School mapping in ESRI imagery

Andrija Gorup, Dominik Jambrović, Marin Kačan, Siniša Šegvić

October 24th, 2024

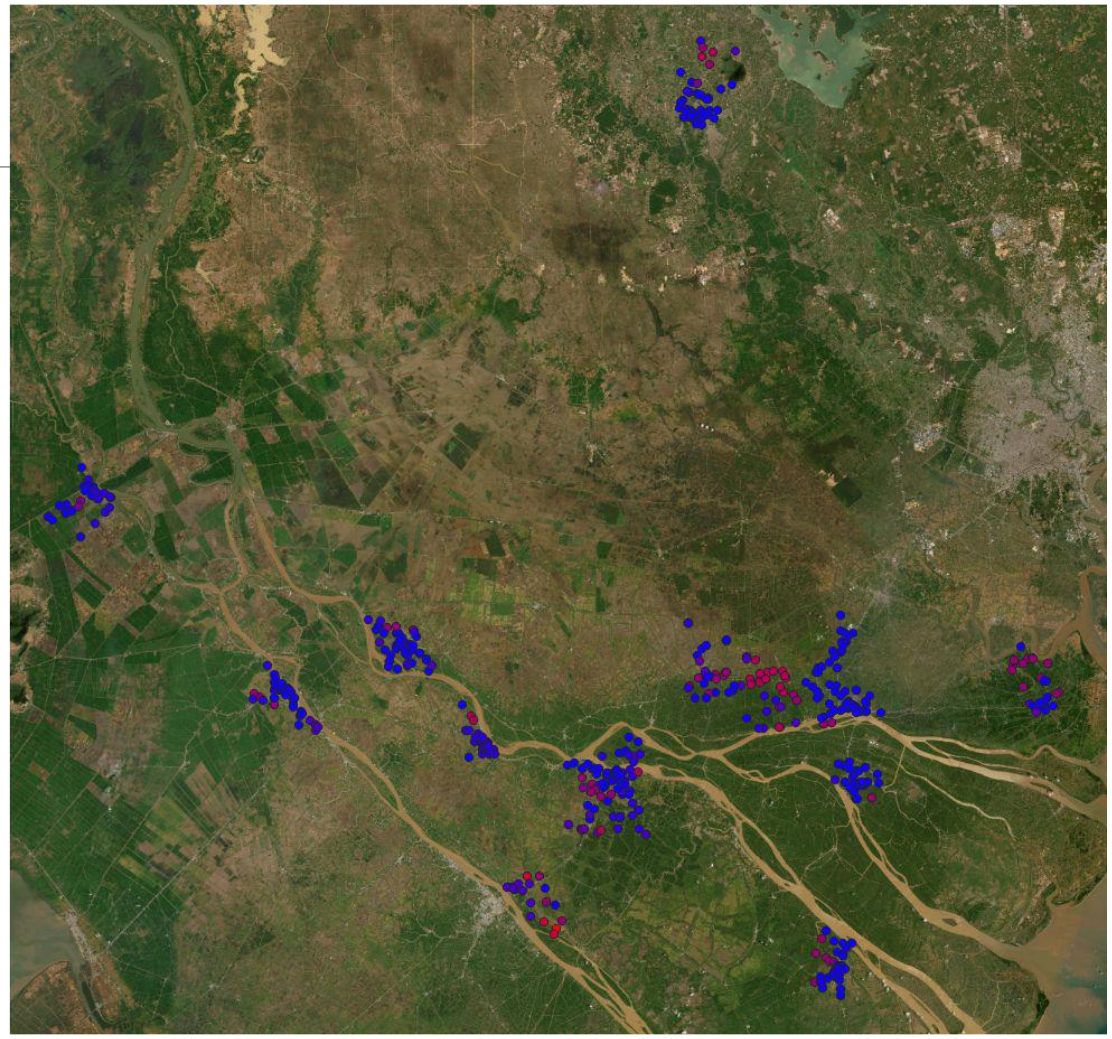
Recap from last meeting

2-stage training, fine-tune and eval on Anditi schools

- 2-fold cross-validation
  - 50:50 train/val split of school locations
  - Add equal number of non-school locations
    - Sampled randomly throughout Vietnam
  - Unusually high results – **F1: 93.91%**
    - In spite of potentially problematic outdated imagery

Potential problems

- **Outdated imagery**
  - Schools visible on Google Maps, but not in ESRI images
- Possible solutions:
  - Urban growth layer
  - ESRI metadata layer
  - Combination of the 2 methods
  - Compressed file size



Potential problems

- **Distinct appearance**
  - Images from different regions have large variations in appearance

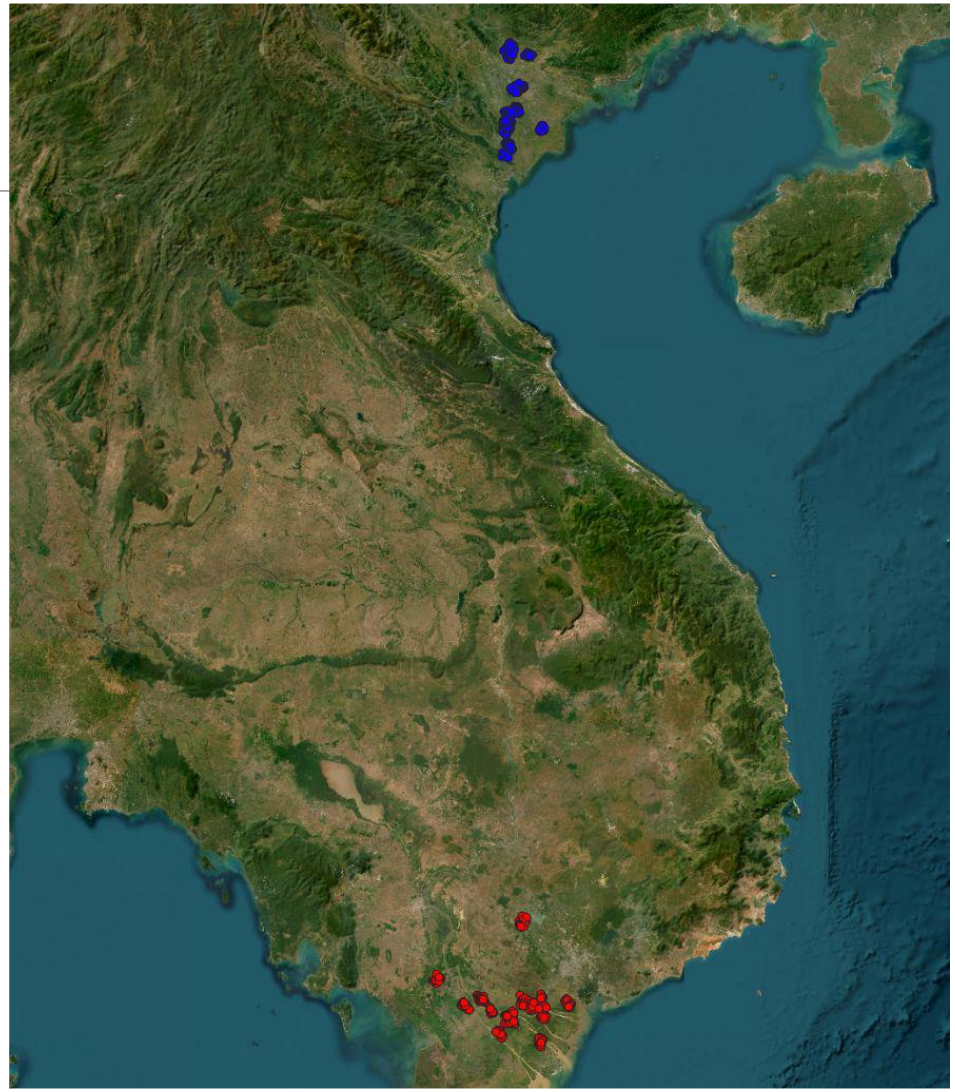
Potential problems

- **Many schools next to each other**
  - For cross-validation experiments, we do a random 50:50 train/val split
  - Data leakage – overlapping images get into both splits

- **Schools in areas with distinct appearance**
  - Similarity due to source of imagery
  - Tiles with very similar appearance get into both splits

Solution – split by clusters

- More realistic results
  - **F1: 81.55%**
  - P: 95.94%
  - R: 73.97%



New contributions

Non-school sampling

- Previously:
  - Non-schools sampled throughout Vietnam
  - Non-schools have no connection to the splits' schools

- **Distance-based non-school sampling**
  - Sample non-schools based on distance from schools

Distance-based non-school sampling

- Schools are split into two clusters
- For each cluster, calculate the corresponding centroid

- For each non-school (OSM Vietnam data), calculate the **Euclidean distance from both centroids**
- Normalize the distances to interval [0, 1]

$$d((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

Distance-based non-school sampling

- Convert distances to probabilities of sampling

$$p_c\,(x) = \frac{1}{d_{c,x}+\varepsilon}$$

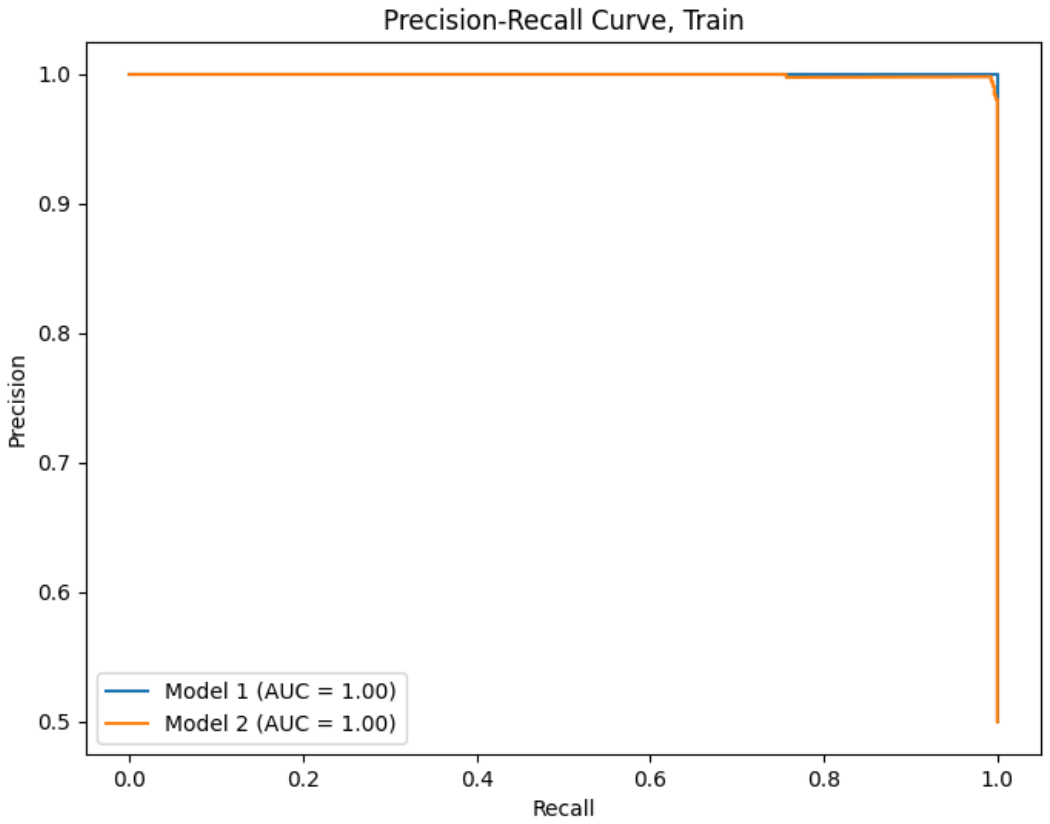
- Sample non-schools for each cluster individually (no repetition)

Results comparison

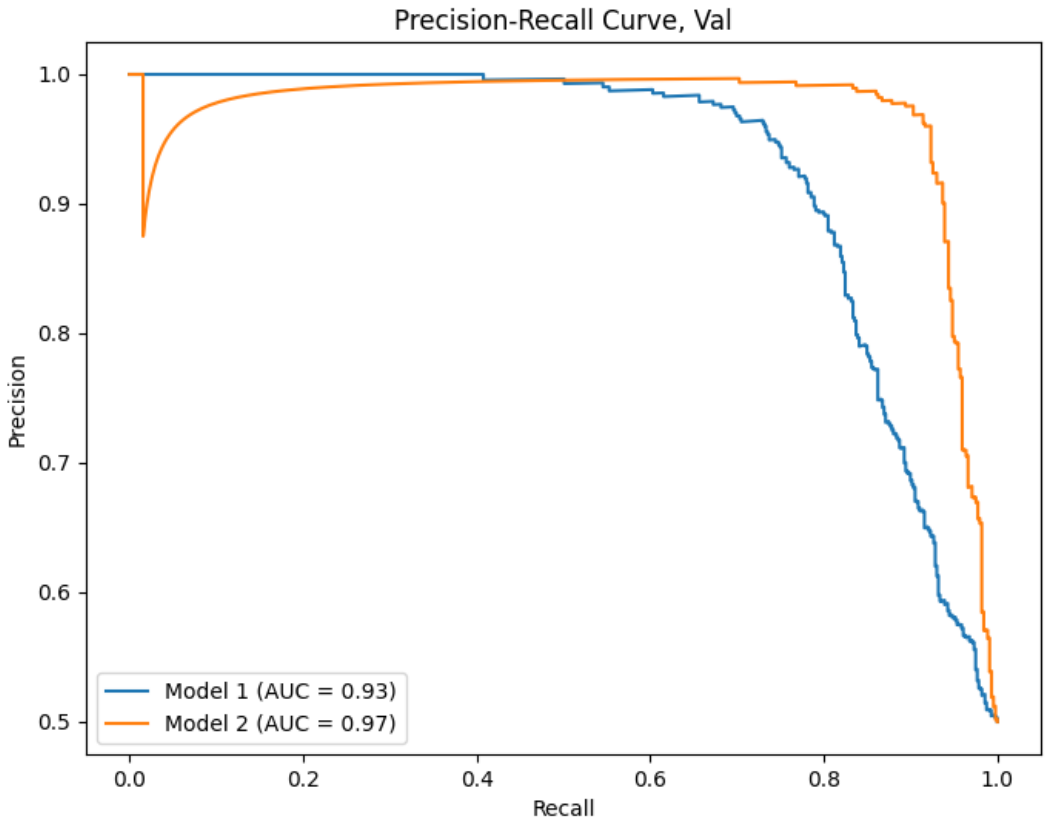
- Random non-school sampling:
  - **F1: 81.55%**
  - P: 95.94%
  - R: 73.97%

- Distance-based non-school sampling
  - **F1: 86.07%**
  - P: 97.71%
  - R: 77.71%

Precision-Recall Curve



Precision-Recall Curve



Future work

- Grayscale images experiment (try to eliminate appearance differences)
- Dense inference (run newly fine-tuned model on tiles covering the chosen 26 districts)