%	94.7%
Boy	100.0%	99.4%	99.5%	91.4%
Girl	90.3%	93.3%	98.6%	93.4%

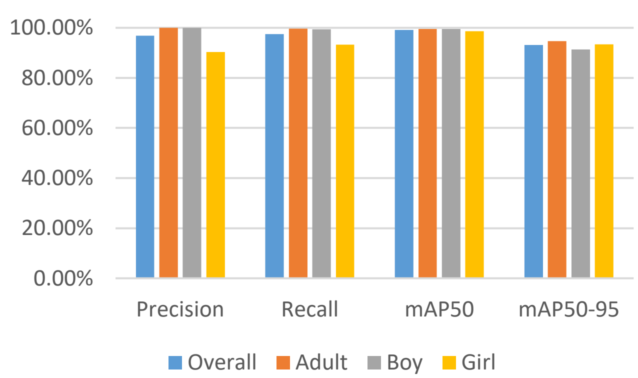

Fig. 6. Bar chart representation of the YOLOv8 performance metrics

TABLE II. CONFUSION MATRIX FOR THE YOLOV8 MODEL

Predicted \ True	Adult	Boy	Girl	BG
Adult	19	0	0	2
Boy	0	22	0	1
Girl	1	0	9	0
BG	0	0	1	0

Fig. 7. Samples of image results from the YOLOv8 model

TABLE III. SPEED METRICS FOR THE BYTETRACK MODEL

Preprocessing	Inference	Postprocessing
4.1 ms	44.2 ms	6.5 ms

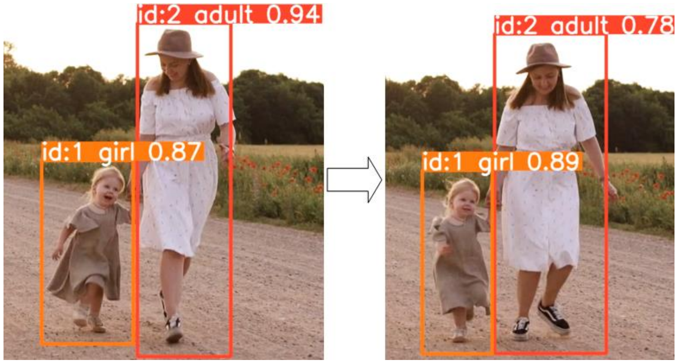

Fig. 8. Sample results from the ByteTrack model

From these results, it is apparent that the YOLOv8 model is very capable and highly accurate at both distinguishing child pedestrians from adults, and also determining the gender. Overall, the system achieves a mAP50 value of 99.2%, with the breakdown of adults, boys, and girls being 99.5%, 99.5%, and 98.6% respectively. The slightly lower accuracy for girls could be due to the fact that there were less images of girls in the original dataset compared to boys and adults. The ByteTrack MOT system demonstrates accurate real-time tracking of adult and child pedestrians.

IV. Conclusion

In this paper, we proposed a pedestrian detection and tracking framework to distinguish between adult and child pedestrians, as well as to determine the gender of the child pedestrians. This framework was based on the YOLOv8 OD method and the ByteTrack MOT technique. It is observed that the OD and MOT model is able to accurately detect and track adult and child pedestrians and determine their gender in the majority of cases. As a result, it is an ideal choice for implementation in AVs.

Future work in this area could revolve around the implementation of speed estimation or trajectory prediction models to further aid the perception capabilities of an AV. Additionally, transfer learning could be implemented into this system to build a system capable of detecting road signs or animals. Another potential area of interest lies in the development of predictive modeling based on the gender of the child pedestrian, such as determining the likelihood of certain actions, such as running out into the road, based on the gender and historical video data.

References

[1] A. Sanghavi, "50 Autonomous vehicle statistics to drive you crazy in 2024," *G2*, Feb. 13, 2024. https://www.g2.com/articles/self-driving-vehicle-statistics/.

[2] J. Wang, L. Zhang, Y. Huang, and J. Zhao, "Safety of autonomous vehicles," *J. Adv. Transp.*, vol. 2020, October 2020, Art. no. 8867757, doi: 10.1155/2020/8867757.

[3] O. Sharma, N.C. Sahoo, and N.B. Puhan, "Recent advances in motion and behavior planning techniques for software architecture of autonomous vehicles: A state-of-the-art survey," *Eng. Appl. Artif. Intell.*, vol. 101, p. 104211, May 2021, doi: 10.1016/j.engappai.2021.104211.

[4] T. Ward, "Development of detection and tracking systems for autonomous vehicles using machine learning." M.S. thesis, Department of Engineering & Technology Management, Morehead State University, Morehead, KY, 2023.

[5] A. Rasouli, I. Kotseruba, and J.K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, Los Angeles, CA, USA, 2017, pp. 264-269, doi: 10.1109/IVS.2017.7995730.

[6] M. Jabbari et al., "The pedestrian network concept: A systematic literature review," *J. Urban Mobil.*, vol. 3, p. 100051, December 2023, doi: 10.1016/j.urbmob.2023.100051.

[7] M.N. Khan and S. Das, "Advancing traffic safety through the safe system approach: A systematic review," *Accid. Anal. Prev.*, vol. 199, p. 107518. May 2024, doi: 10.1016/j.aap.2024.107518.

[8] A. Deluka-Tibljaš, S. Šurdonja, I.I. Otković, and T. Campisi, "Child-pedestrian traffic safety at crosswalks – literature review," *Sustainability*, vol. 14, no. 3, p. 1142, January 2022, doi: 10.3390/su14031142.

[9] H. Wang, D.C. Schwebel, D. Tan, L. Shi, and L. Miao, "Gender differences in children's pedestrian behaviors: Developmental effects," *J. Saf. Res.*, vol. 67, pp. 127-133, December 2018, doi: 10.1016/j.jsr.2018.09.003.

[10] H. Wang, M. Wu, X. Cheng, and D.C. Schwebel, "The road user behaviours of Chinese adolescents: Data from China and a comparison with adolescents in other countries," *Ann. Glob. Health*, vol. 85, no. 1, May 2019, doi: 10.5334/aogh.2452.

[11] B.K. Barton and D.C. Schwebel, 'The roles of age, gender, inhibitory control, and parental supervision in children's pedestrian safety," *J. Pediatr. Psychol.*, vol. 32, no. 5, pp. 517-526, June 2007, doi: 10.1093/jpepsy/jsm014.

[12] M. Simeunović, I. Tanackov, P. Pitka, M. Simeunović, and Z. Papić, "Determination of moving speed of school age children," *Math. Probl. Eng.*, vol. 2021, August 2021, Art. no. 9965753, doi: 10.1155/2021/9965753.

[13] A. Sandygulova, M. Dragone, and G.M. O'Hare, "Real-time adaptive child-robot interaction: Age and gender determination of children based on 3D body metrics," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, Edinburgh, UK, 2014, pp. 826-831, doi: 10.1109/ROMAN.2014.6926355.

[14] A. Sandygulova, Y. Absattar, D. Doszhan, and G.I. Parisi, "Child-centered motion-based age and gender estimation with neural network learning," in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, USA, 2016, pp. 47-52.

[15] A. Iftikhar, R. Ashraf, A. Saeed, S. Alva, and R. Pardeshi, "Automatic gender identification from children facial images using texture descriptors," in *Recent Trends in Signal and Image Processing: Proceedings of ISSIP 2018*, 2019, pp. 43-42, doi: 10.1007/978-981-13-6783-0_5.

[16] N.S. Russel and A. Selvaraj, "Gender discrimination, age group classification and carried object recognition from gait energy image using fusion of parallel convolutional neural network," *IET Image Process.*, vol. 15, no. 1, pp. 239-251, January 2021, doi: 10.1049/ipr2.12024.

[17] S.K. Gupta, S.H. Yesuf, and N. Nain, "Real-time gender recognition for juvenile and adult faces," *Comput. Intell. Neurosci.*, vol. 2022, March 2022, Art. no. 1503188, doi: 10.1155/2022/1503188.

[18] T. Ward, S. Rashad, and H. Elgazzar, "Machine learning based pedestrian detection and tracking for autonomous vehicles," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2023, pp. 1294-1299, doi: 10.1109/CCWC57344.2023.10099089.

[19] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augementation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91-99, June 2022, doi: 10.1016/j.gltp.2022.04.020.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recogntiion (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[21] G. Jocher, A. Chaurasia, and J. Qiu, (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics.

[22] Y. Zhang et al., "ByteTrack: Multi-object tracking by associating every detection box," in *Computer Vision – ECCV 2022*, Tel Aviv, Israel, 2022, doi: 10.1007/978-3-031-20047-2_