Comments:

Most of the questions have been addressed carefully, and I appreciate the well-documented answers prepared by the author. Given this, I recommend this paper for publication.

However, I still have one concern regarding the response to my comment 9 about how the Phase 2 model can overcome the inaccurate boundaries in the Phase 1 model. I believe from Phase 1 to Phase 2, there is a huge improvement (Figure 13, 15, 16, and 17), and it deserves more detailed explanations. That is why I was looking for a theoretical explanation in my previous comment. However, the author only claims it is due to the robustness of the Phase 2 model. Again, it would be beneficial for me and other readers to know the mechanism behind the robustness.

My initial guess is that the robustness of the Phase 2 network is owing to the early stopping. The author mentioned that the training data are not all correct. Maybe the early stopping could prevent the Phase 2 network from being overfitted to the incorrect training information.

Also, I am curious about how much percent of the training data is incorrect. I suppose that the F1 scores for Phase 1 CNN tile classification are tile-based ones. If I am correct, what the pixel-based F1 score would be after converting the tile labels to a full-size class raster (e.g., Figure 16c & Figure 17c)?