AGILE: GIScience Series, 7, 32, 2026. https://doi.org/10.5194/agile-giss-7-32-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

Object-Level Detection of Hand-Drawn Annotations in Participatory Sketch Maps Using Paired Clean and Annotated Basemaps

Clemens Langer [1,2], Celina Thomé[1,2], Nir Fulman [2], Steffen Knoblauch [1,2,3], Alexander Zipf [1,2,3], and Yulia Grinblat [1]

[1]Heidelberg Institute for Geoinformation Technology (HeiGIT) gGmbH, Heidelberg, Germany
[2]Heidelberg University, Heidelberg, Germany
[3]Interdisciplinary Center for Scientific Computing (IWR/ICSC), Heidelberg University, Heidelberg, Germany

Correspondence: Yulia Grinblat (yulia.grinblat@heigit.org)

Abstract.

Automatic extraction of hand-drawn annotations from participatory sketch maps is essential for digitising community-generated spatial information but remains challenging due to heterogeneous drawing styles, scanning artefacts, and complex basemap content. Existing approaches typically treat markup extraction as pixel-level segmentation or simple image differencing, which struggles under real-world variability. To address this, we formulate annotation extraction as an object-level task using a YOLO-based detector applied to RGB images of annotated maps. In addition, change detection is performed using paired RGB images of annotated and clean maps to isolate user-drawn content from the underlying basemap. Experiments on ∼2,300 real sketch maps and ∼18,000 synthetic samples show strong performance across diverse conditions. Object detection on annotated maps alone achieves mAP@50 of 91.5% on satellite imagery and 97.3% on OSM basemaps, while incorporating paired clean maps for change detection improves performance to 97.4% and 98.1%, respectively. Synthetic pretraining further enhances results on real hand-drawn data, indicating that simulated annotations effectively supplement limited labelled samples.

Submission Type. Model, Algorithm

BoK Concepts. [IP3-4-7] Machine learning; [GC3-12] AI algorithms; [GS4-3b] Citizens and volunteered geographic information

Keywords. Participatory Mapping, Sketch Map Digitization, Object Detection, Change Detection, GeoAI

1 Introduction

Participatory Mapping (PM) complements traditional geospatial data production by allowing local communities to contribute local knowledge and insights that are often overlooked by top-down data collection and therefore absent from official datasets (Laituri et al., 2023). In contexts ranging from urban development to climate adaptation and disaster risk reduction, such inputs help identify hazards, infrastructure needs, and social vulnerabilities. While some participatory mapping workflows use structured acquisition setups that separate sketch content from the basemap and thereby facilitate subsequent digitisation (Chipofya et al., 2021), many real-world applications still rely on participants drawing directly on printed basemaps during field activities. The resulting annotated paper maps are then scanned or photographed, creating *sketch maps* (SMs) that highlight hazards, infrastructure, or places of interest.

Converting these hand-drawn markings into digital geospatial data is challenging due to heterogeneous drawing styles and colours, inconsistent scanning conditions and scales, scanning artefacts, and the visual complexity of underlying basemaps (Boschmann and Cubbon, 2014). Most existing approaches treat markup extraction as a pixel-level problem: the goal is to decide, for each pixel, whether it belongs to the basemap or to a user annotation drawn on it. Early solutions rely on image thresholding, which assumes a uniform annotation style and therefore does not generalize well. More recent work employs pixel-wise segmentation, which classifies every pixel into a category (e.g., “annotation” vs. “background”) by learning patterns in colour and texture directly from training data. An example is SmartLandMaps (Lindner et al., 2023), which segments the annotated image

and then applies rule-based refinement for removing implausible regions.

When a clean basemap is available alongside the annotated version—as is typical in sketch map workflows—the problem can be framed as change detection between two nearly identical images, that is, the objective is to localize content that appears only in the annotated image. Having both images is inherently advantageous: it becomes easier to highlight new markings because the unchanged map content aligns across both images and effectively cancels out. Simple approaches apply pixel-based image differencing between the clean and annotated maps. Examples include the version of the Sketch Map Tool described in Klonner et al. (2021) and the Paper2GIS system (Denwood et al., 2023), both of which align the clean and annotated maps, compute differences, and then use classical segmentation and rule-based noise cleaning to isolate the hand-drawn markings. These methods can work well in controlled conditions but still operate directly on raw pixels, making them sensitive to illumination changes, colour shifts, and scan artefacts.

Another natural step forward is to pose sketch-map markup extraction as an object detection problem. Instead of classifying every pixel as "basemap" or "annotation", an object detector learns to localise instances of hand-drawn content—symbols, short text notes, and line segments—by predicting bounding boxes (and, where applicable, classes) for each marking. This framing aligns well with participatory mapping practice, where annotations are typically interpreted as discrete map elements that can be reviewed, digitised, and linked to downstream GIS workflows. Modern single-stage detectors such as the YOLO family (Redmon et al., 2016) are particularly attractive because they can efficiently detect small, sparse foreground objects against complex backgrounds, and YOLO-style architectures have been successfully adapted in settings where the signal of interest corresponds to changes or additions in paired imagery (e.g. Zhang et al. (2024); Guo et al. (2024)). In the sketch-map context, object detection can be applied in two ways: single-image detection, where the model is given only the annotated map and must separate markings from basemap clutter and scan artefacts, and dual-image detection, where a clean basemap is available in addition to the annotated version and the model can leverage the pair to focus explicitly on what has been added by the participant.

In this work, we investigate how the availability of a clean basemap affects sketch-map annotation detection when the task is framed as object detection. Specifically, we compare single-image detection, where the model receives only the annotated sketch map, with dual-image detection, where the model receives both the clean and annotated maps and can exploit their correspondence to focus on additions made by participants. We ask: (i) How does dual-image detection compare to single-image detection for detecting sketch-map annotations under realistic scanning and photographic conditions? (ii) How well do single- and dual-image detectors generalise across heterogeneous basemaps, such as vector-rendered (OSM-style) maps and satellite imagery (e.g., Esri World Imagery, EWI)? (iii) Can large-scale synthetic pretraining improve performance and reduce reliance on scarce manually annotated data, and does it benefit single- and dual-image settings differently?

Based on the above considerations, we hypothesise that dual-image detection will be more robust than single-image detection in the presence of basemap clutter, illumination variation, and scan artefacts, and that this advantage will be particularly pronounced when transferring between vector and imagery basemaps. We further hypothesise that synthetic pretraining will mitigate data scarcity and improve generalisation in both settings, with the largest gains expected where annotation signals are weakest or most confounded by background variability.

To address these questions, we generate approximately 18,000 synthetic sketch maps over OpenStreetMap (OSM) and Esri World Imagery (EWI) basemaps by simulating hand-drawn markings and realistic photographic artefacts, and we evaluate performance on about 2,300 real printed sketch maps that were scanned or photographed and annotated with bounding boxes. We instantiate YOLOv9e-based detectors in two configurations: a single-image model operating on the annotated map alone ($\text{RGB}_{\text{annotated}}$), and a dual-image Siamese model that takes the clean and annotated maps as paired inputs ($\text{RGB}_{\text{clean}}$, $\text{RGB}_{\text{annotated}}$) using shared-weight feature extraction and feature-level comparison to emphasise annotation-specific changes. To disentangle the effects of input regime and training data, we evaluate each configuration both when trained only on real data and when pretrained on synthetic data and fine-tuned on real maps.

2 Methodology and Dataset

2.1 Datasets

Each sample in our dataset consists of a pair of maps:

(1) a clean basemap (no user markup) from an OSM (vector-rendered) and EWI (satellite imagery).

(2) an annotated map where hand-drawn markings appear on top of the same basemap.

Hand-drawn data. We use ∼2,300 printed sketch maps created during two internal data-curation rounds conducted in 2024–2025, involving 5—7 student assistants in Germany. The maps are based on both vector and satellite basemaps from regions across the world and selected to cover different spatial scales, from neighbourhood to city or district level, with a balanced mix of urban and natural settings. Annotations were created with different pens and markers, resulting in heterogeneous styles, colours, and stroke thicknesses.

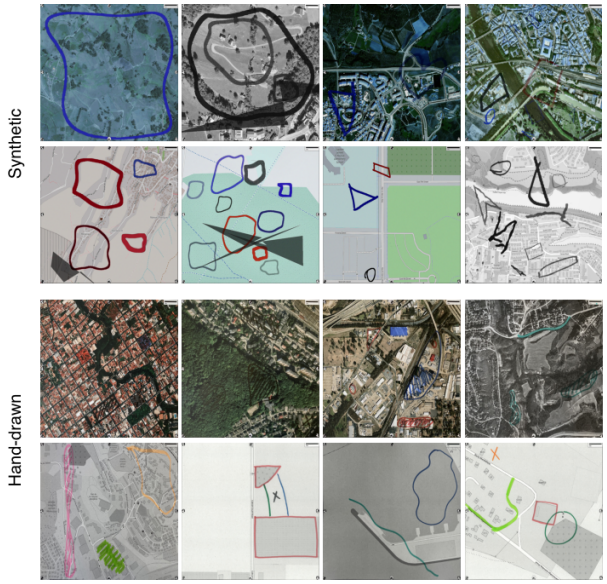

Figure 1. Synthetic (top) and real hand-drawn (bottom) SMs on OSM and EWI basemaps.

The annotated maps are scanned or photographed and the markings are labelled with bounding boxes using CVAT (CVAT.ai Corporation, 2022), while the corresponding clean basemap is kept as the second image in the pair. Multiple annotators followed shared guidelines and labels were cross-checked for consistency. We deliberately included irregular, small, partially overlapping, and off-basemap markings to reflect realistic sketch-map behaviour and to prevent the detector from implicitly relying on perfect geographic alignment. Example real-world basemap–annotation pairs are illustrated in Figure 1 (bottom row).

Synthetic data. To mitigate data scarcity and expose the detector to diverse basemaps and annotation styles, we also generated ∼18k synthetic sketch-map pairs over OSM and EWI basemaps. For each basemap, we create a corresponding annotated map by overlaying hand-drawn-like shapes. We obtain these from the Hand-drawn-shape dataset (Robert, 2022), a collection of isolated hand-drawn basic shapes (rectangles, ellipses, triangles, and irregular blobs) on plain backgrounds. We recolour, scale, and distort these shapes on top of the basemaps with random line widths and opacities, then apply morphological operations and standard image augmentations (noise, blur, lens distortion, perspective warping, brightness/colour shifts, shadows) so that the resulting markings and artefacts resemble real pen strokes on scanned printouts. Samples are split geographically by country into 80/10/10 train/validation/test sets to evaluate spatial generalisation, and representative clean–annotated pairs are shown in Figure 1 (top row).

2.2 Detection architectures

The extraction of sketch-map annotations is formulated as an object-detection task, either on the annotated map alone or on paired clean–annotated inputs, where the latter corresponds to bi-temporal change detection by identifying markings present only in the annotated map. We employ a YOLOv9e object detector, a single-stage convolutional architecture consisting of a backbone for feature extraction, a neck for multi-scale feature aggregation, and a detection head for bounding-box prediction (Wang and Liao, 2024). We evaluate two input regimes: single-image detection from the annotated map alone, and dual-image detection from clean–annotated pairs (Fig. 2).

Single-image detection. The model receives only the annotated RGB map ($\text{RGB}_{\text{annotated}}$) and predicts bounding boxes for hand-drawn markings directly from this image.

Dual-image Siamese detection. The clean and annotated RGB images are processed independently by two weight-sharing backbones, producing corresponding multi-scale feature pyramids. At pyramid levels P3–P5 (strides 8, 16, and 32), element-wise feature differences are computed to emphasise changes and suppress shared structure. The resulting change-enhanced features are forwarded to a shared neck and detection head for bounding-box prediction. This differencing operation introduces no additional learnable parameters and maintains computational complexity comparable to the single-image baseline while enabling change reasoning at the feature level. By encoding each map separately before comparison, the model can first form a stable representation of the underlying basemap in each input and then focus explicitly on what differs between them. This is expected to improve robustness primarily to appearance-related variation, such as shadows, colour shifts, and scanning artefacts, but it does not explicitly correct geometric misalignment and still assumes approximate spatial alignment between the paired inputs.

During training, geometric and photometric augmentations are applied to the annotated image in the single-image setting, and synchronously to both images in the paired setting to preserve alignment. In both settings, additional degradations (e.g., blur, noise, shadows, and illumination changes) are applied only to the annotated image to simulate realistic scanning and photography artefacts (distinct from the distortions used during synthetic data generation).

2.3 Training and evaluation

We evaluate three model variants. First, a single-image model is trained only on the real hand-drawn data. Second, the same single-image model is pretrained on synthetic data and subsequently fine-tuned on real-world maps to assess the benefit of synthetic pretraining. Third, a dual-

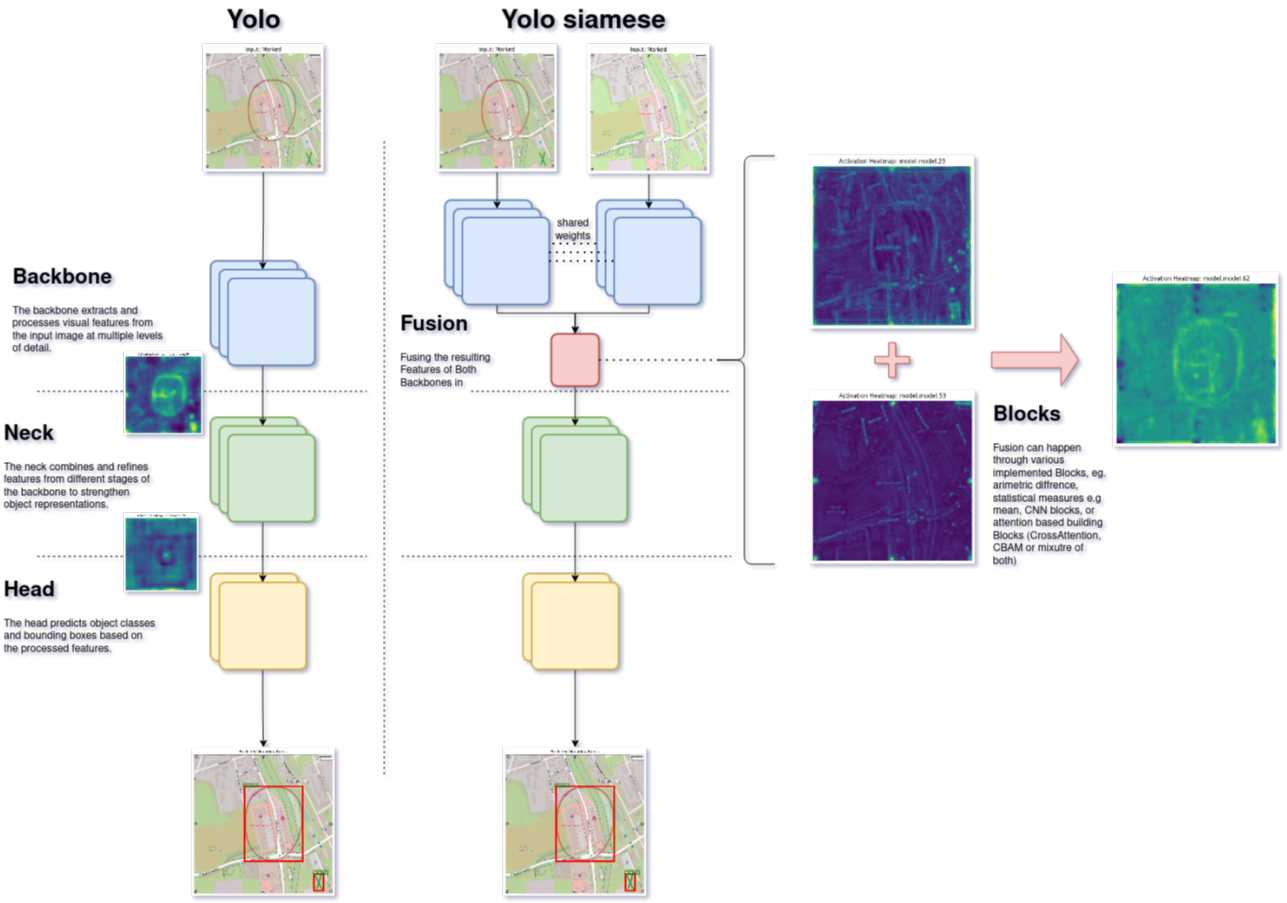

Figure 2. YOLOv9e single-image (left) and Siamese YOLOv9e dual-image (right) detectors.

image Siamese model is trained under the same synthetic-pretraining and real-world fine-tuning regime to assess the benefit of using paired clean–annotated inputs. For synthetic pretraining of the single-image models, only the synthetic annotated maps are used, whereas the Siamese model uses the full synthetic clean–annotated pairs. Real-world evaluation is conducted separately on the hand-drawn OSM and EWI datasets to assess performance across distinct basemap styles.

Training uses 1024×1024 inputs, batch size 7, 200 epochs, SGD with lr0 = 0.01 and cosine decay, momentum 0.937, weight decay 5×10^{-4}, and early stopping (patience 20). Performance is reported using Precision (p), Recall (r), mean Average Precision at IoU = 0.5 (mAP_{50}), and COCO-style mean Average Precision averaged over IoU thresholds 0.50–0.95 ($\mathrm{mAP}_{50\text{–}95}$), which emphasizes localization quality and penalizes loose bounding boxes.

Training was conducted on a single NVIDIA A100 GPU (50 GB VRAM).

3 Results

Across all three models – early fusion trained on real data (model i), early fusion with synthetic pretraining (model ii), and late fusion (model iii) – training and qualitative analyses indicate stable learning behaviour and effective localisation of hand-drawn annotations. Training curves (Fig. 3) show smooth convergence without evident overfitting. Activation maps (Fig. 4) suggest that the models focus on newly added markings while largely suppressing static background content. Representative failure cases are shown in Figure 5.

Table 1 summarizes detection performance for three models by basemap type: vector-rendered (OSM) and satellite imagery (EWI). The single-image model trained only on real data achieves mAP_{50} scores of 90.6% on EWI maps and 93.2% on OSM maps. Synthetic pretraining improves performance, with the single-image model achieving higher mAP_{50} on both EWI (90.6%⟶91.5%) and OSM (93.2%⟶97.3%). The Siamese dual-image model further improves performance on both datasets, with mAP_{50} increasing substantially on EWI maps (91.5⟶97.4) and more modestly on OSM maps (97.3%⟶98.1%), alongside a large precision increase on EWI from 91.2 to 96.9. On the synthetic (combined) split, both models trained with synthetic data achieve very high accuracy (mAP_{50} 97.2–98.6%), with only modest differences between models.

Table 1. Detection performance across datasets. Precision (p), Recall (r), mAP_{50}, and $mAP_{50–95}$.

Model	Synthetic (combined)				Hand-drawn EWI				Hand-drawn OSM			
	p	r	mAP_{50}	$mAP_{50–95}$	p	r	mAP_{50}	$mAP_{50–95}$	p	r	mAP_{50}	$mAP_{50–95}$
i. YOLOv9e (no synthetic pretrain)	–	–	–	–	94.6	87.8	90.6	73.0	95.5	95.5	93.2	80.4
ii. YOLOv9e (with synthetic pretrain)	99.1	95.7	97.2	95.7	91.2	86.5	91.5	76.2	98.0	96.2	97.3	81.7
iii. **Siamese-YOLOv9e** (with synthetic pretrain)	99.1	96.8	**98.6**	96.7	96.9	89.6	**97.4**	77.2	98.9	97.0	**98.1**	84.5

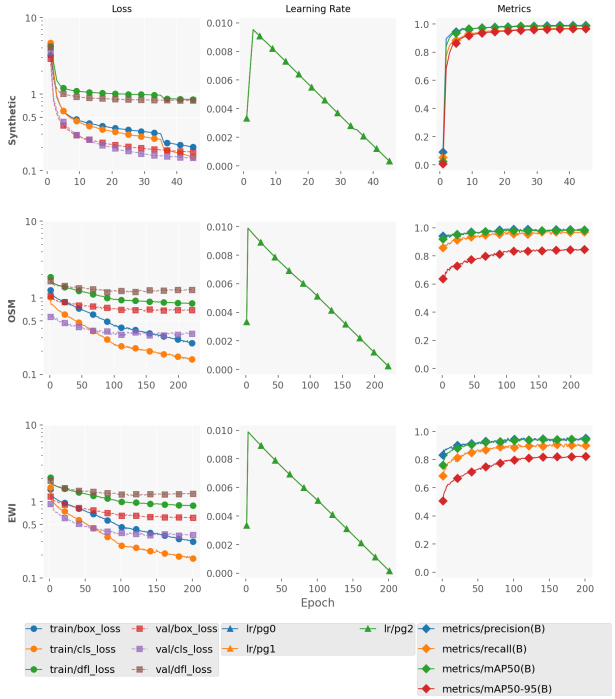

Figure 3. Training and validation metrics for the Siamese YOLOv9e model.

4 Discussion

Digitising participatory sketch maps remains challenging because the source material is inherently messy: hand-drawn annotations vary widely in style, thickness, and placement, while scanned or photographed maps introduce noise, distortions, illumination variations, and background clutter. Addressing this problem therefore requires methods that can tolerate substantial visual variability rather than relying on clean, well-structured inputs. Although all of our models achieve high accuracy in absolute terms—reaching roughly 90%–98% mAP_{50} depending on data conditions—even modest improvements in this task are meaningful because real participatory sketch maps can be noisier still than the datasets used here. Moreover, this application effectively demands near-perfect performance: in practice, moving from "reasonable" accuracy to the 95% + range represents a qualitative shift, where outputs become reliable enough for GIS use with minimal manual correction.

We hypothesised that using the clean basemap as an explicit reference and reasoning about differences between the clean and annotated maps would improve robustness under such conditions. The results support this expectation: dual-image Siamese detection improves performance on hand-drawn maps, with the largest gains observed on visually complex satellite imagery. The improvements are smaller on vector-rendered basemaps, where single-image detection already performs strongly, suggesting that the paired-input advantage is most pronounced when background texture and appearance variability make annotations harder to separate from the map content. Beyond this specific case study, the findings suggest that approaches that explicitly model change between paired inputs may be broadly beneficial for digitising noisy, user-generated spatial data, where annotations often resemble or overlap with background features.

Synthetic pretraining improves performance on real data, indicating that simulated annotations can provide useful supervision when manually labelled examples are limited. However, accuracy remains higher on synthetic than on real-world maps, suggesting a realism gap: despite extensive augmentation to approximate scanning and photography artefacts, synthetic samples still rely on cleanly generated annotations and controlled pairings, whereas real maps exhibit uneven illumination, paper texture, registration errors, and labelling inconsistencies that are difficult to reproduce with random transformations alone. This points to a clear next step: improving the synthetic data generation process—by better modelling acquisition artefacts and drawing behaviour—so that synthetic pretraining transfers more effectively to real participatory sketch maps.

From a workflow perspective, the detector should be understood as a decision-support tool for semi-automatic digitisation. Manual correction remains necessary whenever false positives or missed annotations affect the faithful representation of participant input, especially for thin, overlapping, or poorly captured markings. In practice, false positives increase review time by adding non-existent elements, while false negatives are more critical where complete capture of community-provided information is required.

Several limitations also point to directions for future work. The manually annotated dataset is relatively small and cannot capture the full diversity of drawing styles, marker types, and capture conditions encountered in practice. Highly overlapping or very thin strokes remain difficult to localise accurately, indicating failure modes that are particularly relevant for real participatory mapping workflows. In terms of data provenance, the hand-drawn

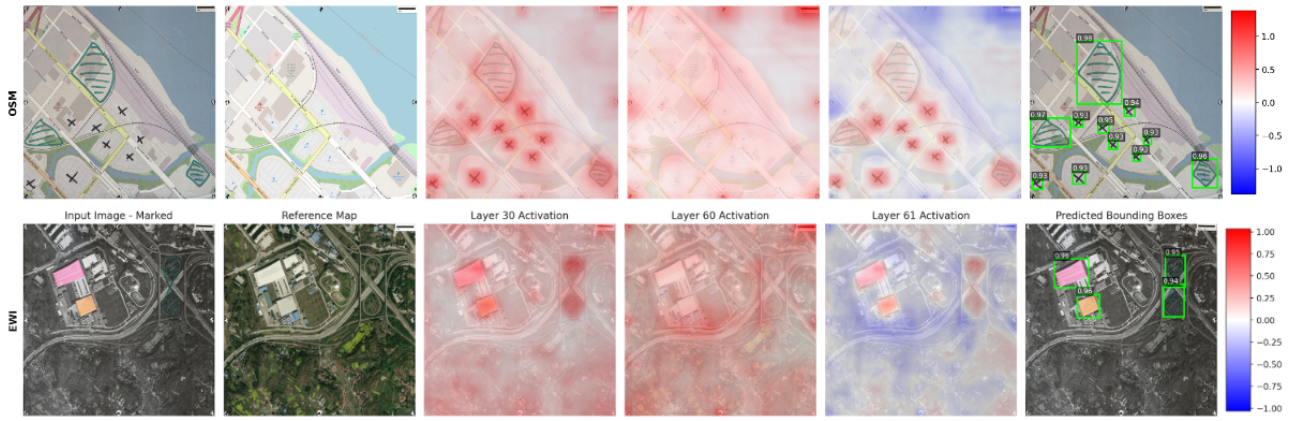

Figure 4. Class activations maps for the fusion of the P5-level features in layer 30 (marked), 60 (clean) and 61 (Fused).

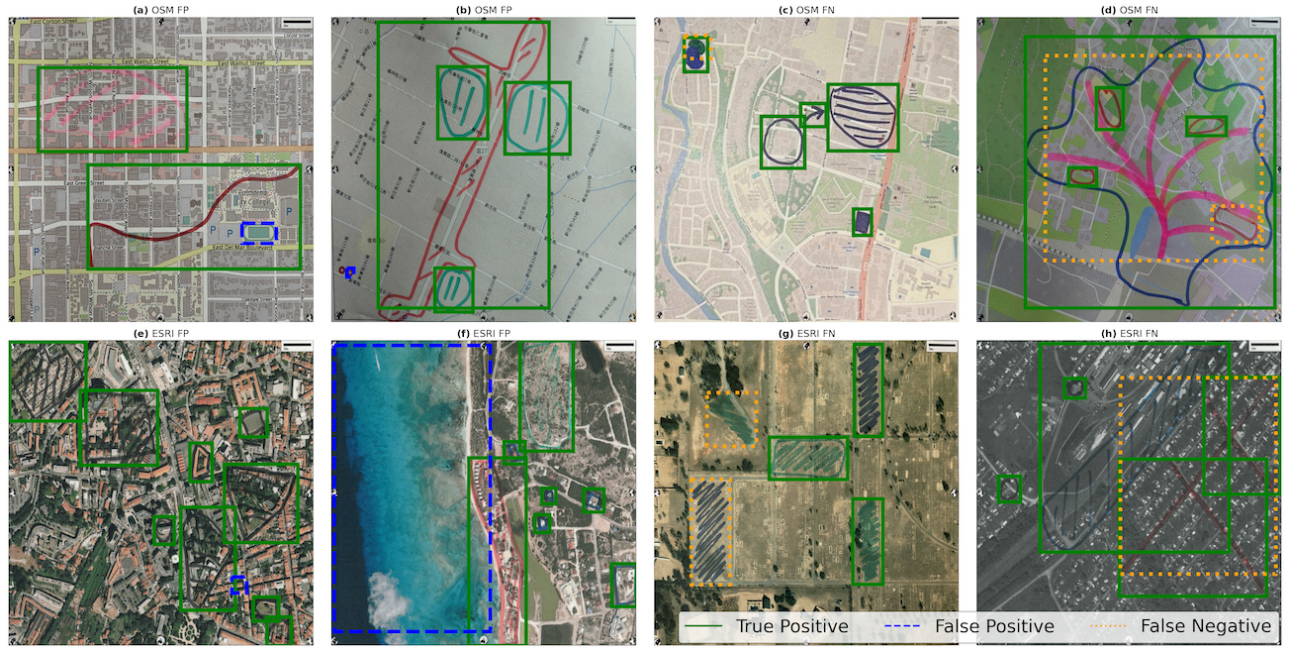

Figure 5. Common Detection Errors on OSM (a–d) and EWI (e–h): Examples of False Positives (FP) and False Negatives (FN).

dataset was created in internal data-curation rounds rather than collected from live participatory mapping campaigns, even though it was designed to reflect variation in basemap context and annotation style. Methodologically, the benefits of dual-image change reasoning should also be evaluated beyond the YOLO family; in particular, testing transformer-based detectors (e.g., DETR-style models such as RT-DETR (Lv et al., 2024)) would help determine whether the observed paired-input benefit generalises across architectural paradigms. Finally, exploring smaller and more efficient backbones could reduce computational demands and support deployment in resource-constrained environments.

Declaration of Generative AI in writing

The authors used generative AI tools for language editing only. No AI tools were used to generate scientific content, data, analysis, or conclusions.

Data and Software Availability (DASA)

The hand-drawn annotated OSM-based data are available at https://doi.org/10.5281/zenodo.19367883. EWI-based data cannot be redistributed due to licensing restrictions. The synthetic dataset is not publicly released, as it is derived from a combination of OSM and EWI basemaps.

The source code, training scripts, configuration files, and pretrained model weights (SMT-OSM and SMT-ESRI) are publicly available at https://github.com/GIScience/ultralytics_siamese_smt. Documentation and

usage instructions are provided to enable inspection and validation of the reported results.

Acknowledgements

This work was supported by the Klaus Tschira Stiftung (KTS), the German Foreign Office, and the German Red Cross (GRC). Computational and storage support was provided by Heidelberg University's Computing Center through the SDS@hd hot-data storage service, by the state of Baden-Württemberg through bwHPC and the bwForCluster Helix, and by the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

References

Boschmann, E. E. and Cubbon, E.: Sketch maps and qualitative GIS: Using cartographies of individual spatial narratives in geographic research, The Professional Geographer, 66, 236–248, 2014.

Chipofya, M. C., Sahib, J., and Schwering, A.: SmartSkeMa: Scalable Documentation for Community and Customary Land Tenure, Land, 10, 662, https://doi.org/10.3390/land10070662, 2021.

CVAT.ai Corporation: Computer Vision Annotation Tool (CVAT), https://doi.org/10.5281/zenodo.4009388, https://github.com/opencv/cvat, 2022.

Denwood, T., Huck, J. J., and Lindley, S.: Paper2GIS: improving accessibility without limiting analytical potential in Participatory Mapping, Journal of Geographical Systems, 25, 37–57, 2023.

Guo, H., Sun, C., Zhang, J., Zhang, W., and Zhang, N.: Mmyfnet: Multi-modality yolo fusion network for object detection in remote sensing images, Remote Sensing, 16, 4451, 2024.

Klonner, C., Hartmann, M., Dischl, R., Djami, L., Anderson, L., Raifer, M., Lima-Silva, F., Castro Degrossi, L., Zipf, A., and Porto de Albuquerque, J.: The sketch map tool facilitates the assessment of OpenStreetMap data for participatory mapping, ISPRS International Journal of Geo-Information, 10, 130, 2021.

Laituri, M., Luizza, M. W., Hoover, J. D., and Allegretti, A. M.: Questioning the practice of participation: Critical reflections on participatory mapping as a research tool, Applied Geography, 152, 102 900, 2023.

Lindner, C., Degbelo, A., Vassányi, G., Kundert, K., and Schwering, A.: The SmartLandMaps Approach for Participatory Land Rights Mapping, Land, 12, 2043, 2023.

Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., and Liu, Y.: RTDETRv2: All-in-One Detection Transformer Beats YOLO and DINO, 2024.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection, CVPR, 2016.

Robert, F.: Hand-drawn Shapes (HDS) Dataset, https://github.com/frobertpixto/hand-drawn-shapes-dataset/tree/main, accessed: 2025-05-12, 2022.

Wang, C.-Y. and Liao, H.-Y. M.: YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information, arXiv preprint arXiv:2402.13616, 2024.

Zhang, Y., Pang, J., Li, B., and Luo, J.: Siamese YOLO V5 with Structure coefficient for object-level change detection, in: 2024 4th International Conference on Electronic Information Engineering and Computer (EIECT), pp. 232–242, IEEE, 2024.