Comparative Discussion of Model Performance

YOLOv26n Pedestrian Behavior Detection — Per-Class Analysis

Authors:

John Bernard B. Barce

Luis Arnold N. Respecio

1. Overview of the Models and Their Configurations

Three YOLOv26n models were trained and evaluated on a pedestrian behavior dataset composed of four classes: crossing, jaywalking, pedestrian, and waiting. Each model was trained with a distinct combination of optimizer, learning rate, batch size, and number of epochs, allowing for a direct comparison of how different training configurations affect per-class detection performance.

Model 1 used the AdamW optimizer with a learning rate of 0.01, a batch size of 4, and was trained for 25 epochs. Model 2 used the SGD optimizer with a lower learning rate of 0.001, a larger batch size of 20, and was trained for 30 epochs. Model 3 used automatic optimizer selection with the lowest learning rate of 0.0001, an auto-determined batch size, and was trained the longest at 40 epochs. The following sections compare their performance across overall and per-class metrics.

2. Overall Performance Comparison

The table below summarizes the overall validation metrics for all three models across the entire dataset.

Model	mAP50	mAP50-95	Precision	Recall	F1 Score
Model 1 (AdamW)	0.5430	0.3180	0.5000	0.6140	0.5510
Model 2 (SGD)	0.4340	0.2540	0.4080	0.5390	0.4660
Model 3 (Auto)	0.6980	0.4470	0.6750	0.6600	0.6670

Model 3 achieved the best results across all metrics, with a mAP50 of 0.698, precision of 0.675, recall of 0.660, and an F1 Score of 0.667. Model 1 came in second, reaching a mAP50 of 0.543 and showing higher recall (0.614) than precision (0.500), which suggests it tended to over-predict. Model 2 performed the worst, with a mAP50 of 0.434 and an F1 Score of 0.466, indicating that using the SGD optimizer with a low learning rate and large batch size did not work well on this dataset.

All three models showed a large gap between mAP50 and mAP50-95 (ranging from 0.225 to 0.251). This is common in object detection and means that, although the models can localize objects fairly well at a lower IoU threshold, accurately drawing bounding boxes at stricter thresholds is still difficult for all model setups.

3. Per-Class Performance and Struggle Analysis

The table below provides a detailed breakdown of per-class metrics across all three models.

Model	Class	AP50	Precision	Recall	F1 Score
Model 1	Crossing	0.5140	0.4790	0.6750	0.5540
Model 1	Jaywalking	0.8340	0.6590	0.8590	0.7440
Model 1	Pedestrian	0.2510	0.3700	0.2090	0.2710

Model 1	Waiting	0.5720	0.4930	0.7120	0.5800
Model 2	Crossing	0.3960	0.3720	0.6070	0.4630
Model 2	Jaywalking	0.7340	0.6060	0.7490	0.6710
Model 2	Pedestrian	0.2060	0.3110	0.1970	0.2430
Model 2	Waiting	0.3990	0.3430	0.6030	0.4360
Model 3	Crossing	0.7350	0.6740	0.7170	0.6940
Model 3	Jaywalking	0.9050	0.8510	0.8650	0.8580
Model 3	Pedestrian	0.4530	0.5150	0.3690	0.4300
Model 3	Waiting	0.6990	0.6590	0.6890	0.6740

3.1 Jaywalking (The Strongest Class Across All Models)

Jaywalking was consistently the best-detected class for all three models. Model 3 led with an AP50 of 0.905, a precision of 0.851, a recall of 0.865, and an F1 Score of 0.858. Model 1 achieved an AP50 of 0.834 and an F1 Score of 0.744, while Model 2 posted an AP50 of 0.734 and an F1 Score of 0.671. The relatively strong performance on jaywalking across all models is likely attributable to its high instance count in the validation set (331 instances) and the fact that jaywalking behavior may exhibit more visually distinctive characteristics such as mid-road positioning or a particular posture that the model can latch onto.

Even so, the precision-recall balance for jaywalking is not perfect. In Model 1, recall (0.859) far exceeded precision (0.659), suggesting the model detected most true jaywalking instances but also produced a notable number of false positives. Model 3 corrected this imbalance considerably, with precision (0.851) and recall (0.865) nearly aligned, reflecting a more calibrated detector.

3.2 Pedestrian (The Most Consistently Struggled Class)

The pedestrian class was the weakest-performing class for all three models by a significant margin, and represents the most critical area of failure in this experiment. Despite having the highest instance count in the validation set (396 instances), all three models struggled dramatically to detect and localize pedestrians accurately.

Model 1 produced an AP50 of only 0.251 and an F1 Score of 0.271. The recall of 0.209 is particularly alarming, it means the model detected fewer than 21% of actual pedestrian instances. Model 2 performed even worse with an AP50 of 0.206 and a recall of 0.197, effectively missing over 80% of pedestrians. Model 3 improved substantially but still underperformed relative to all other classes, reaching an AP50 of 0.453 and an F1 Score of 0.430, with a recall of 0.369.

These results collectively indicate that the pedestrian class poses a fundamental challenge. A key reason for this is likely semantic ambiguity: a pedestrian standing on a sidewalk could visually overlap with a person classified as waiting, or a person stepping off a curb could be confused with either jaywalking or crossing. The pedestrian label may therefore be competing with the other three classes in ways that confuse the model, causing it to misclassify pedestrian instances into one of the more behaviorally specific categories. This hypothesis is supported by the consistently low recall across all models even the best model (Model 3) still missed roughly 63% of pedestrian instances.

The precision values for pedestrians also merit attention. Model 1 shows a precision of 0.370 and Model 3 shows 0.515, meaning that even when the models do predict pedestrians, a large fraction of those predictions are incorrect. This two-sided failure of low precision and low recall confirms that the pedestrian class is not merely underrepresented in predictions, but is genuinely difficult for the model to differentiate from the other behavioral classes.

3.3 Crossing (Moderate Performance with Consistent Improvement)

The crossing class showed moderate but improving performance across the three models. Model 1 achieved an AP50 of 0.514 and an F1 Score of 0.554, Model 2 dropped to an AP50 of 0.396 and an F1 Score of 0.463, and Model 3 recovered strongly with an AP50 of 0.735 and an F1 Score of 0.694. The crossing class has 272 validation instances, and its detection difficulty likely stems from its visual resemblance to jaywalking, both involve individuals traversing a roadway, and the distinction between the two depends on contextual cues such as crosswalk markings that may be partially visible or absent in some images.

Model 2's decline in crossing performance relative to Model 1 is notable. Despite training for more epochs, the SGD optimizer at lr=0.001 and batch=20 appears to have underfit the crossing class, producing lower precision (0.372 vs. 0.479) and lower recall (0.607 vs. 0.675). Model 3's superior crossing performance, nearly doubling Model 2's AP50 suggests that the combination of longer training (40 epochs) and automatic optimizer/batch selection was critical for learning the nuances of this class.

3.4 Waiting (Second Best, But With Recall Dependency)

The waiting class occupies a middle tier across all models. Model 1 achieved an AP50 of 0.572 and an F1 Score of 0.580, while Model 2 dropped to an AP50 of 0.399 and an F1 Score of 0.436. Model 3 again led with an AP50 of 0.699 and an F1 Score of 0.674. Across all models, recall for waiting was substantially higher than precision, which means the models tended to correctly identify most waiting instances but also produced false positives, likely by misclassifying some stationary pedestrians as waiting when they were engaged in other behaviors.

The waiting class (267 instances) shares a visual characteristic with the pedestrian class: both involve relatively stationary individuals. The key difference is context waiting is typically at a crossing point, while pedestrians may involve walking. This contextual dependency may explain why recall for waiting is higher than for pedestrian; the model appears to assign stationary-person detections to waiting more readily than to the general pedestrian label.

4. How the Metric Scores Reflect These Struggles

4.1 Precision

Precision measures how many of the model's positive predictions for a class were correct. The pedestrian class consistently showed the lowest precision of any class across all models (0.370, 0.311, and 0.515 for Models 1, 2, and 3 respectively). This means that even when a model predicted 'pedestrian,' it was wrong more often than it was right in the case of Models 1 and 2. This directly reflects the difficulty of the class boundary between pedestrians and the other three behavioral categories. Jaywalking, by contrast, maintained the highest precision across all models (0.659, 0.606, 0.851), confirming that the model's jaywalking predictions were more reliable.

4.2 Recall

Recall measures how many of the actual instances of a class were successfully detected. The pedestrian class produced recall scores of 0.209, 0.197, and 0.369 across the three models, consistently the lowest recall in the entire per-class breakdown. This means the models were failing to detect the vast majority of actual pedestrian instances in the validation images. For jaywalking, recall was markedly better (0.859, 0.749, 0.865), reinforcing that the model's overall detection ability is not the problem the pedestrian class specifically presents features that are difficult to recognize or distinguish from other classes at inference time.

4.3 F1 Score

The F1 Score is the harmonic mean of precision and recall, and provides a balanced view of a model's performance on a given class. The gap between jaywalking F1 scores (0.744, 0.671, 0.858) and pedestrian F1 scores (0.271, 0.243, 0.430) across the three models is striking and confirms the severity of the pedestrian detection problem. The F1 Score for pedestrian never exceeded 0.43, while jaywalking never fell below 0.67. This consistent divergence across all three models which differ substantially in optimizer, batch size, learning rate, and epochs suggests that the challenge is rooted in the data and class definition rather than any single hyperparameter choice.

4.4 AP50 (Average Precision at IoU = 0.5)

AP50 integrates precision and recall across all confidence thresholds for a class at a 50% IoU overlap requirement. It is the most holistic single-number summary of detection quality per class. The pedestrian AP50 scores (0.251, 0.206, 0.453) were the lowest of all four classes in every model by a substantial margin. The crossing class (0.514, 0.396, 0.735) and waiting class (0.572, 0.399, 0.699) both showed moderate AP50 scores, while jaywalking dominated (0.834, 0.734, 0.905). The AP50 results thus paint a consistent picture: the model architectures and training procedures explored in this experiment are capable of learning behavioral detection reasonably well for three of the four classes, but all fail to adequately capture the pedestrian class.

6. Conclusion

Across all three models and all four evaluation metrics, a clear and consistent pattern emerges: jaywalking was the easiest class to detect, crossing and waiting showed moderate and improvable performance, and pedestrian was the most problematic class for every model tested. The pedestrian class's persistent low precision and recall despite varying optimizers, learning rates, batch sizes, and training durations indicates a structural challenge rooted in class ambiguity and potentially insufficient discriminative visual features that separate a generic pedestrian from someone crossing, waiting, or jaywalking.

The metric scores directly reflect these struggles. Low AP50, low recall, and low F1 Scores for pedestrians across all models confirm that the issue is not simply a matter of insufficient training, but a deeper problem of class overlap in the feature space. Future work should consider revising the class taxonomy, enriching the pedestrian label with more diverse and unambiguous examples, or applying techniques such as focal loss or class-balanced sampling to address the detection disparity. Model 3's overall superiority also confirms that automatic hyperparameter selection with conservative learning rates and extended training is a strong baseline strategy for this type of multi-class pedestrian detection task.