Assessment #3

Research Assessment

Name: Aryan Patel

Date: September 25, 2026

Subject: Machine Learning

MLA Citation:

Yunganina Mamani, Fernando G., et al. "Comparative Evaluation of Deep Learning Object Detectors for Real-Time Parking Occupancy Detection Under Variable Lighting Conditions." *Sensors*, vol. 26, no. 17, 22 Aug. 2026, article 5329, https://doi.org/10.3390/s26175329.

Analysis:

The team was able to train five object detectors with one stationary camera during the day and at night on a college parking lot in Peru to detect free, occupied and unavailable parking spots. All of the models achieved an mAP@0.5:0.95 above 0.985 on the standard detection score with a random train/test split and were within 0.0055 of each other. The scores fell when the test images were taken on different dates and from a different camera angle, and the ranking was almost reversed (Spearman's correlation of -0.80), with Faster R-CNN going from last place to first place. A complete system with YOLOv8s showed 87.5 percent per-slot accuracy in the daytime and 91.8 percent per-slot accuracy at night, on days it was not trained on.

What I remember is that the problem of the random split was not obvious, and no one would ever notice it unless someone is measuring it. Both the camera and the cars remained in their location for hours, with a frame captured every 15 minutes. The authors noticed that 50.2 percent of the test pictures were collected within 15 minutes of a training picture. About half of

the test set was a close copy of the images the model had previously seen, which is why the 0.99 scores were primarily due to the model's ability to remember a single parking lot from a single perspective.

This applies to what I've done too. In my bibliography for Assessment #2, I decided to split my robotics vision tests based on distance and lighting because an average can be a misleading failure indicator as Gender Shades revealed. This paper demonstrates that the split is another area where failure can lie. If it is done from the same clip on which the model has been trained, then slicing by lighting is not beneficial. If the frames in my training sessions and test sessions were recorded from the same practice sessions, it is possible that my scores are higher than they should be, and I'm not sure if this is the case or not at the moment. I had been contemplating it as a set-up a tool would solve.

I would separate the study into three findings: leakage (the authors used the number of images that were similar to the training images to determine the amount of leakage); ranking flip (the occurrence when they compared the models under each split); the difference between the detector and the deployed system. The two are different, but with the strict split the detector scored 0.8828, and the entire system was 87.5 percent accurate during the day. The system also has to map boxes to slots, track cars between frames, and smooth results over time, and mAP never sees these things. It's a tiny slice of the gap I noticed in my first assessment, when 80 percent of workers indicated they felt they were more productive, while only 37 percent of companies reported any impact to earnings. It also alters my belief. I thought that the best architecture would be the most important aspect to get the best vision system, but in this case, the evaluation protocol decided which architecture would be the best. The split is the one thing that was overlooked in the design, like the label in Obermeyer's health algorithm.

The most surprising thing was that the model that was last in the random split was the most accurate in the strict one. I would have given the first table blind trust. It was also encouraging, since it was easy to resplit the data, as opposed to training a larger model. I do have doubts. This ranking change is determined by one piece of data and one set of camera angles and, the authors write, must be repeated. It's noted in the title that it's variable lighting, which is also noted in the limitations section as being between day and night with clear sky conditions. There was another oddity in the paper that I noticed. It says in the text that the gap between the best and worst models increased from 0.0013 to 0.0562, but in the abstract, it is 0.0055 for the random split, and 0.0013 is the gap for mAP@0.5. It does not alter the outcome, but it was a reminder to verify it myself.

I had a question about catching the model that fails before I ship it, after my bibliography. Now I have to wonder, how do I know that my test set is a true test? First, I will check the time or hash of images from my team for near duplicates. Second, I want to train my own object detector twice, once using a random split and once using a split by session or date, but in all other respects keeping the same, and compare the scores. That would be something that I could show a mentor. Third, I will fill out the evaluation data field of the model card that I will write based on the Mitchell and co-authors paper and record the split used. There are some questions that I still have. There are some questions that I still have. Will the ranking vary in other cameras and sites? Are two-stage detectors like Faster R-CNN better at dealing with new viewpoints than one-stage detectors, or is this specific to this data? So how do they break up video when there are only a few hours worth of video?

Annotated article: Annotated Article - Assessment #3.pdf