Highlights

How high are we?
Large-Scale Building Height Estimation using Sentinel-1 SAR and Sentinel-2 MSI Time Series

Ritu Yadav, Andrea Nascetti, Yifang Ban

- Proposed T-SwinUNet for joint building height estimation and footprint segmentation
- Learning temporal correlation of building features across time series
- Dataset across Netherlands, Switzerland, Estonia and parts of Germany
- Building height prediction with 1.89 m RMSE, footprint with 0.69 F1 score
- Demonstrated model's generalizability on unseen data

How high are we?
Large-Scale Building Height Estimation using Sentinel-1 SAR and Sentinel-2 MSI Time Series

Ritu Yadav, Andrea Nascetti, Yifang Ban

[a]*Division of Geoinformatics, KTH Royal Institute of Technology, Sweden*

Abstract

Accurate building height estimation is essential to support urbanization monitoring, environmental impact analysis and sustainable urban planning. However, conducting large-scale building height estimation is a challenging task. While Deep Learning (DL) has proven effective for large-scale mapping, the lack of advanced DL models specifically tailored for height estimation remains a challenge, particularly when using open source Earth Observation data. In this study, we propose an advanced DL model (T-SwinUNet) for large-scale building height estimation leveraging Sentinel-1 Synthetic Aperture Radar and Sentinel-2 MultiSpectral Instrument time series. In the proposed T-SwinUNet, the semantic feature learning capabilities of the efficientnet encoder are combined with the local/global feature comprehension capabilities of Swin transformers. A temporal attention module is added to learn the correlation between constant and variable features of building objects over time which not only helps in differentiating building objects from the surroundings but also in learning salient features for building height estimation. The model is trained on a multi-task to predict both building height and footprint at 10 m spatial resolution. The model is evaluated on data from the Netherlands, Switzerland, Estonia, and Germany. The extensive evaluation and comparison with state-of-the-art DL models show that our proposed T-SwinUNet model yields Root Mean Square Error (RMSE) of 1.89 m, surpassing the state-of-the-art at 10m spatial resolution. Further assessment at 100 m resolution shows that our predicted building heights (0.29 m RMSE, 0.75 R^2) also outperformed the global building height product GHSL-Built-H R2023A product(0.56 m RMSE and 0.37 R^2). Our implementation is available at: https://github.com/RituYadav92/Building-Height-Estimation

Preprint submitted to Remote Sensing of Enviornment *February 19, 2024*

Keywords: Height Estimation, Footprint Segmentation, Sentinel, Spatiotemporal, Regression, MultiTask Learning.

1. Introduction

More than half of the world's population currently lives in cities. By 2050, an estimated 7 out of 10 people will likely live in urban areas. While cities contribute more than 80% of global GDP they are also accountable for major energy consumption and carbon emission [UN, 2022]. Therefore, urbanization monitoring is essential to support sustainable development. In the last decade, 2-dimensional (2D) urban monitoring such as building footprint extraction has received considerable attention and resulted in many high-resolution and global products [Li et al., 2020b, Marconcini et al., 2021, Hafner et al., 2022, Huang et al., 2022b, Hu et al., 2023, Chen et al., 2023b]. Despite being the essential component of urbanization, the third dimension (3D) or height has not been equally investigated. There are relatively few studies on building height estimation, and most of them focus only on a few sites, e.g. [Huang et al., 2020, Liu et al., 2022, Yadav et al., 2022, Chen et al., 2023a, Dong et al., 2024]. Accurate estimation of building height plays an important role in urban planning, as it is correlated with transportation, telecommunications, energy consumption [Marconcini et al., 2020], population [Leichtle et al., 2019], urban heat island effect [Wu et al., 2022], and urban climate [Xi et al., 2021] and is also one of the key parameters in their quantification.

While airborne laser scanning (ALS) and high-resolution aerial images offer detailed information ideal for accurate building height estimation, especially in dense urban areas, they are not suitable for large-scale mapping due to their high cost and time-consuming data collection processes [Cao and Huang, 2021, Liu et al., 2022]. Earth observation, on the other hand, is an effective and promising tool for large-scale mapping and monitoring. Although some studies have explored building height estimation using very high-resolution satellite images [Recla and Schmitt, 2022, Liu et al., 2022, Chen et al., 2023a], the restricted accessibility of their data sources limits scalability for large-scale applications.

In contrast, satellite missions such as Sentinel-1 and Sentinel-2 provide open access to global SAR and optical data free of cost. Their frequent revisit cycles coupled with a spatial resolution of 10 meters (m) and open data access, not only make them suitable candidates for large-scale 3D mapping but also allow for frequent update cycles. In recent years, several studies have tried to fill this gap and estimate building heights using these free-of-cost satellite imagery. For example, [Li et al., 2020c] proposed to estimate building height using Sentinel-1 SAR data. The authors developed a new VVH indicator, which was evaluated in seven major cities in the US to estimate building heights at 500m resolution with RMSE of 1.5 m. In general, Sentinel-1 SAR is useful for estimating building height, as there is a positive correlation between the derived backscatter coefficient and the height of the buildings [Koppel et al., 2017]. However, apart from building height, the backscatter coefficient can be influenced by other factors adding uncertainties to the estimate [Li et al., 2020c]. These factors can be a metal surface with high reflectivity, certain types of building structure that cause double bouncing [Li et al., 2016], scattering variation from the tree canopy and building density [Corbane et al., 2008]. Adding optical data in the height estimation process helps to overcome some of these factors, for example, [Li et al., 2020a] proposed using Sentinel-1 SAR and optical data from Landsat-8 OLI incorporating auxiliary data (OSM, cadastral data and commercial maps). The authors estimated building height at the continental scale by applying a random forest model. However, the model overestimated the heights of small buildings and the coarse spatial resolution of 1 Km made it impossible to examine height differences in various building structures. Both resolution and scale are improved by the GHSL-Built-H R2023A product [Pesaresi et al., 2021], providing global building height at 100 m spatial resolution. The building height is derived using a regression method on multiple statistics calculated from ALOS Global Digital Surface Model - 30 m, the NASA Shuttle Radar Topographic Mission data - 30 m, and the Sentinel-2 MSI global pixel-based image composite from L1C data for the period 2017-2018. The estimations are referred to the year 2018. The resolution is further improved by [Esch et al., 2022], where the global scale building height is estimated at 90 m resolution extending the World Settlement Footprint (WFS) [Marconcini et al., 2021] to 3D. The estimated building heights have been validated showing a promising accuracy with an RMSE of 6.01 m.

Meanwhile, [Huang et al., 2022a] estimated building heights in China at a better spatial resolution of 30 m, achieving a RMSE of 4.98 m. However, it is worth noting that both [Esch et al., 2022] and [Huang et al., 2022a] rely on commercial DSMs collected by the TanDEM-X and ALOS missions, respectively, which require processing stereo satellite image pair, making it a complex and expensive process thereby frequent updation of building height can be challenging. The spatial resolution of building height estimation maps is further improved to 10 m by [Frantz et al., 2021], where the authors proposed using the change in the length of building shadows with each month. They derived an exhaustive number of spatial, spectral and temporal statistical features along with several handcrafted features from Sentinel-1 SAR and Sentinel-2 MSI time series data and trained a support vector machine regression model to predict building height. This approach was tested in five major areas in Germany and the derived building heights show an RMSE of 6.07 m. Another recent study by [Wu et al., 2023] focused on estimating building heights in China at 10m resolution. They adopted a combined approach that integrated elements from both [Li et al., 2020a] and [Frantz et al., 2021], supplementing their methodology with additional data sources such as ALOS PALSAR, LUOJIA 1–01, WFS footprints, and DEM data. Although this study resulted in a similar RMSE of 6.1 m, comparable to that of [Frantz et al., 2021], it operated in more complex urban regions characterized by a wider distribution of high-rise buildings. [Dong et al., 2024] also estimated building height in a complex urban area of Hangzhou, China, combining indices from Sentinel-1/2 and a physical model where several statistics, such as the orientation angle of the building, number of vertices, the distance from neighboring buildings, the road, and many others are calculated based on prior ground-based knowledge. They trained an XGBoost model with these features, achieving an RMSE of 6.64 m at the individual building level. However, the method relies on prior ground-based knowledge, which poses challenges for large-scale applications.

In the last decade, compared to machine learning algorithms, DL models became popular in remote sensing due to their powerful discriminative ability and rich representation learning [Asokan and Anitha, 2019]. The approaches mentioned above use machine learning algorithms with handcrafted features, which often have limitations in capturing the complex and high-dimensional nature of remote sensing data. In contrast, DL models can directly learn from raw features without relying on handcrafted ones [Zhou et al., 2018, Yan et al., 2020]. For instance, [Cai et al., 2023] proposed a dual branch DL network (BHE-Net) that outputs building footprints majorly using Sentinel-1 SAR in one branch and building height using Sentinel-2 MSI in the second branch. The outputs are then combined to estimate building height at 10 m spatial resolution. Their model was evaluated in three regions of China and the results show an RMSE of 4.65 m. Meanwhile, [Yadav et al., 2023] developed another deep fusion network (MBHR-Net) and proposed using time series data of Sentinel-1 SAR and Sentinel-2 MSI data. The model was evaluated across ten cities of Netherlands and exhibited an RMSE of 3.73 m. Although [Yadav et al., 2023] used time series data, the model did not utilize the spatio-temporal features of the time series as they considered the time series images as augmented images with different seasonal effects. A more advanced DL model is required to exploit spectral and spatio-temporal features of rich SAR and MSI time series. Furthermore, the scale of the studies can be improved by including available training data from different countries.

Given the gaps, we propose T-SwinUNet model, which utilizes free Sentinel-1 SAR and Sentinel-2 MSI data to achieve scalability and frequent update cycle. We propose using time series data to learn from the temporal correlation of the features, as it can differentiate between the building and surroundings while capturing height features like building shadow over time. Our T-SwinUNet model is embedded with temporal attention and window based multi-head attention to efficiently learn salient spatial, spectral and temporal features. The proposed model improved building height estimation accuracy, and application scale at fine spatial resolution. The main contributions of this work are summarized as follows:

- We proposed T-SwinUNet, a novel model that integrates fine-grained pattern capturing capabilities of efficientnet with temporal attention to extract spatio-temporal features of multimodal time series data. The model is further integrated with the brilliant global/local feature learning abilities of Swin Transformer.

- We introduced a multitask decoder that takes advantage of the complementary tasks of building height estimation and footprint segmentation. The model not only learns two tasks simultaneously, but also improves overall performance through a consistency loss.

- We conducted comprehensive experiments and ablation to demonstrate the contribution of different parts of the proposed model. The results show that our proposed model achieved state-of-the-art building height estimation results at 10 m spatial resolution and also outperformed GHSL-Built-H R2023A, a global building height product at 100 m spatial resolution.

- We demonstrate that merging predicted building heights with existing building footprints yields precise instance-level building height estimates achieving an improved RMSE of 1.60 m.

2. Study Area and Data Collection

This study is conducted on building data across four countries, Netherlands, Switzerland, Estonia and parts of Germany i.e., Hamburg, Brandenburg, Sachsen and North Rhine-Westphalia. The defined training and test areas are shown in Figure 1. The test areas are kept separate to perform un-

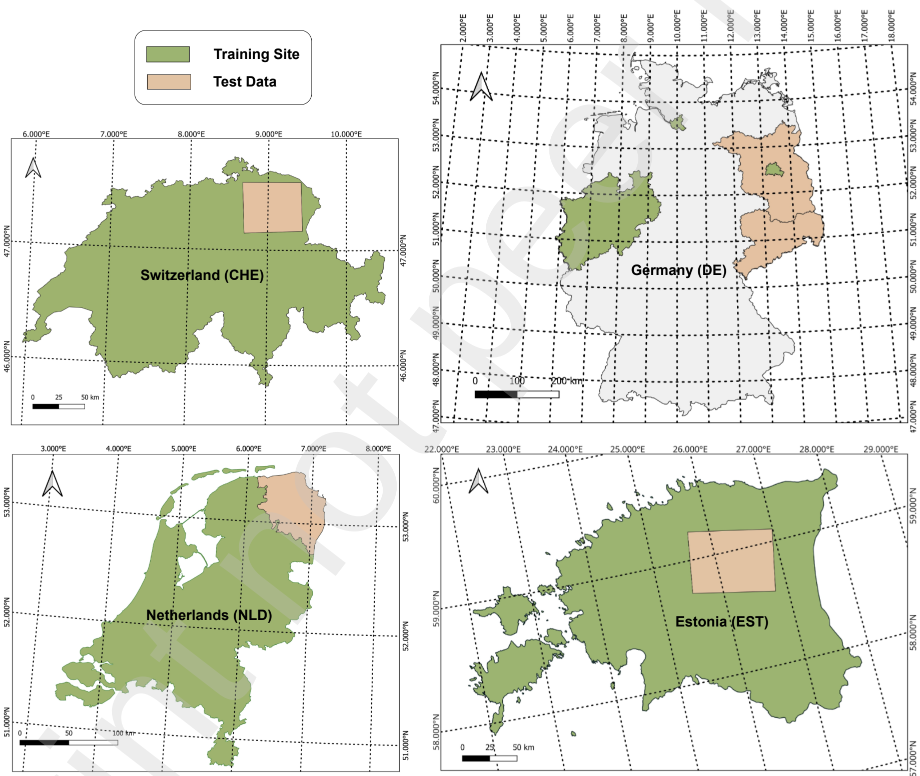

Figure 1: Study site map (CRS 3035)

biased evaluation. Our dataset comprehensively covers these areas, encompassing not only dense buildings in major cities but also sparsely distributed buildings in rural regions. It is worth noting that the urbanization patterns in Switzerland and Estonia are heterogeneous. Therefore, for these countries, test areas were selected to represent a mix of dense cities and sparse rural settlements. In Germany, the test areas span Brandenburg and Sachsen states, where urban density varies, yet many cities are densely populated. Given the relatively consistent urbanization density in the Netherlands, a random area (Groningen province) was chosen as the test area. The specific statistics on train and test areas are provided in Figure 3 and explained later in data filtering and splits 2.3 subsection.

In both the training and test sites, random patches of size 1280 m× 1280 m are sampled using the area random sampling method. These patches are sampled with a 20% overlap to ensure comprehensive coverage. Data collection involves gathering reference data which contains building heights and building footprints and input data which contains Sentinel-1 SAR, and Sentinel-2 MSI time series data. Both reference and input data are collected for each sample patch. While reference data are sourced from multiple providers listed in Table 1, Sentinel-1 SAR and Sentinel-2 MSI data are collected through the Google Earth Engine Python API. The entire dataset adheres to the European terrestrial reference system EPSG:3035, and all data sources are publicly available at no cost. Figure 2 illustrates our data collection process.

2.1. Reference Data

The building Height references provided at the sources(Table 1) are derived from either aerial stereo images or airborne LiDAR data collected over many years. These references include the height of each individual building.

Table 1: Reference data specifications.

Site	Year	Sensors	Resolution	#Patches(train+test)	Data Provider
Netherlands	2014-19	ALS	2m	14835, 1440	TUDelft3d
Germany	2018-21	Stereo Aerial Photo, ALS	1m	8278, 1794	German State Government
Switzerland	2018-21	Stereo Aerial Photo	<1m	12735, 1623	Swiss Federal Office of Topography
Estonia	2017-20	ALS	1m	6314, 620	Estonian Land Board

The reference building heights for all four sites are available at 1 to 2 m spatial resolution. We collected the reference data for all sampled patches, given that each patch has a minimum of 10 buildings. This filtering process helps to avoid numerous patches with rare to no buildings, resulting in a more balanced dataset.

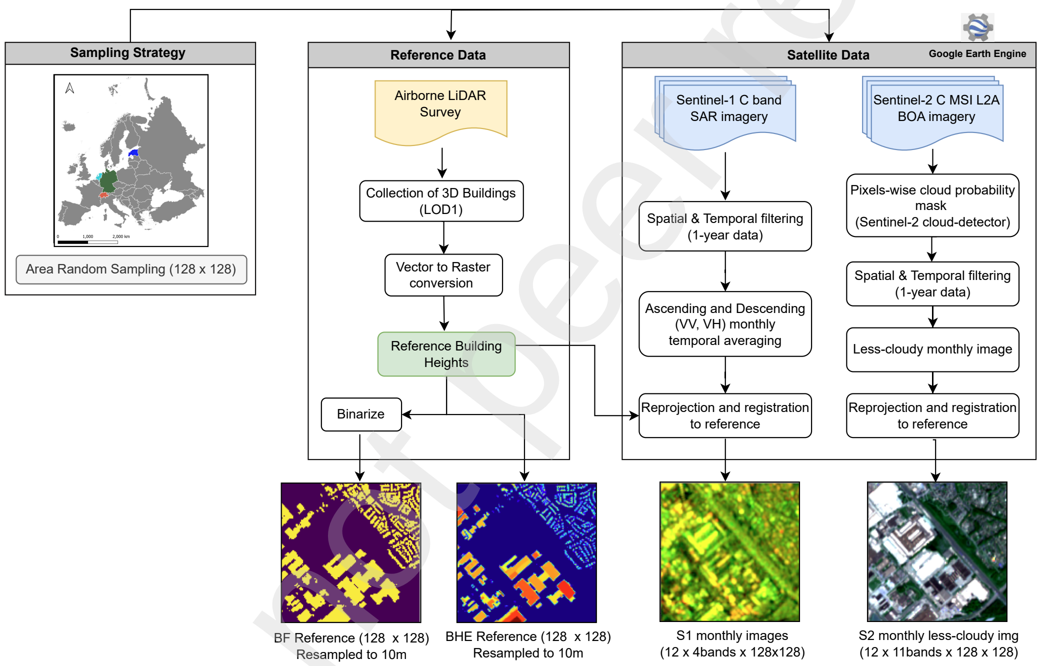

Figure 2: Data Collection Framework.

2.2. Sentinel-1 SAR and Sentinel-2 MSI Time Series Data

The input data used in this study consists of time series data from Sentinel-1 SAR ground range detected and Sentinel-2 MSI Level-2A. These two Copernicus Sentinel missions provide free data with global coverage. The data is useful for large-scale analysis because of their ability to acquire images with large swaths and at good temporal resolution. For each reference patch, 12 Sentinel-1 SAR images (one image for each month) and 12 Sentinel-2 MSI images are collected, all at 10-meter resolution. The year of Sentinel-1 SAR and Sentinel-2 MSI data for each site is based on the acquisition year of the corresponding reference data (see Table 1). For the Netherlands, Estonia and Germany we chose 2019 while for Switzerland the Sentinel-1 SAR, Sentinel-2 MSI data from the year 2021 was collected. After automatic preprocessing i.e. thermal noise removal, radiometric calibration, and terrain correction, the monthly average is computed for both ascending and descending orbits, which helps in reducing the speckle. The data is downloaded with 4 bands i.e. VV and VH polarizations for both orbits. The Sentinel-2 MSI data undergoes radiometric calibration and atmospheric correction to produce Bottom-Of-Atmosphere (BOA) reflectance data. Then the monthly less cloudy composites are generated and the data is downloaded with 5 bands i.e. Band 2 (blue), Band 3 (green), Band 4 (red), Band 8 (near-infrared), and Band 12 (short-wave-infrared).

2.3. Data Processing and splits

The reference building height maps collected from the source consist of continuous values starting from zero. Since it is improbable to have a building with less than 1.0 m of height, we adjusted any values below 1.0 m to zero. Subsequently, the building footprint references were generated by binarizing the building height maps with a threshold of 1.0 m. Both building height and footprint references were then resampled to a spatial resolution of 10 m using an inter-area resampling technique. Also, the backscatter of Sentinel-1 SAR and the reflectance values of Sentinel-2 MSI bands are normalized using 2 and 98 percentiles computed over all data samples. The total number of train and test samples per site are specified in Table 1. The train patches are further split into train and validation sets using 80/20 splits. The test set was kept separate from the training process. The distributions of the reference building heights across the train and test sets are compared in Figure 3. The first histogram displays the distribution of reference heights across all sites, while the subsequent four histograms illustrate the distribution for individual sites.

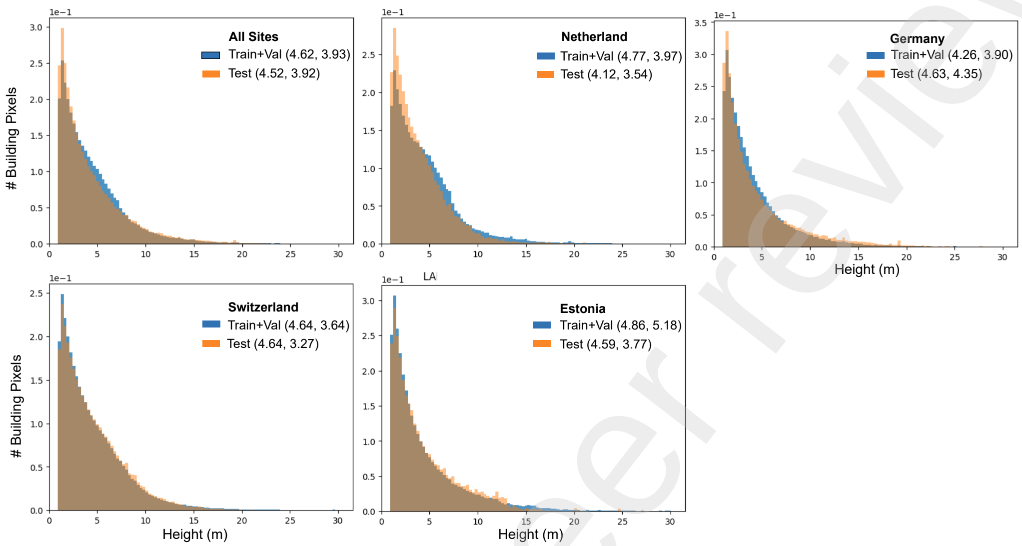

Figure 3: Normalized histograms to show the distribution of building height reference on each site. The values in the legend are the mean and standard deviation of train + validation data and test data respectively.

3. Methodology

To estimate building height, a deep multi-task supervised model is employed that takes coregistered Sentinel-1 SAR and Sentinel-2 MSI time series images as input, and outputs building height along with building footprints. While the model can be focused on only one task of building height estimation, the complementary task of building footprint segmentation, i.e. the existence of a building or no building, is helpful to avoid height estimations of non-building objects. Figure 4 depicts the network architecture of the proposed **T**emporally attentive and **Swin** transformer enhanced dual task **UNet** model named as **T-SwinUNet**. The following subsections explain the architecture, training and implementation details of the network.

3.1. Network Architecture

The proposed network, T-SwinUNet, follows an encoder-decoder structure where the decoder is composed of two branches, a regression branch to estimate building heights and a segmentation branch to predict building footprints. Network input is a time series of co-registered Sentinel-1 SAR and Sentinel-2 MSI images representing the same geographical area. The input has dimension $x \varepsilon R^{t \times H \times W \times C}$, where $H \times W \times C$ represents the spatial resolution and t represents the temporal range of the input.

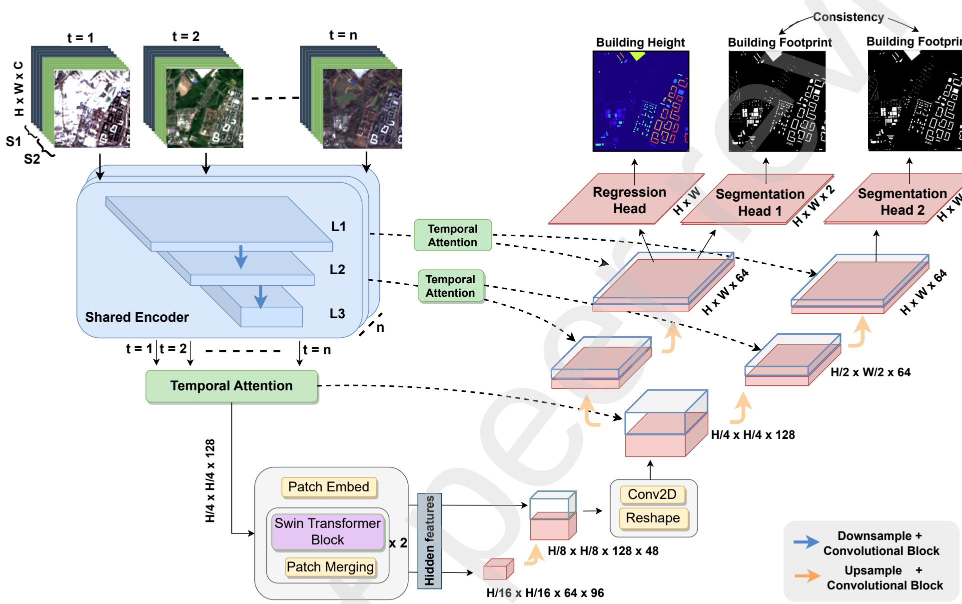

Figure 4: The proposed T-SwinUNet for building height estimation and footprint segmentation.

The input is fed into a shared efficientnet-B4 encoder that processes the time series input with an effective compound scaling that captures fine-grained features [Tan and Le, 2019]. The output features are extracted at three stages or levels $(i \varepsilon L1, L2, L3)$ at resolutions $H \times W \times 64$, $\frac{H}{2} \times \frac{W}{2} \times 64$, and $\frac{H}{4} \times \frac{W}{4} \times 128$. At each resolution level, we get n = 12 sets of features corresponding to 12 time stamps. A temporal attention module is applied at each level to correlate temporal features across time stamps. To implement temporal attention, We used a multi-head self-attention based module proposed in L-TAE (Lightweight-Temporal Attention Encoder) [Garnot and Landrieu, 2020]. The module generates an attention mask of input shape. The input features from the encoder are multiplied by the generated attention mask and added along the temporal dimension. After temporal attention on the third stage (L3) output features of the shared encoder, the output features are of size $\frac{H}{4} \times \frac{W}{4} \times 128$. A patch embedding layer is applied to generate 3D tokens which are then mapped to latent embedding space of size $D = 48$. With window size [7, 7, 7], patch size [2, 2, 2] and number of heads [3, 6], two consecutive Swin transformer blocks [Liu et al., 2021] are employed to apply multi-head attention with the shifted window technique. After each Swin transformer block a patch merging layer is applied to downsample the output by a factor of 2. A patch merging layer concatenates the 2x2 neighboring patches and applies a linear layer on top. The output hidden features are upsampled, concatenated, reshaped and fed into the upsampling decoder.

The decoder contains two branches, the first branch outputs building height as well as building footprints whereas the second branch outputs only building footprints. Both branches process features in three levels as follows. At each level ($L\varepsilon 1, 2, 3$), the output features from the encoder are enhanced through temporal attention and concatenated with the same level features of the decoder through skip connection. After each concatenation, a convolutional block is applied followed by the convolutional transpose layer that upsamples the features by a factor of two.

At the end of the first decoder branch, a regression head and a segmentation head (segmentation head1) are applied which are convolutional layers with one and two output channels respectively. The regression head is followed by a relu and the segmentation head by a sigmoid activation function. The outputs of the first branch are one building height map and one building footprint map each of size $H \times W$. At the end of the second decoder branch, a segmentation head (segmentation head2) similar to the first branch is applied followed by a sigmoid function. The output from the second branch is a building footprint map of size $H \times W$. A consistency is maintained between the building footprint outputs from the two branches so that the second branch can guide the first branch to efficiently learn the presence or absence of the building and avoid estimating height of non-building objects.

3.2. Training

The network is trained using a supervised regression loss (L_{reg}), two supervised segmentation losses (L_{seg}, L_{rseg}) and a unsupervised consistency loss (L_{consis}). The regression loss L_{reg} is used to train the model for building height regression task. The loss contains an RMSE loss to calculate the loss over all pixels (L_{rmse}) and an RMSE loss specific to nonzero label pixels

(L_{nz_rmse}). The regression loss is expressed as follows :

$$L_{reg} = 0.5 * L_{rmse} + 0.5 * L_{nz\_rmse}$$
$$where, L_{rmse} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{H}_i - H_i)^2}{n}} \tag{1}$$

Here $\hat{H}_i$ is the reference height value at pixel i and H_i is the corresponding predicted height. The two supervised segmentation losses (L_{seg}, L_{rseg}) are used to train the model for building footprint segmentation tasks. Both L_{seg} and L_{rseg} are composed of dice loss (L_{dice}) Sudre et al. [2017] and focal loss (L_{focal}) Lin et al. [2017] given as follows :

$$\begin{aligned} L_{seg} &= L_{dice} + L_{focal} \\ L_{rseg} &= L_{dice} + L_{focal} \\ where, L_{dice} &= 1 - \frac{2\hat{F}F}{\hat{F} + F}, \\ L_{focal} &= -(1-p)^{foc} log(p) \end{aligned} \tag{2}$$

Here $\hat{F}$ is the reference segmentation class, F is the predicted segmentation class, p is the class probability and foc is the focusing parameter. Finally, an unsupervised consistency loss (L_{consis}) is used to maintain consistency between the two building footprint segmentation outputs. The loss is implemented as well known IoU (Intersection over Union) loss between the two segmentation outputs. The following equation gives the combined objective function (L_{obj}), where α, β, and γ, are the weight parameters for the four losses.

$$L_{obj} = \alpha * L_{reg} + \beta * L_{rseg} + \beta * L_{seg} + \gamma * L_{consis} \tag{3}$$

3.3. Implementation Details

The time series was augmented by introducing a random channel drop (noise) with a probability of 0.2. The added noise has a regularization effect during training which in turn helps to reduce overfitting. In the objective function (equation 3), the weight parameter α was set to 2.0 to prioritize the building height estimation task, while the weight parameters for the BF detection task, β and γ, were set to 1.0, giving equal importance to footprint segmentation and the consistency between the two segmentation outputs.

The focusing parameter in the focal loss was set to two. All hyperparameters were fine-tuned based on the training and validation datasets, and the evaluation was done on the test set. The network was trained for 100 epochs with a batch size of 4, using the AdamW optimizer, an initial learning rate of 0.0001, and a decay rate of 0.5. The learning rate was decreased to 0.000001, with the "reduce on plateau" method controlling the decay steps. All the implementation was done in PyTorch and the experiments were carried out on an NVIDIA GeForce RTX 3080 GPU.

3.4. Evaluation Metrics

The predicted building heights are evaluated using two metrics RMSE and R^2 score. The RMSE indicates the accuracy of predicted heights with respect to reference and R^2 score measures the effectiveness of the model in capturing the variance in building heights. These two metrics are strategically calculated on building pixels (those with nonzero labels) to provide a more precise and focused assessment of building height predictions, mitigating any background bias. The RMSE and R^2 score formulas are given in Eq. 4, 5, where n is the number of validation samples, $BH_{est,i}$ is the estimated building height and $BH_{ref,i}$ is reference building height.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(BH_{ref,i} - BH_{pred,i})^2}{n}} \tag{4}$$

$$R^2 = 1 - \frac{(n-1)\sum_{i=1}^{n}(BH_{ref,i} - BH_{pred,i})^2}{(n-2)\sum_{i=1}^{n}(BH_{ref,i} - BH_{pred,i})^2} \tag{5}$$

The predicted building footprints are evaluated using four well-known metrics, recall, precision, F1 score and Intersection over Union (IoU). The recall and precision evaluate the completeness and accuracy, respectively, of predicted building pixels compared to reference building pixels whereas the F1 score, being the harmonic mean of recall and precision, provides a balance between minimizing false positives and false negatives. The IoU metric measures the overlap between predicted and reference building footprint pixels.

Furthermore, it is particularly important to ensure that the model predicts the height of buildings and not some surrounding object. To ensure this correspondence, the recall, precision, F1 score and IoU metric are also calculated between the reference building footprints and predicted building height binarized with a threshold of 1.0 meter i.e. pixel with predicted height >= 1.0 is categorized as a building pixel (1.0) otherwise background (0.0).

A good model should predict building height with a low RMSE score and a high R^2 value, and building footprints with high recall, precision, and IoU. All these five metrics follow the range [0,1].

4. Results

The proposed T-SwinUNet model is evaluated on the test set, which comprises four distinct regions, each corresponding to a different country. These regions are depicted in Figure 1. A detailed quantitative evaluation, ablation study and qualitative evaluation are presented in the subsections 4.1, 4.2 and 4.3, respectively. The evaluation is further followed by the generalizability test (subsection 4.4) and comparison of our results with the state-of-the-art global building height product GHSL-Built-H R2023A at 100 m (subsection 4.5).

As shown in Table 2, T-SwinUNet predicted building heights (BH) with a good RMSE score of 1.89m and R^2 of 0.534. Building footprints are predicted with 0.59 IoU. A threshold of 0.5 was used to separate the background and building footprint classes. We also evaluated the direct correspondence

Table 2: Building Height (BH) and Building Footprint (BF) evaluation over test set at 10 m spatial resolution (results over 5 runs).

	RMSE (m)$\downarrow$	R^2 $\uparrow$	Recall $\uparrow$	Precision $\uparrow$	IoU $\uparrow$	F1 $\uparrow$
BH	**1.89**$\pm$0.016	**0.53**$\pm$0.009	0.71$\pm$0.014	0.66$\pm$0.009	**0.58**$\pm$0.010	**0.69**$\pm$0.013
BF			0.72$\pm$0.011	0.67$\pm$0.006	**0.59**$\pm$0.007	**0.69**$\pm$0.008

between the predicted heights and reference building footprints by measuring the overlap between the binarized building height prediction (threshold 1.0 m) and the reference building footprints. T-SwinUNet gave 0.58 IoU, which shows good alignment between the predicted height and reference building footprints. Also, the IoU of binarized building height is close to IoU of predicted building footprint. This shows good consistency between the two learned objectives.

The histogram plots in Figure 5 provide insights into the building height prediction at each site, where the predicted height distribution is compared with the corresponding reference building height distribution. Both reference and predicted building heights saturates between 10 m to 15 m. In each site, there are certain building pixels with height values greater than 1m but their

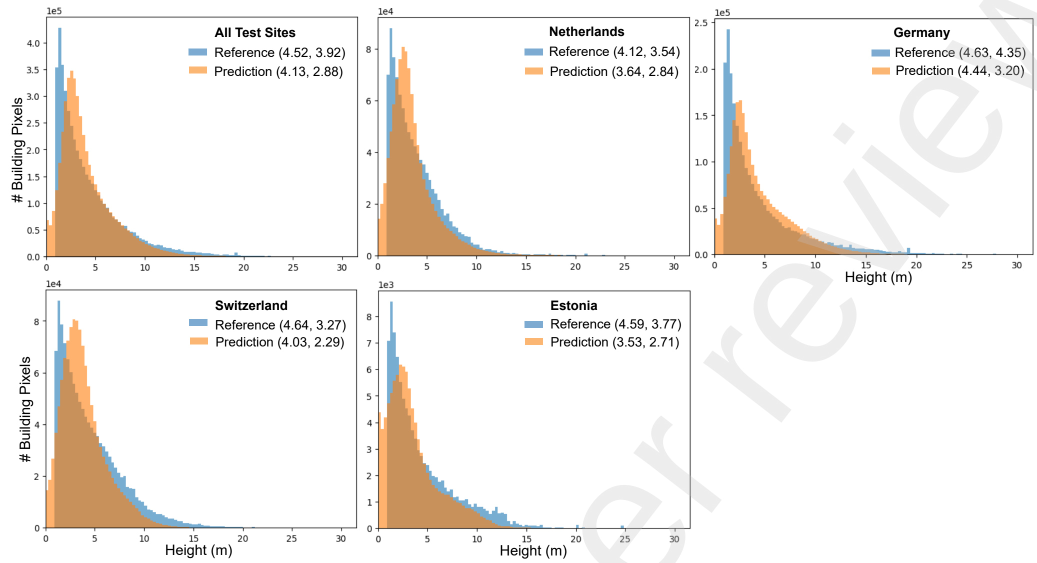

Figure 5: Sitewise histogram comparison of reference and predicted building height on the test set. The values in the legend are the mean and standard deviation values of reference and predicted building heights.

predicted building height values are between 0 and 1 m. Overall, there is a good overlap between the predicted and reference distributions.

For further evaluation, Figure 6 presents pixel-wise correlation between predicted and reference building heights. The prediction on the Netherlands test set shows the best correlation with 0.63 R^2 and 1.66 m RMSE and Switzerland shows the least correlation with 0.45 R^2 and 2.05 m RMSE. Both Germany and Switzerland's train sets have approximately the same number of building pixels. Still, the building height is better predicted in Germany which probably indicates higher complexity in learning heights in Switzerland than in Germany.

It is essential to derive individual (or instance-level) building height from the pixel-wise regressed height values, as they are more interpretable, easy to analyze and monitor. To do so, the predicted pixel-wise building heights are post-processed using reference building footprint polygons. Where, the building height values were smoothed using 70 percentile height value over each building polygon. This also improved the pixel-wise correlation between the predicted height of the building and the reference height shown in Figure

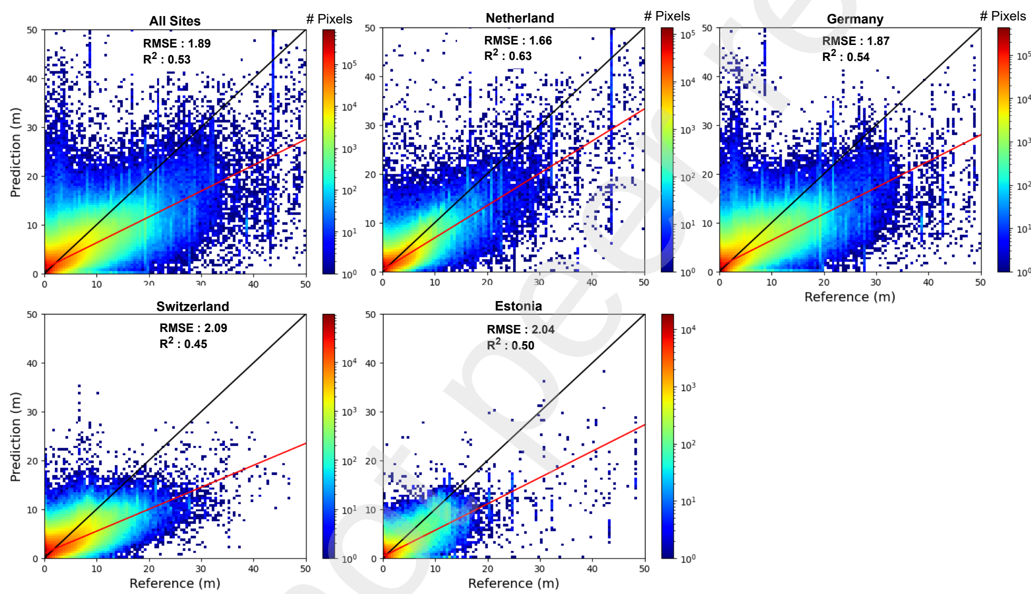

Figure 6: Correlation between predicted building height and reference building height. The first plot is on full test set while the other four plots are on individual test sites. The black diagonal plot y=x represents the best possible fit and the red line is the actual fit to the plot.

7. The overall RMSE score reduced from 1.89 m to 1.60 m and the overall R^2 improved from 0.53 to 0.66. The improvement is evidently consistent on each study site.

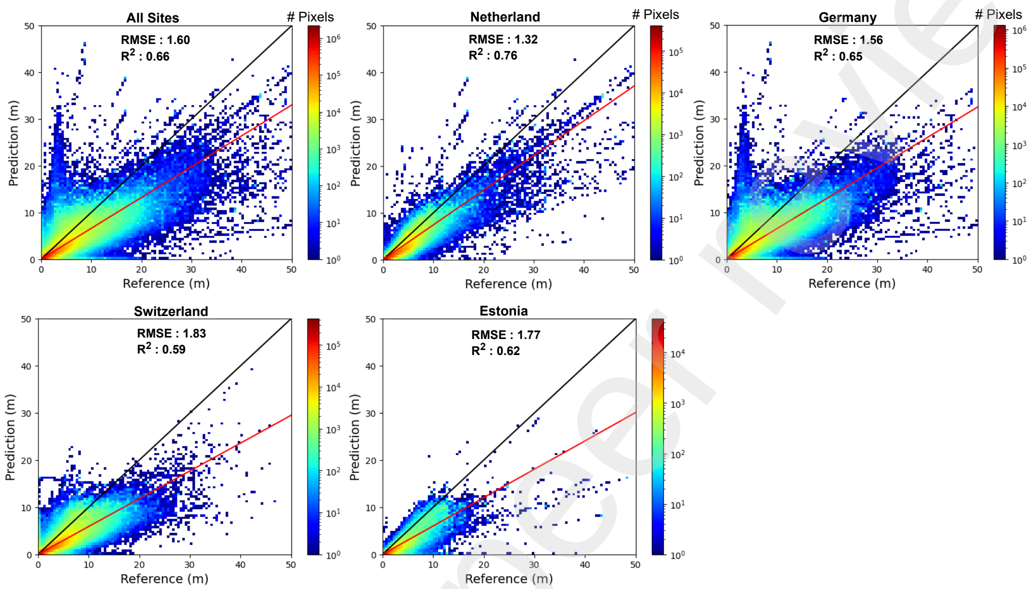

Figure 7: Instance-wise smoothed correlation between predicted and reference building height using 70 percentile of building pixels.

4.1. Comparison with other models

The performance of the proposed model is compared with four other models, a basic U-Net [Ronneberger et al., 2015] model, two recent transformer-based networks TransUNet [Chen et al., 2021] and SwinUNETR [Hatamizadeh et al., 2021] and a recent satellite time series network UTAE [Garnot and Landrieu, 2021]. To make a fair comparison, we implemented these four networks in a multitask setting. The quantitative comparison is shown in Table 3. The UNet model shows comparatively low scores. Both SwinUNETR and UTAE gave similar scores with a difference in R^2 score. The UTAE model learns from the temporal dimension resulting in 3% better R^2. The best results come from the proposed T-SwinUNet, which efficiently learns spatio-temporal features of time-series data to predict building height and footprints with at least 0.16 lower RMSE, 4.5% better R^2 and 7% better IoU score.

Table 3: Comparison with the UNet baseline and three competing models.

	RMSE (m)↓	R^2 ↑	IoU ↑
UNet	3.02	0.369	0.481
TransUnet	2.49	0.422	0.50
SwinUNETR	2.05	0.456	0.51
UTAE	2.20	0.489	0.53
T-SwinUNet	**1.89**	**0.533**	**0.58**
MBHR-Net	4.64	0.42	0.500
BHE-Net	4.21	0.397	0.518

Our proposed T-SwinUNet model is also compared with the MBHR-Net [Yadav et al., 2023] and BHE-Net [Cai et al., 2023] DL models, proposed in two recent studies to estimate building height using Sentinel-1 SAR and Sentinel-2 MSI data. Both the models are dual stream models where one stream extracts Sentinel-1 SAR features and other extracts Sentinel-2 MSI features. [Yadav et al., 2023] and [Cai et al., 2023] tested MBHR-Net and BHE-Net on small test sets from Netherlands and China respectively. We implemented these two models and trained them on our dataset as proposed by their authors. The results are given in Table 3. Compared to MBHR-Net and BHE-Net, T-SwinUNet predicts building height more accurately (atleast 2.32 lower RMSE, 11% better R^2) and predicted height show better alignment with the building footprints (atleast 6% better IoU).

4.2. Ablation

In this section, we evaluate the contribution of Multi-Task Learning (MTL), Time Series (TS) input and individual modality i.e. Sentinel-1 SAR and Sentinel-2 MSI on building height estimation results of the proposed T-SwinUNet. Quantitative ablation results are given in Table 4. To assess the impact of MTL on the performance of T-SwinUNet, the segmentation branch was removed from the decoder including the segmentation head1. The building footprints were derived by binarizing the regression output using a threshold of 1.0 meter. The results in Table 4 demonstrate that without MTL, the model yields 2% lower R^2 and slightly (0.02) high RMSE. The improved recall and reduced precision indicate false building detection and possibly the cause of the drop in R^2 value. On the other hand, T-SwinUNet trained with the complementary task of segmentation learns to avoid estimating the height of a non-building object i.e., avoid false detections.

Table 4: Ablation study to evaluate the contribution of different parts of the proposed approach on the accuracy of building height estimation. The metrics shows Building Height (BH) evaluation.

	RMSE (m)↓	R^2 ↑	Recall ↑	Precision ↑	IoU ↑	F1 ↑
T-SwinUNet	**1.89**	**0.53**	0.71	0.66	**0.58**	**0.69**
W/O MTL	1.92	0.51	0.81	0.56	0.52	0.66
1 TS	2.22	0.36	0.68	0.53	0.48	0.59
SAR	2.25	0.37	0.57	0.58	0.47	0.57
Optical	1.94	0.47	0.67	0.68	0.56	0.67

For estimating the impact of time series input, the T-SwinUNet was trained following a methodology similar to that of [Yadav et al., 2023]. where the 12 time series images of one data point are used as 12 augmented images with seasonal effects. To adapt this in our implementation, instead of feeding 12 time series images stacked as one input, only one image is given to the model. The images is selected randomly for each data sample. The results in Table 4 show that the 12 month time series input improved the performance of T-SwinUNet, reflected in all metrics. Without time series input, the building height estimation shows an accuracy drop with 17% lower R^2 and 0.33 higher RMSE. Similarly, the predicted heights show low alignment (10% drop in IoU score) with the building footprints.

To evaluate the contribution of the two modalities we trained two T-SwinUNet one with Sentinel-1 SAR time series only and the other with Sentinel-2 MSI time series. The quantitative results show that Sentinel-2 MSI input provided significantly better height estimates with good alignment with building footprint than Sentinel-1 SAR. Although Sentinel-1 SAR is positively correlated with the height of the building, the relation can become weak due to metallic surface, building density, complex tree canopy scattering and others. Sentinel-2 MSI can capture shadows and seasonal effects very well, which are important features for estimating building heights. Also, MSI data is beneficial in distinguishing different land cover surfaces. When both inputs are incorporated into the model, the results reflect further enhancement, particularly in the R^2 score, indicating the model's improved ability to precisely capture height variations.

4.3. Qualitative Evaluation

For the qualitative analysis, some samples from the four test sites are visualized in Figure 8, 9 and 10, where Figure 8 and 9 visualize building height predictions in small areas while Figure 10 provides visualization at a larger scale. These samples showcase diverse urban areas with varying building densities and architectural styles. For instance, the Netherlands and Germany samples are high-building density areas, while the Switzerland samples are medium density and Estonia samples are examples of low-building density areas. The predicted building heights capture variations in building heights (from tall to short) and demonstrate a strong correlation with reference height values. The samples show an accurate prediction of the building footprint with fine spatial details of building structures and their boundaries, ensuring fine details in building height maps. This highlights the robustness of the T-SwinUNet model in accurately estimating building heights in areas with diverse characteristics, including geographic location and architectural style.

Apart from 128 $\times$ 128 samples, the predicted building height maps are visualized on a larger scale. Figure 10 presents city-scale predicted building height maps, showcasing one city from each test site. The three cities, Groningen (Netherlands), Leipzig (Germany) and Winterthur (Switzerland) are dense cities with approximately 103482, 22472, 13254 buildings respectively whereas Tamsalu (Estonia) is a small city with only 1517 buildings. The majority of the tall buildings are towards the center of the cities, while the shorter buildings are on the outskirts. Figure 10 demonstrates that our building height predictions are accurate not only for small areas but also extend to large-scale building height mapping.

4.4. Generalizability

The motive behind the experiment is to test the generalizability of the model to another country within Europe. In this experiment, we trained our proposed T-SwinUNet model on Netherlands, Estonia and Switzerland and evaluated on test data from Germany. The results are then compared with our previous test results on Germany where the T-SwinUNet was trained on all four sites including Germany. Table 6 enlists the evaluation metric from the two settings, Figure 11 shows the predicted height distributions and Figure 12 shows two samples to qualitatively compare the height estimations in the two settings.

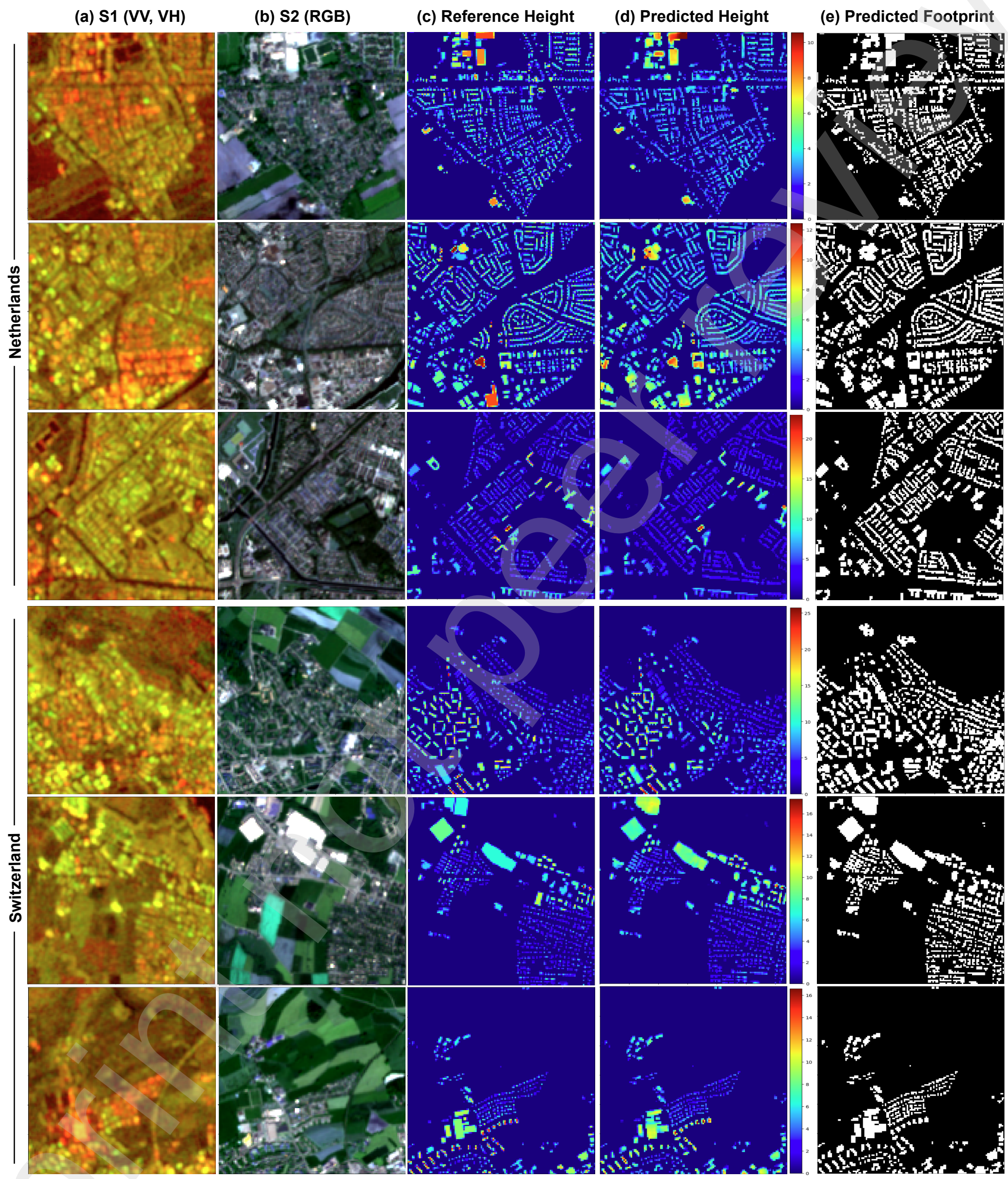

Figure 8: Qualitative comparison: Samples of building height and footprint predictions from Netherlands and Switzerland test set at 10 m resolution.

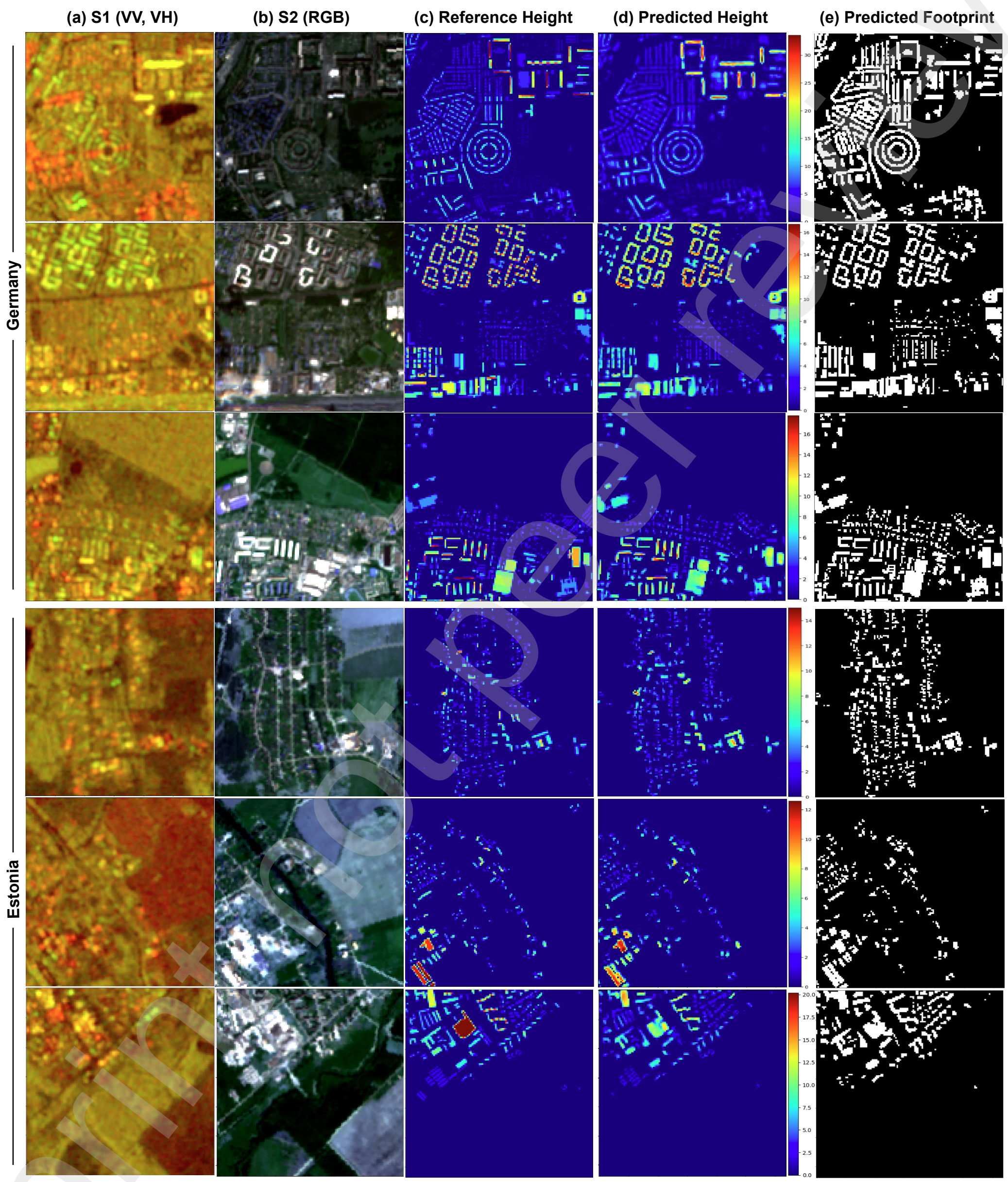

Figure 9: Qualitative comparison: Samples of building height and footprint predictions from Germany and Estonia test set at 10 m resolution.

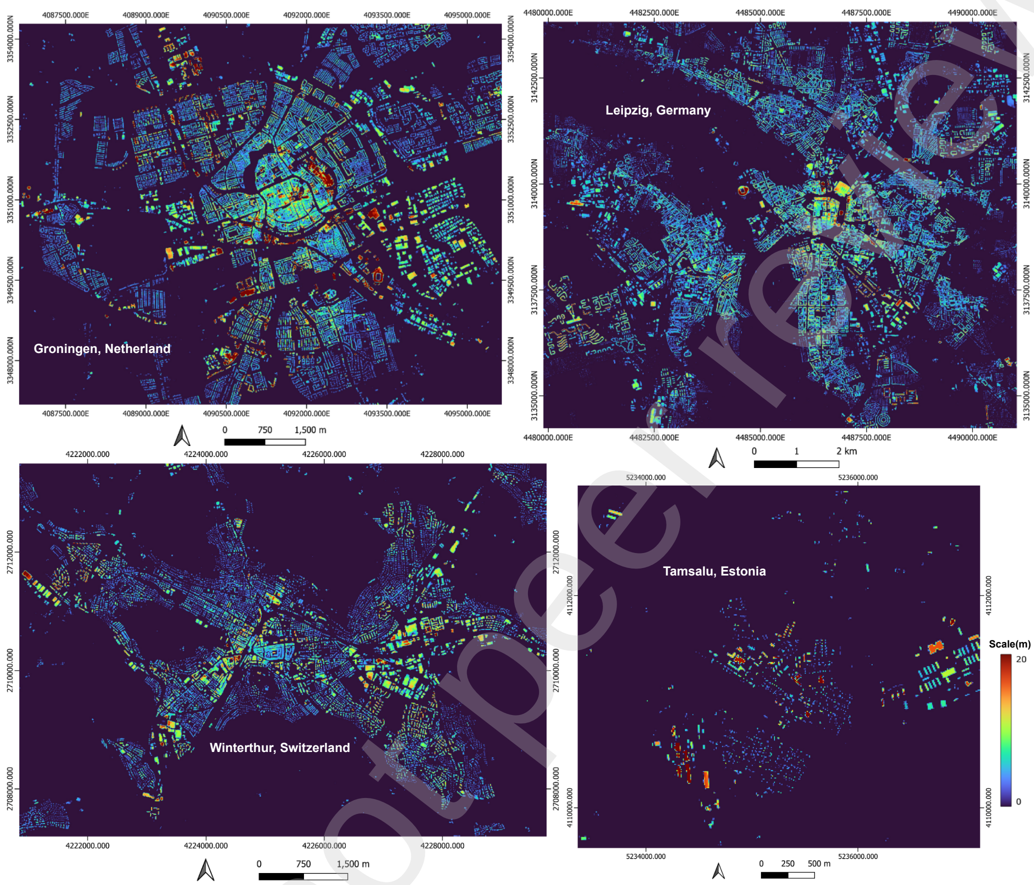

Figure 10: Building height visualizations at a larger scale, one city from each test site.

Table 5: Generalizability of T-SwinUNet on Germany test set. Compare model's performance when trained on all four sites (Trained on Full data) with it's performance when trained without Germany data (Trained W/O Germany data).

T-SwinUNet	RMSE (m)↓	R^2 ↑	IoU ↑
Trained on Full data	**1.87**	**0.54**	**0.53**
Trained W/O Germany data	2.20	0.47	0.52

When the model is not familiar with the Germany building data distribution (not trained on Germany data), the height estimation evaluation metric scores dropped showing an increase in error in the results. Both RMSE and R^2 dropped significantly while IoU score or building segmenting capability of the model remained the same. The predicted distribution in Figure 11 shows a drop in the expected peak adding both underestimations (height less than 1m) and overestimations with respect to the reference building heights. However, the mean and variance of predicted height distribution in Figure 11 (a) and Figure 11 (b) does not show a big change.

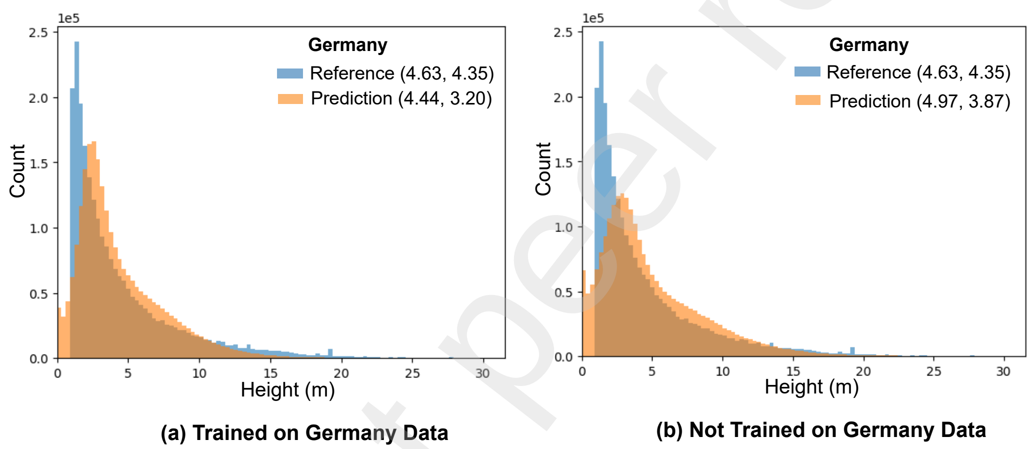

Figure 11: Histogram comparing predicted and reference heights on Germany test data. (a) Prediction by T-SwinUNet trained on all four sites (Netherlands, Estonia, Switzerland and Germany), (b) Prediction by T-SwinUNet trained on three sites (Netherlands, Estonia and Switzerland).

The samples in Figure 12 show that both the predictions have good estimations of building heights but clearly there are a few overestimations in the prediction (e) from the model not trained on Germany data. To summarize, overall we see that the model shows lower performance when it is not familiar to similar building architectures but the building heights are still estimated with good accuracy.

4.5. Comparison with GHSL-Built-H R2023A Global product at 100 m

For the Netherlands, Estonia, and Germany, our building height predictions are based on data from 2019, while for Switzerland, it is based on data

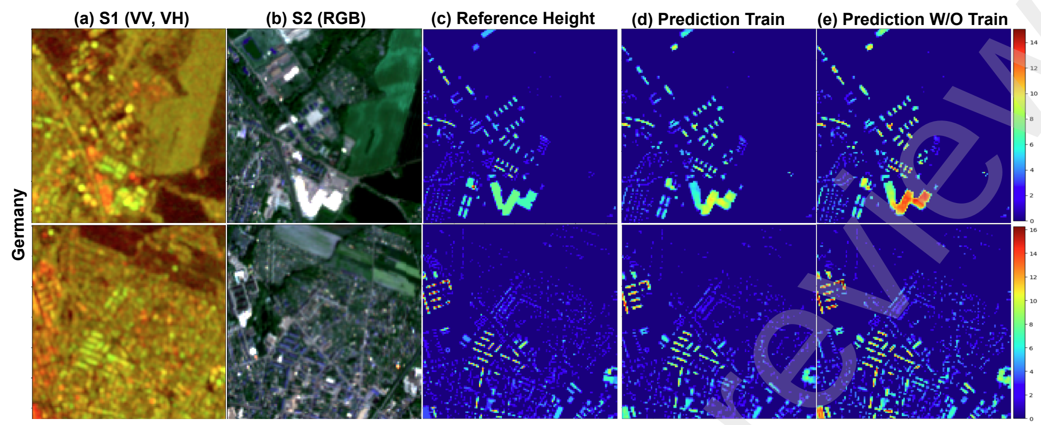

Figure 12: Result samples for qualitative comparison of building height estimation in Germany by model which is trained on Germany data (d), versus estimations by model not trained on Germany data (e).

from 2021. Despite the fact that the building heights from GHSL-Built-H R2023A are derived from 2018 data, they are still valid in 2019 for the Netherlands, Estonia and Germany due to the low rate of urban development (approximately 1%) in Europe [CBS, 2023, ELB, 2023, FIEC, 2023]. For Switzerland, the 2018 GHSL height estimates are also relatively valid in 2021, as only 4.6% of the buildings were newly constructed over a period of 5 years (2016-2021), resulting in a build-up growth of merely 2.7% between 2018 and 2021 [BFS, 2023]. To make a fair comparison, the predicted building heights by T-SwinUNet were downsampled from 10m to 100m spatial resolution using average resampling, and both quantitative and qualitative comparisons were performed.

The quantitative results at 100 m resolution are given in Table 6. The results show that our building height predictions are consistently accurate at both 10 m resolution and 100 m resolution. Compared to the GHSL-

Table 6: RMSE, R^2 and IoU over test set for proposed T-SwinUNet and GHSL-Built-H R2023A product [Pesaresi et al., 2021] at 100 m.

	RMSE (m)↓	R^2 ↑
GHSL-Built-H R2023A (100m)	0.56	0.68
T-SwinUNet (100m)	**0.33**	**0.86**

Built-H R2023A product (0.56 m RMSE and 0.36 R^2), the building height estimations from the proposed T-SwinUNet model (0.29 m RMSE and 0.73 R^2) are more accurate in terms of both RMSE and R^2 metrics. Figure 13

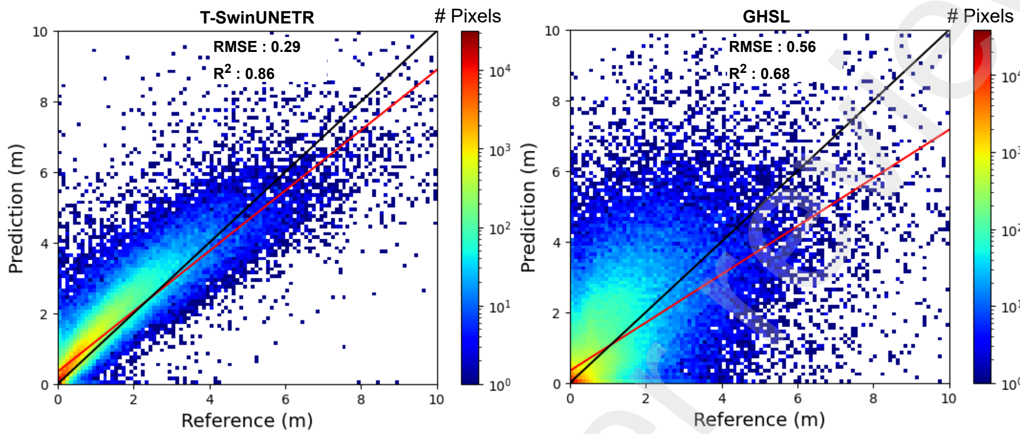

Figure 13: Building height evaluation at 100 m using correlation plots. The black diagonal plot y=x represents the best possible fit and the red line is the actual fit to the plot.

depicts similar behavior. Building height predictions from T-SwinUNet are highly correlated with the reference height values, as the plotted points are close to the diagonal (y = x), while the correlation is weak for the GHSL-Built-H R2023A product as the plot is more scattered.

For qualitative comparison, two samples from each test site are visualized in Figure 14 and 15. Similar to the evaluation at 10m resolution, the selected samples cover both low and high-building-density areas. Visualized samples indicate that the building heights from the GHSL-Built-H R2023A product frequently underestimate or overestimate the actual building height (reference values). For instance, in the first-row sample from the Netherlands test site (Figure 14) and the two samples from the Germany test site (Figure 15), the majority of building heights from the GHSL-Built-H R2023A product tend to overestimate the reference values. On the contrary, in the Swiss and Estonia test sites, the building heights are underestimated. On the other hand, the predicted building heights by the T-SwinUNet model closely approximate the reference building heights. Although there are slight discrepancies in the estimated height values, the model maintains a strong correlation of both tall and short building heights with reference values, supporting the correlation presented in Figure 13.

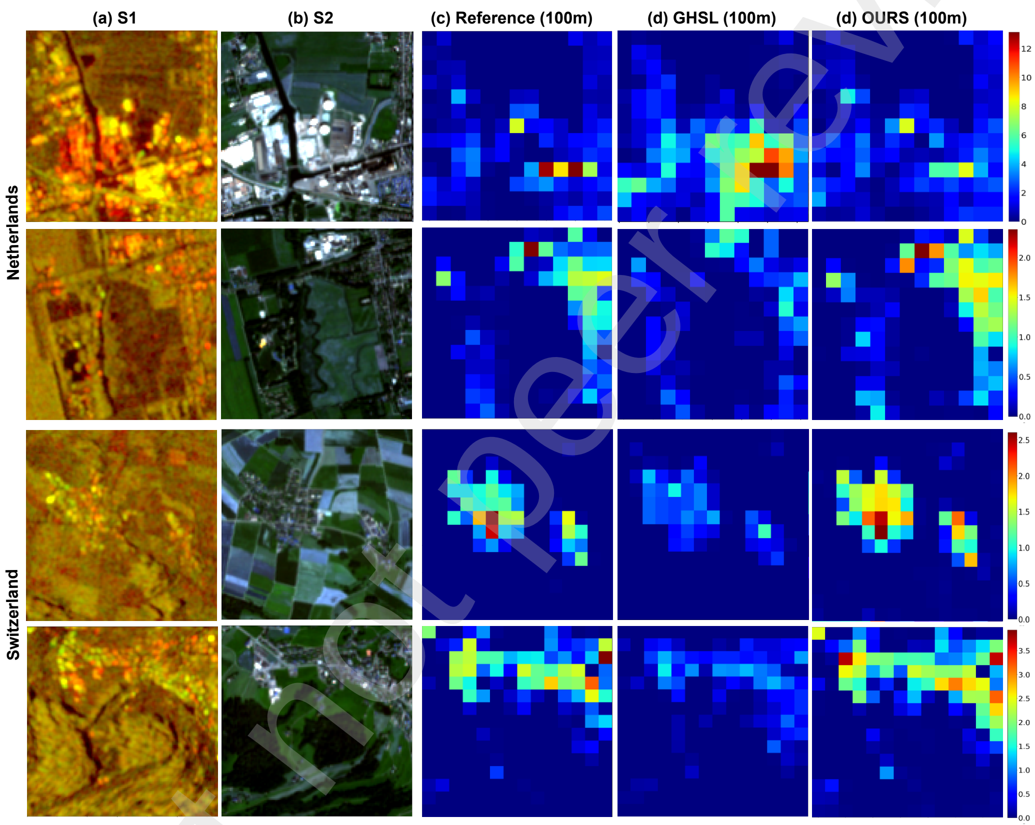

Figure 14: Qualitative comparison of building height predicted by T-SwinUNet (e) with building height from GHSL-Built-H R2023A product (d) at 100 m resolution. The samples are from the Netherlands and Switzerland test sites.

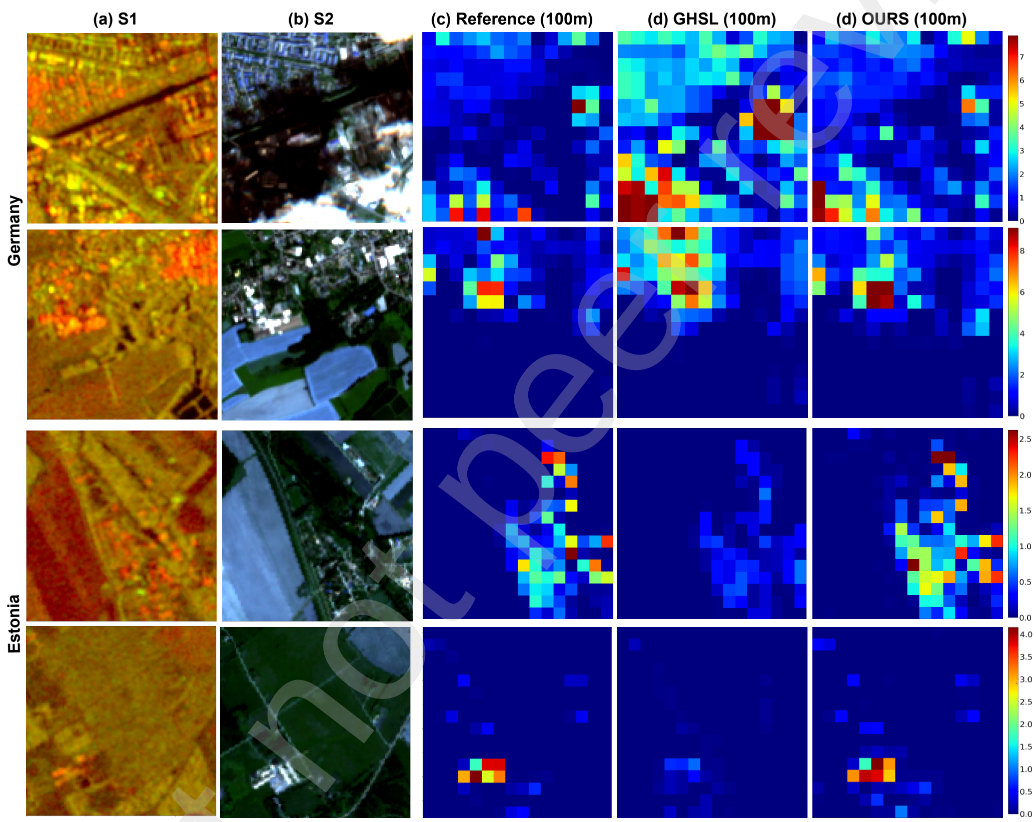

Figure 15: Qualitative comparison of building height predicted by T-SwinUNet (e) with building height from GHSL-Built-H R2023A product (d) at 100 m resolution. The samples are from Germany and Estonia test sites.

5. Conclusions

In this study, we addressed the complex challenge of building height estimation by exploring advanced DL models. Our proposed T-SwinUNet model effectively processes combined Sentinel-1 SAR and Sentinel-2 MSI time series images, providing precise building height and footprint estimations at a 10 m resolution. Comprehensive evaluations across multiple regions, including the Netherlands, Switzerland, Estonia, and specific areas of Germany, demonstrate the model's performance, achieving RMSE of 1.89 m and IoU of 0.58.

Our comprehensive analysis, including the ablation study, highlighted the contributions of various components within our proposed approach. The results emphasized the role of MTL in enhancing the model's overall performance, leading to accurate height estimations and refined building footprint delineations. Notably, the inclusion of Sentinel-1/2 temporal information through time series data significantly improved model's accuracy, enabling it to capture building shadow and height features under seasonal variations. The complementary nature of the Sentinel-1 SAR and Sentinel-2 MSI data further solidified the model's capabilities, with Sentinel-2 MSI contributing significantly to enhanced height estimates and precise footprint segmentation.

Our findings highlight the broad applicability and scalability of the propose T-SwinUNet model, as evidenced by its success in both small-scale and large-scale settings. This study has the potential for global extension and frequent height map updates, as we are using frequently and globally available free-of-cost data. Through the successful development and rigorous evaluation of our T-SwinUNet model, this study contributes significantly to the advancement of accurate and scalable building height estimation, with utilization in diverse urban development monitoring applications, ranging from regulatory assessments and disaster impact analyses to population dynamics and energy consumption evaluations.

6. Acknowledgements

This research is part of the EO-AI4GlobalChange project funded by Digital Futures.

References

A. Asokan and J. Anitha. Change detection techniques for remote sensing applications: a survey. *Earth Science Informatics*, 12(2):143–160, 2019.

BFS. Federal statistical office. 2023. (accessed 01 January 2024).

B. Cai, Z. Shao, X. Huang, X. Zhou, and S. Fang. Deep learning-based building height mapping using sentinel-1 and sentienl-2 data. *International Journal of Applied Earth Observation and Geoinformation*, 122:103399, 2023.

Y. Cao and X. Huang. A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 chinese cities. *Remote Sensing of Environment*, 264:112590, 2021.

CBS. Netherlans dwellings and non-residential stock. 2023. (accessed 01 January 2024).

J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

P. Chen, H. Huang, J. Liu, J. Wang, C. Liu, N. Zhang, M. Su, and D. Zhang. Leveraging chinese gaofen-7 imagery for high-resolution building height estimation in multiple cities. *Remote Sensing of Environment*, 298:113802, 2023a.

S. Chen, Y. Ogawa, C. Zhao, and Y. Sekimoto. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:129–152, 2023b.

C. Corbane, J.-F. Faure, N. Baghdadi, N. Villeneuve, and M. Petit. Rapid urban mapping using sar/optical imagery synergy. *Sensors*, 8(11):7125–7143, 2008.

B. Dong, Q. Zheng, Y. Lin, B. Chen, Z. Ye, C. Huang, C. Tong, S. Li, J. Deng, and K. Wang. Integrating physical model-based features and spatial contextual information to estimate building height in complex urban areas. *International Journal of Applied Earth Observation and Geoinformation*, 126:103625, 2024.

ELB. Estonian land board geo3d. 2023. (accessed 01 January 2024).

T. Esch, E. Brzoska, S. Dech, B. Leutner, D. Palacios-Lopez, A. Metz-Marconcini, M. Marconcini, A. Roth, and J. Zeidler. World settlement footprint 3d-a first three-dimensional survey of the global building stock. *Remote sensing of environment*, 2022.

FIEC. Germany new construction report. 2023. (accessed 01 January 2024).

D. Frantz, F. Schug, A. Okujeni, C. Navacchi, W. Wagner, S. van der Linden, and P. Hostert. National-scale mapping of building height using sentinel-1 and sentinel-2 time series. *Remote Sensing of Environment*, 2021.

V. S. F. Garnot and L. Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6*, pages 171–181. Springer, 2020.

V. S. F. Garnot and L. Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

S. Hafner, Y. Ban, and A. Nascetti. Unsupervised domain adaptation for global urban extraction using sentinel-1 sar and sentinel-2 msi data. *Remote Sensing of Environment*, 2022.

A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

A. Hu, L. Wu, S. Chen, Y. Xu, H. Wang, and Z. Xie. Boundary shape-preserving model for building mapping from high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

H. Huang, P. Chen, X. Xu, C. Liu, J. Wang, C. Liu, N. Clinton, and P. Gong. Estimating building height in china from alos aw3d30. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185:146–157, 2022a.

X. Huang, Y. Cao, and J. Li. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. *Remote Sensing of Environment*, 244:111802, 2020.

X. Huang, J. Yang, W. Wang, and Z. Liu. Mapping 10 m global impervious surface area (gisa-10m) using multi-source geospatial data. *Earth System Science Data*, 14(8):3649–3672, 2022b.

K. Koppel, K. Zalite, K. Voormansik, and T. Jagdhuber. Sensitivity of sentinel-1 backscatter to characteristics of buildings. *International Journal of Remote Sensing*, 38(22):6298–6318, 2017.

T. Leichtle, T. Lakes, X. X. Zhu, and H. Taubenboeck. Has dongying developed to a ghost city?-evidence from multi-temporal population estimation based on vhr remote sensing and census counts. *Computers, Environment and Urban Systems*, 78:101372, 2019.

H. Li, Q. Li, G. Wu, J. Chen, and S. Liang. The impacts of building orientation on polarimetric orientation angle estimation and model-based decomposition for multilook polarimetric sar data in urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 54(9):5520–5532, 2016.

M. Li, E. Koks, H. Taubenböck, and J. van Vliet. Continental-scale mapping and analysis of 3d building structure. *Remote Sensing of Environment*, 2020a.

X. Li, P. Gong, Y. Zhou, J. Wang, Y. Bai, B. Chen, T. Hu, Y. Xiao, B. Xu, J. Yang, et al. Mapping global urban boundaries from the global artificial impervious area (gaia) data. *Environmental Research Letters*, 15(9): 094044, 2020b.

X. Li, Y. Zhou, P. Gong, K. C. Seto, and N. Clinton. Developing a method to estimate building height from sentinel-1 data. *Remote Sensing of Environment*, 2020c.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017.

W. Liu, X. Sun, W. Zhang, Z. Guo, and K. Fu. Associatively segmenting semantics and estimating height from monocular remote-sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

M. Marconcini, A. Metz-Marconcini, S. Üreyen, D. Palacios-Lopez, W. Hanke, F. Bachofer, J. Zeidler, T. Esch, N. Gorelick, A. Kakarla, et al. Outlining where humans live, the world settlement footprint 2015. *Scientific Data*, 7(1):242, 2020.

M. Marconcini, A. Metz-Marconcini, T. Esch, and N. Gorelick. Understanding current trends in global urbanisation-the world settlement footprint suite. *GI_Forum*, 2021.

M. Pesaresi, C. Corbane, C. Ren, and N. Edward. Generalized vertical components of built-up areas from global digital elevation models by multi-scale linear regression modelling. *Plos one*, 16(2):e0244478, 2021.

M. Recla and M. Schmitt. Deep-learning-based single-image height reconstruction from very-high-resolution sar intensity data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:496–509, 2022.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

C. Sudre, W. Li, T. Vercauteren, S. Ourselin, and J. C. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*. 2017.

M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

UN. The sustainable development goals report 2022. 2022. (accessed 01 January 2024).

W.-B. Wu, Z.-W. Yu, J. Ma, and B. Zhao. Quantifying the influence of 2d and 3d urban morphology on the thermal environment across climatic zones. *Landscape and Urban Planning*, 226:104499, 2022.

W.-B. Wu, J. Ma, E. Banzhaf, M. E. Meadows, Z.-W. Yu, F.-X. Guo, D. Sengupta, X.-X. Cai, and B. Zhao. A first chinese building height estimate at 10 m resolution (cnbh-10 m) using multi-source earth observations and machine learning. *Remote Sensing of Environment*, 2023.

C. Xi, C. Ren, J. Wang, Z. Feng, and S.-J. Cao. Impacts of urban-scale building height diversity on urban climates: A case study of nanjing, china. *Energy and Buildings*, 251:111350, 2021.

R. Yadav, A. Nascetti, and Y. Ban. Building change detection using multi-temporal airborne lidar data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B3-2022:1377–1383, 2022.

R. Yadav, A. Nascetti, and Y. Ban. A cnn regression model to estimate buildings height maps using sentinel-1 sar and sentinel-2 msi time series. *arXiv preprint arXiv:2307.01378*, 2023.

P. Yan, F. He, Y. Yang, and F. Hu. Semi-supervised representation learning for remote sensing image classification based on generative adversarial networks. *IEEE Access*, 8:54135–54144, 2020.

W. Zhou, S. Newsam, C. Li, and Z. Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018.