ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-3-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission III, 4–11 July 2026, Toronto, Canada

Calibrated U-Net with HELIX-Based Label Enrichment for Ageing-Aware Spatio-Temporal Urban Change Detection

Sarah Hauser[1,2,3], Stephanie Dachsberger[2], Andreas Schmitt[2,3], Stefan Hinz[1]
[1] Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), Germany – sarah.hauser@kit.edu
[2] Geoinformatics Department, Munich University of Applied Sciences (HM), Germany
[3] Institute for Applications of Machine Learning and Intelligent Systems (IAMLIS), Germany

Keywords: Semantic segmentation, legacy supervision, label quality, probabilistic calibration, uncertainty-aware mapping, urban change detection

Abstract

Urbanisation and land-use change increase the demand for temporally consistent urban maps from high-resolution Earth observation imagery. A key obstacle is *label ageing*: benchmark annotations are often years older than current true orthophotos (TOP), causing semantic and geometric mismatches (e.g., demolished/new buildings, shifted vegetation boundaries) that degrade supervised learning, calibration, and transfer. This paper presents a probabilistic, quality-aware segmentation framework based on a compact U-Net. Legacy annotations are converted into edge-adaptive soft labels to encode boundary uncertainty. A HELIX-derived per-pixel supervision quality score Q is computed and integrated as a weight in a Q-weighted Kullback–Leibler objective with an agreement-focal component, reducing the influence of unreliable or outdated regions. Global temperature scaling is then applied to obtain calibrated per-class probability fields with comparable confidence magnitudes. Experiments on ISPRS Potsdam and Vaihingen combined with recent (2024) TOPs evaluate temporal transfer (archival supervision vs. updated imagery of the same area) and spatial transfer (cross-city application). Finally, calibrated probability fields are used to derive probabilistic semantic transitions and temporal reliability scores, supporting uncertainty-aware mapping of urban change such as construction, sealing, and vegetation loss.

1. Introduction

Urbanisation continues to reshape land cover and settlement structure worldwide. The share of the global population living in urban areas has risen from roughly 34% in 1960 to over 57% today and is still growing at about 1.7% per year (World Bank, 2024, United Nations, 2022). This expansion increases pressure on ecosystems, vegetated areas, and permeable surfaces, creating demand for spatially explicit and temporally current information on processes such as construction, sealing, greening, and canopy loss (Seto et al., 2012). Earth Observation (EO), combined with recent advances in artificial intelligence and Deep Learning (DL), enables dense, pixel-based mapping of urban land cover from high-resolution remote sensing data (Li et al., 2024). Benchmark datasets such as ISPRS Vaihingen and Potsdam (ISPRS, 2013) have therefore been central to the development and evaluation of semantic segmentation models. A central limitation of these benchmarks is temporal obsolescence (Hauser et al., 2025). The imagery and annotations in ISPRS-style datasets were acquired years ago, while current true orthophotos (TOPs) capture cities that have since changed. This temporal mismatch between legacy labels and present-day imagery is referred to as *label ageing*. As buildings are demolished or constructed, natural surfaces are sealed or restored, and vegetation structure shifts, the “ground truth” stops reflecting reality. Figure 1 contrasts legacy ISPRS Potsdam labels (ISPRS, 2013) with a recent (2024) TOP of the same area, illustrating missing and newly appearing structures as well as shifted vegetation boundaries. Label ageing injects contradictory supervision during training (especially at boundaries) and degrades probability calibration by encouraging confidence in outdated structures (Guo et al., 2017). Consequently, conventional metrics (e.g., overall accuracy) can misrepresent practical

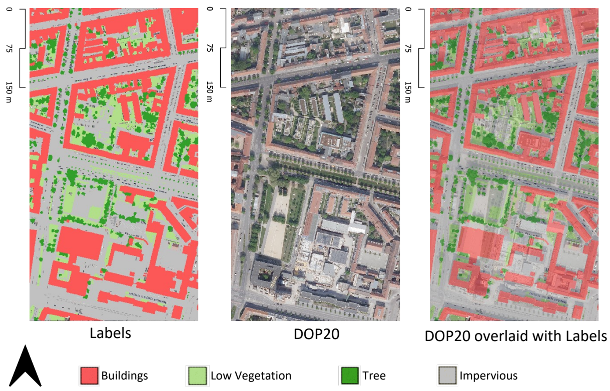


Figure 1. Comparison of reference labels and current aerial imagery for a sub-area of ISPRS Potsdam: (a) original benchmark labels, (b) recent 2024 TOP, (c) overlay highlighting demolished / new buildings and shifted vegetation boundaries.

performance and transferability (Northcutt et al., 2021). These challenges motivate three design choices in the present study: (i) edge-adaptive soft labels to encode boundary ambiguity instead of enforcing artificially crisp targets, (ii) HELIX-derived per-pixel quality scores to down-weight unreliable or ambiguous supervision during optimisation, and (iii) post-hoc calibration to obtain probability fields with interpretable and comparable confidence scales across tiles, cities, and acquisition dates. Accordingly, a compact U-Net is trained to produce calibrated, uncertainty-aware urban surface probability maps from legacy benchmark data, and the resulting probability fields are analysed under spatial and temporal transfer to support consistent urban change interpretation.

1.1 State of the Art

Urban land-cover mapping in EO has evolved from traditional classifiers such as random forests (RF) and support vector machines (SVM) (Amini et al., 2022, Lemenkova, 2024) to deep convolutional encoder–decoder architectures. U-Net-style models (Ronneberger et al., 2015, Zhang et al., 2023, Goswami et al., 2025) are now widely used for dense, per-pixel labelling in high-resolution aerial imagery, including the ISPRS Vaihingen and Potsdam benchmarks (ISPRS, 2013). Such architectures capture multi-scale spatial context while preserving fine spatial detail via skip connections (Majidizadeh et al., 2023), and can delineate classes like buildings, sealed surfaces, and vegetation with high nominal accuracy. However, these models typically assume deterministic supervision: each pixel is assigned a single "true" class, and the network outputs one winning class at inference (Zhang et al., 2023). This assumption fails in practice. Class boundaries are ambiguous, pixels can be mixed, and annotations can be geometrically misaligned. Penalising every deviation from a crisp boundary encourages overconfident, overly sharp transitions and makes uncertainty at class interfaces difficult to interpret. Soft-label strategies have been proposed to mitigate this. The Enhanced Soft Label (ESL) framework assigns adaptive probabilistic targets to pixels and explicitly preserves uncertainty in high-entropy regions, especially near boundaries (Ma et al., 2023). ESL improves structural coherence and reduces artefacts along fine details. Yet ESL largely treats all pixels symmetrically. It does not encode whether a pixel's label is spatially clean, consistent with its neighbourhood, semantically well separated from alternatives, or even temporally valid. The HELIX framework (Hauser et al., 2025) addresses this by modelling supervision not only as a soft class distribution but also through a per-pixel quality score Q that reflects spatial purity, neighbourhood agreement, and semantic separability. HELIX-inspired formulations have also been applied beyond urban settings, for example in bark beetle disturbance mapping, forest structure assessment, and glacier facies characterisation (Hauser, 2025). In most cases, however, HELIX has been used as a preprocessing and diagnostics tool: Q is computed to describe label trustworthiness, but it is not directly integrated into deep network optimisation, sampling, post-hoc calibration, or explicit spatio-temporal change mapping. Other relevant lines of work focus on probabilistic reliability and transfer. Post-hoc calibration, particularly temperature scaling (Guo et al., 2017) and spatially adaptive variants such as local temperature scaling (Ding et al., 2021), rescales logits so that predicted probabilities better match empirical correctness. This improves interpretability of per-class probability maps. Domain adaptation (DA) aims to transfer models across cities, sensors, seasons, or acquisition conditions (Li et al., 2024) by enforcing domain-invariant representations. However, both calibrated inference and DA typically assume that the source-domain labels are correct and up to date. That assumption breaks down under label ageing, which is not just an input-domain shift but a *reference-domain shift*: the ground truth itself has drifted. A modern urban mapping framework is expected to (i) treat legacy benchmark labels as uncertain and potentially outdated, rather than assuming they are flawless; (ii) assign an explicit, spatially aware, per-pixel quality estimate to that supervision; (iii) incorporate that quality directly into the optimisation process so that unreliable pixels contribute less to training; (iv) produce calibrated per-class probability fields, rather than only hard segmentations, such that predictive confidence is interpretable; and (v) exploit those calibrated probability fields to analyse urban change consistently across space and time. The study presented here operationalises these requirements by combining HELIX-style per-pixel label quality with a quality-weighted training objective for a compact U-Net; by applying global temperature scaling to obtain well-calibrated per-class probability maps (Guo et al., 2017); and by interpreting these calibrated probability fields as spatio-temporal evidence of urban transformation (e.g., construction, sealing and de-sealing, vegetation loss, canopy removal), together with associated confidence.

1.2 Materials and Data

All experiments use publicly available remote sensing data. Legacy benchmark annotations from the ISPRS 2D Semantic Labeling datasets for Potsdam and Vaihingen (ISPRS, 2013) are combined with recent (2024) digital orthophotos to analyse both spatial transfer (cross-city generalisation) and temporal transfer (label ageing within the same city). Potsdam, southwest of Berlin, covers 188 km^2 with about 188,000 inhabitants (Landeshauptstadt Potsdam, 2024). The ISPRS Potsdam benchmark (BSF Swissphoto; released via the ISPRS challenge (ISPRS, 2013)) consists of 38 tiles of 6000×6000 px at 5 cm ground sampling distance (GSD). Each tile includes a true orthophoto (TOP) and a co-registered digital surface model (DSM) on a common UTM WGS84 grid. Imagery is available in multiple spectral configurations (RGB, IRRG, RGBIR), allowing the extraction of $\{\mathrm{NIR}, R, G\}$. A subset of tiles is annotated with six semantic classes: *impervious surface*, *building*, *low vegetation*, *tree*, *car*, and *clutter/background*. To assess temporal robustness, these legacy annotations are contrasted with a recent Potsdam TOP (Landesvermessung und Geobasisinformation Brandenburg, LGB 2024), acquired on 30 April 2024 at 20 cm resolution. This pairing (legacy ISPRS labels vs. current imagery) exposes temporal inconsistencies such as demolished or newly constructed buildings, resurfaced or re-sealed areas, and shifted vegetation boundaries, and is therefore used to study label ageing and temporal transfer. Vaihingen an der Enz, roughly 25 km northwest of Stuttgart, has about 29,000 inhabitants across $\sim$73 km^2 (Stadt Vaihingen an der Enz, 2024). Unlike Potsdam's denser perimeter-block structure, Vaihingen is dominated by low-rise residential areas with detached and small multi-storey buildings, gardens, and tree cover. The ISPRS Vaihingen dataset (ISPRS, 2013) comprises 33 tiles of average size 2494×2064 px at $\sim$9 cm GSD, each providing a TOP (near-infrared, red, green; 8-bit TIFF) and a DSM (32-bit float). The same six semantic classes as in Potsdam are annotated for a subset of tiles, ensuring a consistent ontology across both cities. For spatial transfer and joint spatio-temporal transfer, a recent TOP of Vaihingen from 9 July 2024 at 20 cm resolution (Landesamt für Geoinformation und Landentwicklung Baden-Württemberg, LGL-BW 2024) is additionally used. Comparing legacy ISPRS labels to this 2024 acquisition enables simultaneous analysis of (i) domain shift between two contrasting urban morphologies (dense block structures vs. detached housing) and (ii) temporal mismatch between legacy annotations and the present-day built environment. In all experiments, the *car* class is merged into *impervious surface* to harmonise class definitions and reduce sparsity, resulting in a four-class ontology: *impervious*, *building*, *low vegetation*, *tree*. These datasets jointly support two central objectives: assessing how legacy but ageing labels can still be exploited when their spatial and temporal reliability is modelled explicitly, and evaluating whether the resulting calibrated probability fields can be transferred across cities and acquisition dates to detect, describe, and qualify urban change.

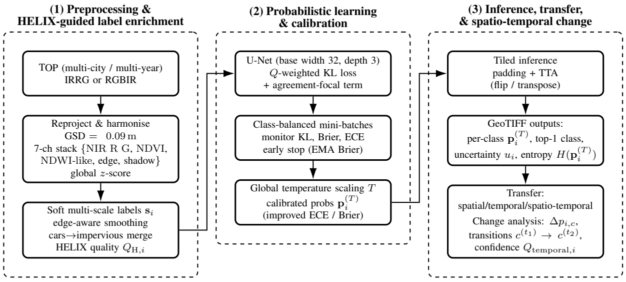


Figure 2. Overview of the framework. (1) Preprocessing and HELIX-guided label enrichment. (2) Quality-weighted probabilistic learning and global calibration. (3) Inference, spatial/temporal transfer, and spatio-temporal change analysis.

2. Methodology

This section describes a three-stage pipeline for calibrated, quality-aware urban surface mapping from high-resolution EO imagery. The pipeline combines (i) HELIX-inspired label enrichment and per-pixel supervision quality estimation (Hauser et al., 2025, Hauser, 2025), (ii) quality-weighted probabilistic learning with global temperature scaling for calibration, and (iii) spatial, temporal, and spatio-temporal transfer for semantic change analysis. The full workflow is summarised in Fig. 2.

2.1 Preprocessing and HELIX-guided label enrichment

TOPs from different acquisition years and cities are first harmonised to a common spatial reference and resampled to a uniform ground sampling distance ($\mathrm{GSD} = 0.09\,\mathrm{m}$). Each scene is then converted into a consistent 7-channel input stack. From the available IRRG (3-band) or RGBIR (4-band) products, the channels $\{\mathrm{NIR}, R, G\}$ are retained, and several derived features are added: NDVI, an NDWI-like index, an edge magnitude from image gradients, and a shadow heuristic from local brightness differences. For each source scene, global per-channel mean and standard deviation are estimated on valid pixels and used to z-score all channels. This yields a compact, normalised feature stack $\{\mathrm{NIR\ R\ G}, \mathrm{NDVI}, \mathrm{NDWI\text{-}like}, \mathrm{edge}, \mathrm{shadow}\}$ that is applied consistently across domains and acquisition dates (ISPRS, 2013). Reference annotations distinguish *impervious surface*, *building*, *low vegetation*, *tree*, *car*, and *clutter/background* (ISPRS, 2013). For a stable and transferable ontology, the *car* class is merged into *impervious surface*, and *clutter/background* is ignored. This results in a four-class label set: *impervious*, *building*, *low vegetation*, *tree*. Hard labels are then converted into soft, edge-aware, multi-scale supervision. For each pixel i, a soft target vector $\mathbf{s}_i \in [0,1]^C$ with $\sum_c s_{i,c} = 1$ is created by blending Gaussian smoothings at multiple spatial scales, guided by an edge map. Small kernels preserve sharp boundaries; larger kernels encode neighbourhood context; edge guidance prevents artificial class bleeding across strong physical edges. This produces spatially coherent targets and explicitly encodes boundary uncertainty. Each pixel also receives a HELIX-style supervision quality score $Q_{\mathrm{H},i}$, which quantifies how trustworthy the supervision is at that location (Hauser et al., 2025, Hauser, 2025). The score $Q_{\mathrm{H},i}$ is defined in Eq. (1) as a weighted combination of spatial purity, neighbourhood agreement, and semantic separability:

$$Q_{\mathrm{H},i} = w_s\, C_{\mathrm{spatial},i} + w_n\, C_{\mathrm{neigh},i} + w_t\, C_{\mathrm{sem},i}, \qquad (1)$$

where $C_{\mathrm{spatial},i}$ penalises high-entropy ambiguity in $\mathbf{s}_i$, $C_{\mathrm{neigh},i}$ measures local consensus via a spatial smoothing of the dominant class probability, and $C_{\mathrm{sem},i}$ is the margin between the top two probabilities in $\mathbf{s}_i$. High $Q_{\mathrm{H},i}$ therefore indicates that the supervision at pixel i is internally consistent, spatially coherent, and semantically unambiguous. Importantly, $Q_{\mathrm{H},i}$ measures the internal reliability of the legacy annotation at its acquisition time, not agreement with the current state of the city. Systematic confusions (e.g., *building* vs. *impervious*) are handled explicitly. In such zones, the two confused classes are gently mixed and the corresponding $Q_{\mathrm{H},i}$ contribution is slightly down-weighted. This encodes that these cases are expected to be ambiguous (e.g., flat roofs vs. sealed courtyards) and prevents them from dominating learning.

2.2 Probabilistic learning and calibration

The supervision $(\mathbf{s}_i, Q_{\mathrm{H},i})$ is used to train a compact U-Net with base width 32 and depth 3. The U-Net follows a standard encoder–decoder structure with skip connections and a 1×1 prediction head, producing per-pixel logits $\mathbf{z}_i$ and probabilities $\mathbf{p}_i = \mathrm{softmax}(\mathbf{z}_i)$. Training minimises a HELIX-quality-weighted Kullback–Leibler divergence between the network output $\mathbf{p}_i$ and the soft supervision $\mathbf{s}_i$. The loss $\mathcal{L}_{\mathrm{KL}}$ is given in Eq. (2):

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{|\Omega|} \sum_{i\in\Omega} Q_{\mathrm{H},i}\, \mathrm{KL}\big(\mathbf{s}_i \,\|\, \mathbf{p}_i\big), \qquad (2)$$

where Ω is the set of sampled training pixels. Here $\mathrm{KL}(\mathbf{s}_i \,\|\, \mathbf{p}_i)$ is computed per pixel as $\sum_c s_{i,c} \log \frac{s_{i,c}}{p_{i,c}}$, i.e. the divergence between the soft supervision $\mathbf{s}_i$ and the network prediction $\mathbf{p}_i$ for pixel i. The factor $Q_{\mathrm{H},i}$ weights this divergence by the estimated supervision quality from Eq. (1), such that spatially coherent and semantically stable pixels contribute more strongly to optimisation than ambiguous or temporally outdated pixels. Mini-batches are constructed from approximately class-balanced patch centres to ensure that all semantic classes are observed during training. Within each batch, each pixel's contribution to the loss is weighted by its HELIX-style supervision quality score Q_{H}, such that spatially unreliable or temporally outdated pixels have reduced influence on optimisation. This produces an effect similar to explicit high-Q sampling, but without discarding uncertain regions entirely. To further emphasise pixels where the model still disagrees with high-quality supervision, an additional agreement-focal weighting is applied. The focal weight w_i^{foc} is defined in Eq. (3):

$$w_i^{\mathrm{foc}} = \big(1 - \langle \mathbf{p}_i, \mathbf{s}_i \rangle\big)^{\gamma}, \qquad (3)$$

where $\langle \mathbf{p}_i, \mathbf{s}_i \rangle$ is the per-pixel agreement between prediction and soft supervision, and γ is a focusing exponent. The final training loss $\mathcal{L}_{\text{U-Net}}$ augments Eq. (2) with this focal term, as shown in Eq. (4):

$$\mathcal{L}_{\text{U-Net}} = \mathcal{L}_{\mathrm{KL}} + \lambda_f\, \frac{1}{|\Omega|} \sum_{i\in\Omega} Q_{\mathrm{H},i}\, w_i^{\mathrm{foc}}\, \mathrm{KL}\big(\mathbf{s}_i \,\|\, \mathbf{p}_i\big), \qquad (4)$$

with $\lambda_f = \alpha/(1-\alpha)$ controlling the focal contribution. Optimisation uses AdamW with a ReduceLROnPlateau scheduler. During training, KL divergence, Brier score, Expected Calibration Error (ECE), soft and hard agreement with supervision, entropy, and per-class RMSE are monitored. Early stopping is triggered by an EMA-smoothed validation Brier score; checkpoints with best Brier and best ECE are retained. After optimisation, global temperature scaling is applied on a held-out validation split. In this context, “temperature” is a single scalar

$T > 0$ that rescales the network's logits before the softmax (cf. Eq. 5). Larger T yields flatter (less confident) probability distributions; smaller T yields sharper (more confident) ones. The procedure is post hoc and model-agnostic: it leaves all network weights and the top-1 class unchanged and only adjusts confidence magnitudes so that predicted probabilities better match the soft labels on the validation split. The calibrated probabilities $\mathbf{p}_i^{(T)}$ are obtained by rescaling the logits $\mathbf{z}_i$ according to Eq. (5):

$$\mathbf{p}_i^{(T)} = \mathrm{softmax}\big(\mathbf{z}_i/T\big). \quad (5)$$

Temperature scaling improves reliability indicators such as ECE and Brier score (Guo et al., 2017), and produces confidence values that are comparable across tiles, cities, and acquisition dates.

2.3 Inference, transfer, and spatio-temporal change analysis

Inference is performed in a tiled, memory-safe forward pass. Each tile is normalised using the same global per-channel statistics as in training, then passed through the network with padding. Test-time augmentations (horizontal/vertical flips and transpose/rotations) are applied and the resulting probability maps are averaged. For each tile, four georeferenced rasters are written directly to disk: (i) per-class calibrated probability maps, (ii) the most likely class per pixel, (iii) per-pixel uncertainty, and (iv) per-pixel entropy. For each pixel i, the calibrated per-class probability vector is denoted $\mathbf{p}_i^{(T)}$, as shown in Eq. (6):

$$\mathbf{p}_i^{(T)} = \big(p_{i,1}^{(T)}, \ldots, p_{i,C}^{(T)}\big), \quad (6)$$

with classes $\{$*impervious*, *building*, *low vegetation*, *tree*$\}$. A simple uncertainty score u_i is defined in Eq. (7) as the complement of the maximum calibrated confidence:

$$u_i = 1 - \max_c p_{i,c}^{(T)}. \quad (7)$$

The Shannon entropy of $\mathbf{p}_i^{(T)}$ is given in Eq. (8):

$$H\big(\mathbf{p}_i^{(T)}\big) = -\sum_{c=1}^{C} p_{i,c}^{(T)} \log p_{i,c}^{(T)}. \quad (8)$$

Because the outputs are globally calibrated (see Eq. (5)), $\mathbf{p}_i^{(T)}$, u_i, and $H(\mathbf{p}_i^{(T)})$ are directly comparable between different areas of interest and different acquisition dates.

2.4 Spatial, temporal, and spatio-temporal transfer

The calibrated network (or an ensemble of networks trained on different tiles) is deployed in three regimes: (i) spatial transfer, where a model trained on Area A (e.g., Potsdam, legacy TOP) is applied to Area B (e.g., Vaihingen, recent TOP); (ii) temporal transfer, where a model trained on time t_1 (e.g., legacy Potsdam) is applied to a later acquisition of the *same* area at t_2 (e.g., Potsdam 2024); and (iii) joint spatio-temporal transfer, where both city and acquisition year differ between training and inference. Because preprocessing is standardised (same 7-channel stack, same cars→impervious fusion, same global normalisation), the calibrated probability vectors $\mathbf{p}_i^{(T)}$ in Eq. (6) are directly comparable across these regimes.

2.5 Semantic change and urban dynamics

Semantic drift between two calibrated maps (e.g., two acquisition times or domains) is quantified from per-class probability deltas. For pixel i and class c,

$$\Delta p_{i,c} \;=\; p_{i,c}^{(T_2)} - p_{i,c}^{(T_1)}, \quad (9)$$

where $p_{i,c}^{(T_k)}$ are calibrated probabilities at t_k. Large positive $\Delta p_{i,building}$ together with negative $\Delta p_{i,low\ vegetation}$ indicates vegetation loss followed by construction or sealing. In parallel to this continuous signal, discrete "transition classes" are formed by the most likely labels at the two times: $c_i^{(t_1)} = \arg\max_c p_{i,c}^{(T_1)}$ and $c_i^{(t_2)} = \arg\max_c p_{i,c}^{(T_2)}$. The ordered pair $(c_i^{(t_1)} \to c_i^{(t_2)})$ yields one of $C \times C$ transitions (for C=4: 16 classes capturing both persistence and change). Beyond hard argmax transitions, a soft transition matrix aggregates probabilistic mass from class i at t_1 to class j at t_2:

$$S_{ij} \;=\; \frac{1}{Z}\sum_{x\in\Omega} p_{x,i}^{(T_1)}\, p_{x,j}^{(T_2)}, \qquad Z \;=\; \sum_{x\in\Omega} 1, \quad (10)$$

normalised so that $\sum_{i,j} S_{ij} = 1$. A reliability-weighted variant uses a temporal weight $Q_{\mathrm{temporal},x}$:

$$\begin{aligned} S_{ij}^{Q} &= \frac{1}{Z_Q}\sum_{x\in\Omega} Q_{\mathrm{temporal},x}\, p_{x,i}^{(T_1)}\, p_{x,j}^{(T_2)}, \\ Z_Q &= \sum_{x\in\Omega} Q_{\mathrm{temporal},x}. \end{aligned} \quad (11)$$

where Ω restricts to pixels whose t_1 labels lie in the four-class ontology (with *cars* merged into *impervious*). The temporal reliability indicator combines the HELIX label quality at t_1 with calibrated confidence at t_2:

$$Q_{\mathrm{temporal},i} \;=\; Q_{\mathrm{H},i}^{(t_1)} \cdot \max_c p_{i,c}^{(T_2)}. \quad (12)$$

High Q_{temporal} emphasises pixels with robust supervision at t_1 and confident predictions at t_2. Low-confidence regions (e.g., high entropy or uncertainty) can optionally be masked to prioritise reliable evidence. This framework yields calibrated, uncertainty-aware transition summaries that localise change (where) and characterise process type (what) together with an explicit confidence signal.

3. Results

In-domain behaviour and intra-city generalisation are evaluated in Potsdam using legacy ISPRS supervision (Sec. 3.1), and the calibrated network is then transferred without retraining across space to a different city (Sec. 3.2) and across time to a more recent TOP of the same area (Sec. 3.3). Finally, calibrated per-class probability fields from different epochs are differenced to derive spatio-temporal semantic change maps, quantifying processes such as sealing, construction, greening, and canopy removal together with associated confidence (Sec. 3.3).

3.1 In-domain segmentation and intra-city generalisation

In-domain behaviour was assessed on Potsdam (t_1) using legacy supervision from the ISPRS 2D Semantic Labelling benchmark (ISPRS, 2013). As outlined in Sec. 2, legacy masks were converted into edge-adaptive, multi-scale soft labels;

(a) Legacy orthophoto (b) Legacy semantic labels

Figure 3. Potsdam, t_1. (a): legacy TOP and (b) ISPRS labels (*building*, *impervious*, *low vegetation*, *tree*).

Setting	KL↓	Brier↓	ECE↓	Hard acc.↑
Potsdam val. (pre-cal)	0.290	0.163	0.056	0.886
Potsdam val. (post-cal)	**0.272**	**0.155**	**0.010**	**0.886**
Unseen Potsdam tile (post-cal, fixed T)	0.668	0.352	0.065	0.703

Table 1. In-domain validation (Potsdam, t_1) and intra-city hold-out (unseen Potsdam tile, $\approx 0.438\,\mathrm{km}^2$). Post-calibration uses a single global temperature fitted on the validation split (argmax unchanged). ECE is computed over 15 bins on the maximum (calibrated) class probability against soft labels.

per-pixel HELIX quality Q_{H} encoded spatial purity, neighbourhood agreement, and semantic separability; and a Q-weighted KL objective with an agreement-focal component was used for optimisation. The training pipeline further employed confusion-aware handling in *building–impervious* near-tie regions, quality-stratified sampling, and synchronised geometric/photometric augmentations, with padding-safe tiled inference and test-time augmentation. A single global temperature was fitted on a held-out split to calibrate probabilities while leaving argmax labels unchanged. The compact U-Net (base 32, depth 3) was trained for 10 epochs on seven-channel inputs $\{\mathrm{NIR\ R\ G}, \mathrm{NDVI}, \mathrm{NDWI\text{-}like}, \mathrm{Sobel}, \mathrm{Shadow}\}$ with patch size 160, batch size 8, a 20% class-balanced validation split, early stopping on EMA(Brier), and ECE computed with 15 bins. Key metrics for the in-domain validation split (pre-/post-calibration) and for a geographically unseen Potsdam tile are summarised in Tab. 1; per-class RMSEs are given in Tab. 3, and additional summary statistics (soft accuracy, entropy, confidence) are listed in Tab. 2. For clarity, *soft accuracy* reports agreement with the soft supervision (the expected probability, under the soft labels, of the predicted class), whereas *hard accuracy* reports argmax agreement against the hard target induced by the soft labels. All accuracies were computed against contemporaneous legacy ISPRS labels; although temporally matched to the TOP, these annotations may exhibit geometric imprecision and boundary coarseness (cf. Fig. 3), which can depress apparent accuracy despite visually coherent and well-calibrated probability fields. In Fig. 4, the probability composite (b) shows saturated red over roofs and sealed courtyards, green over tree crowns, and blue over lawns; lower saturation and mixed hues concentrate along kerbs, roof edges, and under canopy, reflecting boundary uncertainty captured by the soft-label design. Local discrepancies between (b) and the label overlays (a) are most evident at polygon edges and narrow features, consistent with geometric coarseness of legacy masks rather than systematic model failure. On the validation split, calibration yields very low ECE (≈ 0.01) while accuracies remain stable; on the geographically unseen tile, a degradation consistent with a geographic hold-out is observed (hard accuracy ≈ 0.70 and higher KL/Brier/ECE), with errors concentrated

Setting	Soft acc.↑	Entropy↓	Confidence↑
Potsdam val. (pre-cal)	0.879	0.449	0.824
Potsdam val. (post-cal)	**0.880**	**0.322**	**0.873**
Unseen Potsdam tile (post-cal, fixed T)	0.691	0.592	0.755

Table 2. Additional metrics to Tab. 1. Soft accuracy reports agreement with the soft supervision; entropy and confidence summarise the calibrated posterior. The unseen-tile evaluation employed the temperature fitted on the validation split.

Setting	Building	Low veg.	Tree	Imp.
Potsdam val. (post-cal)	**0.209**	**0.162**	**0.155**	0.247
Unseen Potsdam tile (post-cal, fixed T)	0.319	0.276	0.262	0.324

Table 3. Per-class RMSE for Tab. 1. Errors on the unseen tile are higher for built surfaces than for vegetation, in line with known *building–impervious* ambiguity under geographic hold-out.

on built surfaces.

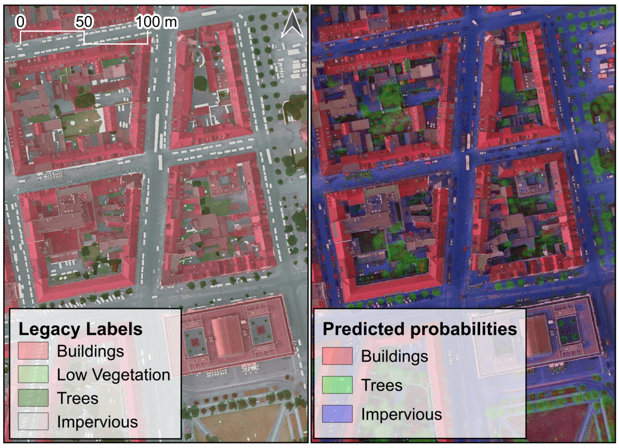
3.2 Spatial transfer

Without re-training or adaptation, the Potsdam-calibrated network was applied to Vaihingen ($\approx 0.513\,\mathrm{km}^2$) using the fixed temperature from Potsdam (T=0.770). As expected under a cross-city morphology shift (detached housing, gardens, narrow streets), performance degraded relative to the in-domain baseline: KL = 0.747, Brier = 0.433, ECE = 0.105, soft/hard accuracies = 0.632/0.643, entropy = 0.631, confidence = 0.737. Errors concentrated on built surfaces, with per-class RMSEs *Impervious* 0.358 and *Building* 0.349 exceeding *Tree* (0.287) and *Low vegetation* (0.317), consistent with the known *building*↔*impervious* ambiguity and the finer-grained residential structure of Vaihingen. Despite the shift, retaining a single scalar T kept argmax labels unchanged and preserved interpretable probability magnitudes across cities, enabling calibrated downstream comparisons.

3.3 Temporal transfer & spatio-temporal semantic change

Temporal transfer is assessed on an unseen Potsdam tile by applying the Potsdam-calibrated network (fixed T=0.770) to two acquisition dates; semantic drift is then quantified via per-class probability deltas Δp (Eq. 9) and 16-class transitions (Eqs. 10 & 11) with reliability weighting Q_{temporal} (Eq. 12).

Temporal transfer: A second, geographically unseen Potsdam tile ($\approx 0.369\,\mathrm{km}^2$) was evaluated at two acquisition times using the Potsdam-trained, globally temperature-scaled network (fixed T=0.770; Sec. 2). On the legacy epoch (t_1; contemporaneous with ISPRS labels), performance was in line with intra-city hold-out behaviour: KL = 0.552, Brier = 0.311, ECE = 0.047, soft/hard accuracies = 0.716/0.731, entropy = 0.572, confidence = 0.763, and class-wise RMSEs $\{0.272, 0.280, 0.273, 0.291\}$ for {*building*, *low vegetation*, *tree*, *impervious*}. Applying the same calibrated model to a more recent TOP (t_2, 2024) without any re-training led to the expected degradation under temporal shift (radiometry/phenology/true change): KL = 1.544, Brier = 0.709, ECE = 0.264, soft/hard accuracies = 0.407/0.410, entropy = 0.751, confidence = 0.671, and RMSEs $\{0.439, 0.361, 0.400, 0.475\}$. Retaining the single scalar T ensured that argmax labels were preserved while the magnitudes of class probabilities remained interpretable and comparable across epochs, an explicit requirement for the change analysis following.



(a) Legacy semantic labels over legacy orthophoto

(b) Predicted per-class probabilities (0-100%) over legacy orthophoto

Figure 4. In-domain example (Potsdam, t_1). (a): legacy TOP with ISPRS labels (*building*, *impervious*, *low vegetation*, *tree*). (b): calibrated per-class probability composite (R=*building*, G=*tree*, B=*impervious*) over legacy TOP

	B	LV	Tree	Imp
Soft transitions (mass fraction; global sum = 1)				
B	0.1240	0.0228	0.0276	0.1282
LV	0.0500	0.0778	0.0678	0.0500
Tree	0.0429	0.0644	0.0594	0.0359
Imp	0.0648	0.0344	0.0437	0.1063
Q^*-weighted soft transitions (global sum = 1)				
B	0.1537	0.0162	0.0210	0.1594
LV	0.0343	0.0705	0.0605	0.0289
Tree	0.0385	0.0612	0.0552	0.0310
Imp	0.0688	0.0288	0.0392	0.1328

Table 4. Soft and Q^*-weighted transition matrices. Entries within each block are globally normalised.

Spatio-temporal semantic change: Semantic drift between the calibrated Potsdam maps at t_1 and t_2 was quantified by per-class probability deltas $\Delta p_{i,c}$ (Eq. 9) and by discrete transition classes obtained from argmax pairs $\left(c_i^{(t_1)} \rightarrow c_i^{(t_2)}\right)$ (Sec. 2.5). With the four-class ontology, this yields a 16-class transition raster that encodes both persistence and change. The Δp layers localised gains and losses in *building*, *impervious*, *low vegetation*, and *tree*; the hard transition map summarised these as categorical processes (e.g., *low vegetation→impervious* for sealing, *impervious→low vegetation* for greening). For aggregated reporting, soft transition summaries S and their reliability-weighted variant S^Q (Eqs. 10 & 11) were computed, normalised to unit mass and restricted to pixels within the four-class ontology at t_1. As expected, diagonal entries (persistence) dominated, with off-diagonal mass primarily along *low vegetation→impervious* and *tree→impervious/low vegetation*, consistent with sealing/densification and canopy loss. Weighting by Q_{temporal} (Eq. 12) reduced the influence of boundary-uncertain pixels and low-confidence predictions at t_2, yielding crisper transition proportions without altering the qualitative pattern. Low-confidence regions (high u_i or entropy; Eqs. 7 & 8) were optionally masked to prioritise reliable evidence. Table 4 reports (i) soft transitions from calibrated probability mass, and (ii) a reliability-weighted variant using $Q^* \propto Q_{\text{HELIX}}^{(t_1)} \cdot \max_c p_c^{(t_2)}$. Rows are sources at t_1; columns are targets at t_2. Figure 5 illustrates the full workflow: calibrated probabilities at t_1 and t_2, the induced 16-class transition raster and the continuous Δp signal from Sec. 2.5. Persistence dominates, while off-diagonal transitions concentrate on *low vegetation→impervious* and *tree→*$\{$*impervious*, *low vegetation*$\}$, consistent with sealing/densification and canopy loss.

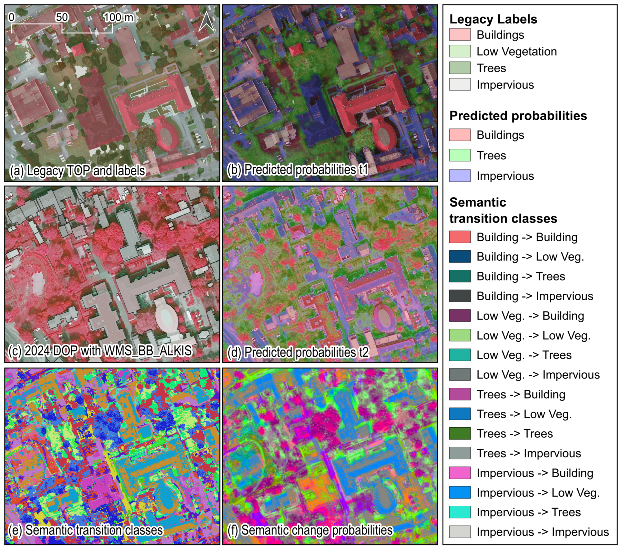


Figure 5. Temporal transfer and semantic change in Potsdam. (a) Legacy TOP with ISPRS labels (t_1). (b) Calibrated per-class probability composite over t_1. (c) Recent TOP (t_2, 2024) with current WMS BB ALKIS building footprints. (d) Calibrated probabilities over t_2 using the same globally temperature-scaled model (T fixed). (e) 16-class transition map from argmax pairs $c_i^{(t_1)} \rightarrow c_i^{(t_2)}$. (f) Per-class probability deltas $\Delta p_{i,c}$ visualising temporal semantic change.

4. Discussion

Temporal generalisation via HELIX and calibration: The combination of HELIX-style label enrichment and post-hoc temperature scaling yielded calibrated probability fields whose magnitudes transfer across tiles and years (Sec. 2). In-domain, calibration reduced ECE to ≈ 0.01 without affecting argmax labels (Sec. 3.1). Under temporal shift ($t_1 \rightarrow t_2$ in Potsdam), absolute accuracy degraded as expected (radiometry/phenology and true change), yet the fixed temperature ensured comparable confidence scales (Sec. 3.3). This comparability enables pixel-wise probability differencing $\Delta p_{i,c}$ and transition summaries that behave consistently over time (Eqs. 9, 10 & 11). In practice, weighting by Q_{temporal} (Eq. 12) concentrates evidence on reliable supervision at t_1 and confident predictions at t_2, stabilising temporal summaries without altering qualitative patterns.

Disambiguating buildings and sealed surfaces: Ambiguity between *building* and *impervious* concentrates most errors (Tabs. 3&1). This is expected given similar spectra/texture for flat roofs and paved courtyards and is amplified when only RGB is available (ablation), whereas NIR strengthens canopy–impervious separation. In many monitoring tasks (e.g., sealed-surface accounting), a macro-class *built surface* mitigates adverse effects; where differentiation is required, height cues or material proxies are needed (Sec. 4). Importantly, calibrated posteriors expose the uncertainty in these ties: near-tie regions carry lower confidence/higher entropy, which can be down-weighted or masked in downstream products.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-959-2026

Distinguishing annotation drift from prediction error: Legacy ISPRS masks (ISPRS, 2013) are temporally matched to t_1 but retain geometric coarseness and boundary offsets (Sec. 3.1). Discrepancies at kerbs, roof edges, and under canopy therefore need not imply model failure. HELIX quality Q_{H} reduces the influence of such regions during training, and calibrated probabilities reveal where supervision is brittle. Over time, persistent Δp signals aligned with urban structure are indicative of genuine drift, while isolated, low-Q_{temporal} changes or entropy spikes more likely trace label geometry or illumination artefacts. This separation is operationally useful: it prioritises polygons likely affected by label ageing for manual review, while preserving genuine process signals for reporting.

Semantic change patterns: The 16-class transition raster and the soft Q^*-weighted transition matrices (Sec. 2.5 with Eqs. 10 & 11) summarise process types in a compact, uncertainty-aware form. In Potsdam, diagonals (persistence) dominate, with off-diagonal mass chiefly along *low vegetation*→*impervious* and *tree*→{*impervious*, *low vegetation*}, consistent with sealing/densification and canopy loss (Sec. 3.3). Q^*-weighting sharpens these proportions by suppressing boundary-uncertain and low-confidence pixels, yielding crisper, more actionable summaries while keeping the qualitative story intact.

Co-registration and seasonality: Sub-pixel misalignment between t_1 and t_2 inflates apparent change along boundaries, disproportionately affecting *building–impervious* and canopy edges. This is mitigated by operating on calibrated probabilities, by reporting soft transitions (Eqs. 10 & 11), and by optionally masking high-entropy regions. In practice, a narrow erosion of boundary bands, or a tolerance buffer on Δp, further stabilises counts. Seasonal/phenological differences (e.g., leaf-on vs. leaf-off) shift spectral indices and can mimic canopy loss; here, per-scene normalisation and the Q^* weighting reduce false positives, but acquisition-date metadata remain essential for interpretation.

Probabilistic supervision without loss of semantic contrast: HELIX enrichment yields a *structured* soft target: each pixel is assigned a distribution $s_i(c)$ whose uncertainty is concentrated at boundaries, mixed pixels, and class interfaces, while remaining sharp in homogeneous regions. Edge guidance limits cross-class bleeding at strong physical discontinuities, and neighbourhood coherence and semantic margins shape the entropy of s_i so that it reflects annotation reliability rather than uniform smoothing. This produces supervision that is robust to geometric imprecision and label ageing without sacrificing semantic contrast where evidence is clear. Training matches the predicted posterior p_θ to this enriched target using a Q_{H}-weighted divergence objective. Expressing the loss in KL form highlights the distribution-matching objective and enables Q_{H} to act as a principled per-pixel weight, reducing the influence of unreliable supervision. Hard-label cross-entropy and unweighted soft-target cross-entropy are included as reference baselines. Table 5 confirms the intended behaviour: mIoU remains comparable across losses, while agreement with the enriched reference improves under soft supervision (SoftAcc) and is further concentrated on high-quality regions by Q-weighting (SoftAcc$_Q$). For the unlabeled t_2 tile, mean MaxProb is reported as a transfer-only confidence proxy.

Ablations and practical implications: Qualitative, precision-oriented effects across configuration choices were examined. When NIR was present (IRRG/RGBIR), the intended 7-channel stack $\mathrm{NIR}, R, G, \mathrm{NDVI}, \mathrm{NDWI\text{-}like}, \mathrm{Sobel}, \mathrm{Shadow}$ sharpened canopy–impervious separation and improved robustness in cast shadow; the RGB-only ablation (radiometric RGB + Sobel + shadow) showed weaker under-canopy separation and slightly more *low vegetation* false positives along tree–street seams. Edge-guided multi-scale soft labels yielded crisper boundaries and reduced overconfidence versus a lighter two-scale variant. Using Q_{H} as a continuous loss weight was more reliable than hard high-Q filtering, which underfit boundaries and nascent change. A mild confusion-aware feedback (*building*↔*impervious*) reduced flips on flat roofs and paved courtyards without blurring edges. Overall, label enrichment (soft+edge+HELIX), spectral richness (when NIR is available), and post-hoc calibration dominated precision and reliability; secondary knobs mainly traded runtime or recall for smaller effects.

Method (post-cal)	mIoU↑	SoftAcc↑	ECE↓	SoftAcc$_Q$↑	MaxProb$_{\mathrm{fut}}$↑
HardCE	**0.635**	**0.769**	0.030	**0.781**	**0.657**
SoftCE	0.599	0.749	**0.006**	0.762	0.635
Ours (Q-KL+AF)	0.618	0.760	0.018	0.773	0.642

Table 5. Post-calibration ablation on Potsdam t_1 validation; $\mathrm{MaxProb_{fut}}$ is mean maximum calibrated probability on unlabeled t_2. SoftAcc is agreement with HELIX soft labels; SoftAcc$_Q$ is Q_{H}-weighted.

Limitations and validation needs: A single global temperature T (Sec. 2) normalises confidence across tiles, cities, and dates while preserving argmax, but class-specific bias can remain; under sensor/season changes a per-sensor scalar is advisable, and per-class calibration can reduce ECE further at the cost of strict comparability and possible argmax flips. Legacy ISPRS masks, though contemporaneous at t_1, are geometrically coarse and age over time; HELIX weights Q_{H} temper their influence, yet bias remains, especially along kerbs and roof edges. Sub-pixel misregistration and phenology/illumination shifts inflate apparent boundary change; operating on Δp and soft transitions (Eqs. 9, 10 & 11) and using small erosions/tolerance bands stabilise counts. Without NIR, canopy–impervious separation weakens (Ablations), increasing near-ties—an operational trade-off in RGB-only deployments. Uncertainty/entropy masking raises precision at the expense of recall and reported volumes (Sec. 3.3); because probabilities are calibrated, it is recommended to report change areas with and without masking, together with reliability summaries (ECE, Q^*). City-scale tiling with light TTA is tractable, but heavier TTA yields diminishing returns and ensembles may be runtime-limited. For deployment, an operational audit should stratify by transition type, $|\Delta p|$, and Q_{temporal}; gather independent evidence (recent TOPs, DSM/nDSM or LiDAR, authoritative footprints); estimate precision/recall with bootstrap CIs; include boundary erode/no-erode sensitivity; and run negative controls in nominally stable areas. Beyond cross-epoch ageing, the U-Net was trained on *legacy* TOPs with *legacy* labels at t_1, which do not always align perfectly even at t_1 (coarse polygons, sub-pixel shifts, cartographic generalisation, class ambiguities). This motivated edge-adaptive soft labels and HELIX weighting to reduce overconfidence near boundaries. In Fig. 3, calibrated probabilities follow image structure more closely than polygons; the yellow inset shows a block labelled *building* although the image mixes roof, paved yard/parking, and vegetation. Such over-extended polygons and small misregistrations bias reported accuracy. The broader *label ageing* issue is illustrated in

Fig. 1: since the ISPRS acquisition, buildings have been demolished/added and surfaces and vegetation changed, explaining part of the disagreement with “ground truth” at t_2. For t_2 the official building footprints from the state of Brandenburg, provided via the WMS BB ALKIS (Official Property Cadastre Information System) service (Landesvermessung und Geobasisinformation Brandenburg, 2024) was used as an external, up-to-date reference for *buildings* only. Panels (c)–(d) show strong spatial agreement between predicted *building* probabilities and the current footprints, including newly built structures and resurfaced courtyards. While BB ALKIS is not a pixelwise ground truth for all classes, it provides an independent sanity check at t_2.

Outlook: With harmonised preprocessing and a single global temperature T, the pipeline is directly deployable across municipalities and acquisition dates. In practice, a light per-sensor recalibration and a small, stratified audit (by transition class and Q_{temporal}) are sufficient to certify fitness-for-use before reporting. Integrating height cues (DSM/nDSM, stereo, or SAR proxies) is expected to further reduce *building*↔*impervious* ambiguities and can be incorporated via early or late fusion followed by recalibration. Calibrated probability differences Δp and Q-weighted transition summaries support scalable, high-frequency monitoring and enable targeted review of uncertain or high-impact areas, providing a practical basis for trustworthy urban change assessment under label ageing.

5. Conclusion

This work demonstrates that legacy supervision can be transformed into calibrated, transferable probability fields for semantic spatio-temporal label-drift analysis, quantifying what changed, where, and with what confidence. Edge-adaptive soft labels with HELIX quality, a Q-weighted KL objective, and a single global temperature make ageing-prone supervision explicit and produce probabilities comparable across tiles, cities, and years without retraining. In-domain, calibration reduces ECE to ≈ 0.01 while preserving argmax and accuracy (Sec. 3.1); under spatial and temporal transfer, the fixed scalar T keeps confidence scales interpretable as scenes change (Secs. 3.2, 3.3). For change analysis, per-class Δp and a 16-class transition raster, supplemented by Q^*-weighted summaries, provide compact, auditable evidence of sealing, construction, greening, and canopy loss (Eqs. 9, 10 & 11). The contribution is therefore not a universal accuracy gain, but a reliability-aware reformulation of legacy supervision for interpretable urban EO change workflows.

References

Amini, S., Saber, M., Rabiei-Dastjerdi, H., Homayouni, S., 2022. Urban Land Use and Land Cover Change Analysis Using Random Forest Classification of Landsat Time Series. *Remote Sensing*, 14(11). doi.org/10.3390/rs14112654.

Ding, Z., Han, X., Liu, P., Niethammer, M., 2021. Local Temperature Scaling for Probability Calibration. *CoRR*, abs/2008.05105. doi.org/10.48550/arXiv.2008.05105.

Goswami, M., Mohanty, S., Dey, S., Mukherjee, A., Pattnaik, P. K., 2025. Convolutional neural network segmentation for satellite imagery data to identify landforms using u-net architecture. A. K. Das, J. Nayak, B. Naik, M. Himabindu, S. Vimal, D. Pelusi (eds), *Computational Intelligence in Pattern Recognition*, Springer, Singapore, 115–126.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q., 2017. On Calibration of Modern Neural Networks. *CoRR*, abs/1706.04599. doi.org/10.48550/arXiv.1706.04599.

Hauser, S., 2025. Automated Feature & Label Refinement in the Context of Environmental Monitoring by Multi-Modal Satellite Data. PhD thesis, Karlsruhe Institute of Technology (KIT). doi.org/10.5445/IR/1000185980.

Hauser, S., Augner, L., Schmitt, A., 2025. Perfect Labelling: A Review and Outlook of Label Optimization Techniques in Dynamic Earth Observation. *Remote Sensing*, 17(7), 1246. doi.org/10.3390/rs17071246.

ISPRS, 2013. Urban modelling and semantic labelling benchmark. `https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/default.aspx`. Accessed: 5 November 2025.

Landeshauptstadt Potsdam, 2024. Statistische grunddaten zur landeshauptstadt potsdam. `https://potsdam.de/de/statistische-grunddaten-zur-landeshauptstadt-potsdam`. Accessed: 8 October 2025.

Landesvermessung und Geobasisinformation Brandenburg, 2024. ALKIS Brandenburg. `https://geobroker.geobasis-bb.de/gbss.php?MODE=GetProductInformation&PRODUCTID=253b7d3d-6b42-47dc-b127-682de078b7ae`. Accessed: 5 November 2025.

Lemenkova, P., 2024. Support Vector Machine Algorithm for Mapping Land Cover Dynamics in Senegal, West Africa, Using Earth Observation Data. *Earth*, 5(3), 420–462. doi.org/10.3390/earth5030024.

Li, J., Zhu, L., Du, Z., 2024. *Introduction to Domain Adaptation*. Springer Nature, Singapore, 1–6.

Ma, J., Wang, C., Liu, Y., Lin, L., Li, G., 2023. Enhanced soft label for semi-supervised semantic segmentation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1185–1195.

Majidizadeh, A., Hasani, H., Jafari, M., 2023. Semantic Segmentation of UAV Images based on U-NET in Urban Area. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, X-4/W1-2022, 451–457. doi.org/10.5194/isprs-annals-X-4-W1-2022-451-2023.

Northcutt, C. G., Athalye, A., Mueller, J., 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *CoRR*, abs/2103.14749. doi.org/10.48550/arXiv.2103.14749.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR*, abs/1505.04597. doi.org/10.48550/arXiv.1505.04597.

Seto, K. C., Güneralp, B., Hutyra, L. R., 2012. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proc. Natl. Acad. Sci.*, 109(40), 16083-16088. doi.org/10.1073/pnas.1211658109.

Stadt Vaihingen an der Enz, 2024. Zahlen & Fakten. `https://www.vaihingen.de/unsere-stadt/stadt-vaihingen/zahlen-fakten`. Accessed: 8 October 2025.

United Nations, 2022. World urbanization prospects: The 2022 revision. `https://population.un.org/wup`. Accessed: 7 October 2025.

World Bank, 2024. Urban population (% of total population). Source: United Nations, Department of Economic and Social Affairs, Population Division, World Urbanization Prospects. https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS. Accessed: 8 October 2025.

Zhang, H., Jiang, Z., Zheng, G., Yao, X., 2023. Semantic Segmentation of High-Resolution Remote Sensing Images with Improved U-Net Based on Transfer Learning. *International Journal of Computational Intelligence Systems*, 16(1), 181. doi.org/10.1007/s44196-023-00364-w.