A Dual-Branch Deep Learning model for Building Height Estimation using Sentinel-1 and Sentinel-2 data

Jenna Mansueto
Stanford University
mansueto@stanford.edu

Spencer Paul
Stanford University
spaul@stanford.edu

Duke Qiu
Stanford University
jiahuiq@stanford.edu

Abstract

Mapping building height is essential to understanding urban development, with important applications in urban planning, sustainability, hazard prevention, and public health. This paper proposes a new deep learning model, Dual-Branch Vision Transformer (Dual-ViT) based on UNet architecture for accurate building height estimation from Sentinel-1 SAR (S1) and Sentinel-2 Optical (S2) data. The model utilizes dual-branch structures to process distinct data streams, trained on Texas S1/S2 datasets using Microsoft Building Footprints as ground truth. We achieved a significant accuracy with an RMSE of 1.078m for all pixels and 3.823m for building-only pixels. Notably, by incorporating building footprint data and adopting a novel building loss function, we reduced the RMSE for building pixels by 55.4% compared to comprehensive loss. Tested for domain adaptation, the Dual-ViT model demonstrated consistent performance across varied regions, including Arizona, Pennsylvania, and Los Angeles, CA, evidencing its robustness and transferability. This correlation of model accuracy with building density points to future work in optimizing loss function architecture and advancing the model's generalizability for diverse urban scenarios.

1. Introduction

Urban development both shapes our environment and reflects characteristics of the populations which inhabit it. Building height is a key feature of these urban systems which can crucially inform our understanding of aspects of their functioning. As a measure of the scale of a structure, building height has been shown to be predictive of building energy consumption and green house gas emissions [14] [3], and has been used as a feature in population-prediction applications [2]. To this end, it can be a useful source of information in urban planning, and even improve our understanding of natural hazard risk and the spread of infectious diseases.

Despite its usefulness in a host of applications, building height data is not readily available globally. There are two main sources of building height data, each with their own challenges: cadastral data and light detection and ranging-based (LiDAR) methods. Often provided by city governments, cadastral data contains building height and footprint information and is generally very accurate. However, it is not publicly available around the world. LiDAR is a laser-based remote sensing technique which can be used to estimate building height with a high degree of precision, but it is cost prohibitive to use on a large scale.

In 2022, Microsoft released a global building footprint dataset, and though it currently contains height information for only areas of the United States and Europe, it contains footprint data for 999M buildings globally [13]. Given the global coverage of this footprint data, we explore a potential model to predict building heights in regions where this footprint data is available, opening the door to inexpensive building height coverage around the world. For scenarios where footprint data isn't yet available –as well as for literature benchmarking purposes– we continue to explore models which do not incorporate footprint information.

Given the value of building height data to applications in urban planning, sustainability, and public health, it is important that this information be universally accessible. In this paper, we propose a deep learning model for estimating building height from publicly available Sentinel-1 and Sentinel-2 imagery, as well as the incorporation of building footprint data.

2. Related Work

There exists a body of literature which focuses on estimating building height from Sentinel-1 and Sentinel-2 imagery – the majority of these studie were conducted in the past five years in China and Europe. Earlier studies within this window generally use random forest models or other ensemble methods, while more recent work has shown the efficacy of deep learning methods for this task.

Previous work emphasizes the importance of using

Sentinel-1 and Sentinel-2 time series imagery in combination to estimate building height [12]. Frantz et al. [7] compared performance for a Support Vector Regression model trained on optical-only (S2), radar-only (S1)and both optical and radar data sources, and found that the latter model significantly outperformed both models trained on only one data source. This study used S1 and S2 features derived from Sentinel-1/2 time series data– a common practice in the literature [4] [16] as it allows for easier segmentation of built up density from the surrounding environment.

Existing work underscores the superiority of deep learning methods to traditional machine learning techniques for the task of building height estimation. Cao and Huang [5] demonstrated that a deep network outperforms a random forest on this task, though they used multi-view, high resolution satellite imagery which is costly for large-scale height estimation.

Of the previously explored deep learning methods, dual-branch architectures appear the most promising for use with Sentinel-1 and Sentinel-2 data. Yadav et al. [16] employed a U-Net architecture with separate ResNet50 encoders for the multispectral and SAR backscatter input from Sentinel-1 (S1) and Sentinel-2 (S2), respectively. This model was trained and tested on 10 cities in the Netherlands, and achieved a promising RMSE of 3.73m. Cai et al. [4] also made use of a dual-branch U-Net-based architecture, creating separate encoders and decoders for S1 and S2 features. A custom feature fusion module is employed after the final encoder layer, giving each branch access to multi-modal feature information. This model was tested on urban areas in China with an RMSE of 4.65m– impressive given the high built up density in the study area.

The majority of previous studies have been conducted in China and Europe, and we found no existing work that applies deep learning for this task in the United States. Li et al. [11] developed an indicator based on Sentinel-1 backscatter data to model building heights in the United States at 500m resolution. Li et al. [10] used a random forest model to predict building heights in the US, China and Europe at 1km resolution. Neither study employed deep learning methods, and both predicted building height at course resolution.

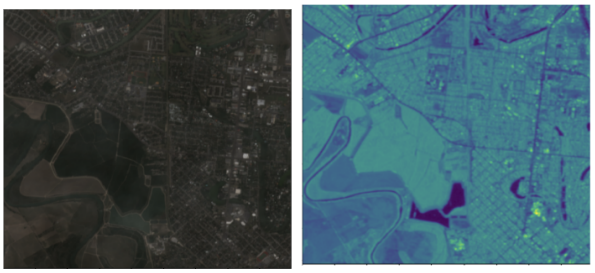
To address this gap in the literature, we propose a deep learning method to estimate building height in the US at 10m spatial resolution.

3. Methodology

3.1. Data

The data used for this study came from various satellite sources, Sentinel-1 Ground Range Detected (GRD) and Sentinel-2, collected between 2017 and 2021. Reference building height and footprint data was obtained from Microsoft's Building Footprint data set [13]. For building associated pixels, our dataset had a mean of 5.3 m and a standard deviation of 2.3 m. For building and non-building pixels, our dataset had a mean value of 0.34 m and a standard deviation of 1.5m.

Sentinel-1 offers free and global SAR data at 10m spacial resolution. The the data is not a single snap-shot, but the aggregation of all recorded readings in the 2017 to 2021 time frame. We extracted per-pixel annual mean of backscatter intensities in VV and VH polarization along with percentile values. The percentile values saved where 10, 25, 75, and 90 with the hope that this would capture a temporal dimension of the data by examining the pixel aggregates when images were both lighter and darker. We hypothesized that this would map to season like winter or summer. Sentinel-2 provides optical imagery at 10m resolution as well. We obtained the same mean and percentile values as our S1 data for RGB bands. We captured the annual percentile pixel and mean pixel values because prior research emphasized the importance of the temporal dimension of data in building height prediction [16]. We randomly



(a) Mean RGB Bands (b) Mean VH Band

(c) Mean VV Band (d) Building Height Mask

Figure 1. Visualization of aligned data bands

sampled 3,000 4.5 km x 4.5 km grid cells across Texas using google earth engine API [1] and paired them with the their corresponding footprint (see figure 1). To ensure significant building representation in our data, we filtered out samples that contained less than 2,000 building pixels. As a result, our data set was dominated by major cities in Texas (Dallas, Houston, San Antonio and Austin).

3.2. Preprocessing

In our study, we chose to crop the images down further to 128x128 before feeding them into our model. This decision

was guided by several practical considerations. Firstly, reducing the image size helps in managing the computational load, allowing the model to train faster and more efficiently . Secondly, there were minor dimensional inconsistencies in our original sampled grid images.

A key decision was made regarding the selection of spectral bands from the available multi-spectral data. After thorough experimentation with various combinations of bands, detailed in the experiment section, we opted to exclusively utilize the mean values of the red, blue, green, VV (vertical transmit and vertical receive polarization), and VH (vertical transmit and horizontal receive polarization) bands. This brought our overall data set size to 128x128x5 images paired with a 128x128x1 building height mask. The data was then split using 80-10-10 train, validation and test split.

3.3. Model Architecture

In Figure 2, we present the architecture of our novel dual-branch ViT-Net, which is meticulously designed to concurrently learn from multispectral imagery of Sentinel-2 (S2) and the Synthetic Aperture Radar (SAR) backscatter characteristics of Sentinel-1 (S1) images. The architecture consists of two distinct pathways: the S1 branch, which processes two input images representing the mean values of VV and VH polarizations, and the S2 branch, dedicated to extracting features from three input images capturing the mean values in the Red, Green, and Blue (RGB) multispectral bands.

The conceptual foundation of our model architecture is rooted in the U-Net framework, originally formulated for medical imaging applications. U-Net's proficiency in segmentation tasks is well-documented and widely acknowledged in the literature [15], serving as an instrumental baseline for our model's design. In this paper, we design our architecture based on a popular segmentation library for easy deployment and adopt various advanced encoders [9]. U-Net's architecture encompasses a dual process: an encoder that distills the image into its most salient features, and a decoder that subsequently upscales these condensed features to generate an output mirroring the input's dimensions. Prior studies [4] [8] [16] have underscored the efficacy of U-Net in generating building height estimation masks. These studies also highlight the benefits of segregating SAR and multispectral data into separate encoders and then fusing their outputs, a strategy that has proven to enhance performance significantly. Building upon this concept, our model adopts a dual-encoder structure. However, diverging from the traditional ResNet backbone, we integrate the more contemporary Vision Transformer (ViT) as our encoder foundation. The adoption of ViT, known for its remarkable success in surpassing conventional convolutional network benchmarks [6], offers a significant enhancement in performance and establishes a new frontier in remote sensing image analysis.

3.4. Loss function

In our model, we employ the Mean Squared Error (MSE) as the loss function, which is a standard choice for regression problems. MSE effectively quantifies the difference between the predicted and reference building heights. It is defined as follows:

$$L_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(H_r - H_p)^2 \tag{1}$$

Here H_r denotes the reference building heights, H_p the predicted heights, and n is the number of observation pixels.

To address the prevalent inconsistencies in benchmarking methodologies within the literature, we adopted a dual-approach training strategy for our model. This involved training two distinct versions:

Building Loss: In the first approach, our loss function was calculated exclusively for pixels within building footprints. This entailed filtering out all pixels with a value of zero with the help of building footprints, thereby focusing solely on the building areas. This version presents a more challenging task as it concentrates on the complex dynamics of building structures without the inclusion of non-building pixels, which could otherwise moderate the mean squared error.

Comprehensive Loss: The second version extends the loss evaluation to encompass all pixels, irrespective of whether they fall within building footprints. This approach not only considers the accuracy in height estimation but also implicitly incorporates building segmentation into the problem, thereby providing a more holistic assessment of model performance.

This dual-strategy allows for a more nuanced understanding of the model's capabilities in building height estimation and segmentation tasks.

3.5. Experimental Strategy

Our experiments were conducted using PyTorch Lightning on an NVIDIA TESLA P4 GPU within a Google Cloud VM. We set the learning rate at 0.0001 and the batch size to 4, utilizing the Adam optimizer for training. An early stopping mechanism was implemented with a patience of 10 epochs.

4. Results

4.1. Performance on Comprehensive Loss

4.1.1 Architecture and Backbone Experimentation

In this paper, we evaluate a novel dual-encoder architecture featuring Vision Transformer (ViT) backbones, contrasting it with traditional backbones like ResNet50 and Vgg16.

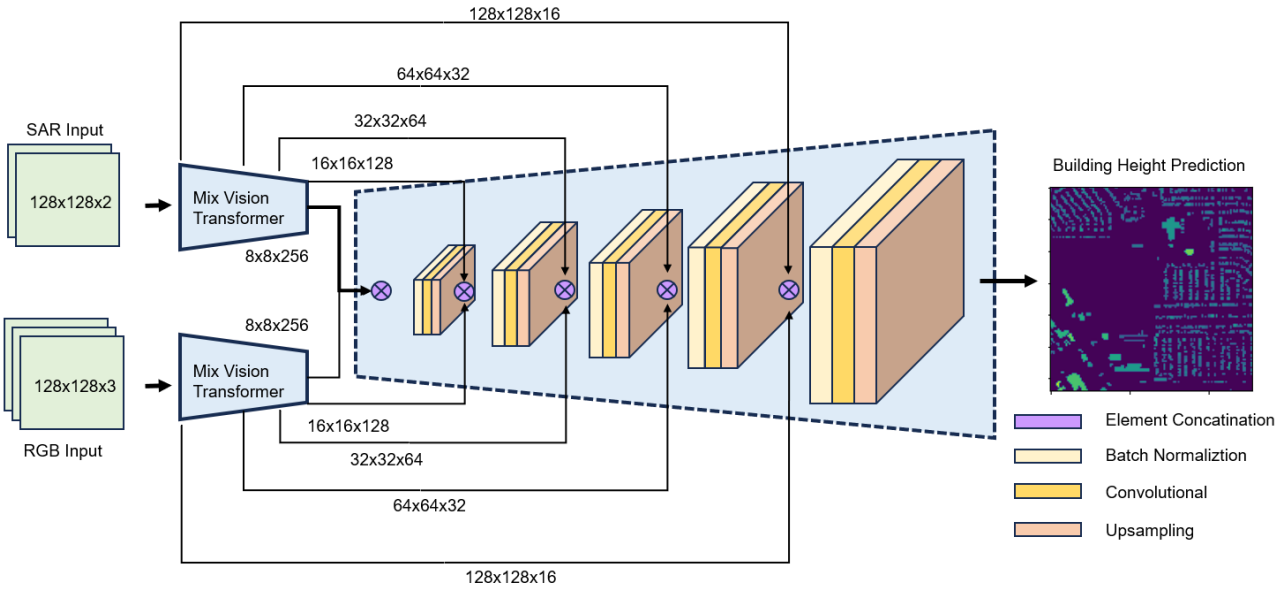


Figure 2. Architecture of our proposed model [this is temporary diagram we need to do one for our specific model once we have ironed out final details].

Additionally, we integrate a top-performing model derived from a single UNet architecture with Vgg16 for comparison. Our experiments utilize a preprocessed dataset from Texas, consisting of 2,744 samples, to assess performance. The results, detailed in Table.1, reveal that Dual-ViT has achieved the best accuracy in both all pixels and building-only pixels scenarios. While the accuracy differences are marginal for all pixels, largely due to a high proportion of zero values in the dataset, the improvement in RMSE for building-only pixels is notably significant. In particular, Dual-ViT exhibits a significant improvement, with a 7.8% reduction in RMSE compared to the Dual-ResNet50 model—the latter being the architecture employed by Yadav et al. in their 2023 study [16].

	Dual ViT	Dual ResNet50	Dual Vgg16	Single Vgg16
RMSE (All)	1.078	1.080	1.084	1.097
RMSE (Bldg)	3.823	4.123	4.227	3.968

Table 1. RMSE of all pixels / building-only pixels with different encoder selections.

4.1.2 Ablation Study

In our study, we conduct ablation experiments on the Dual-ViT model to ascertain the individual contributions of Sentinel-1 (S1) and Sentinel-2 (S2) components. For a fair comparison, we feed the same inputs from each source separately into both branches of the model. From Table.2, S1 & S2 have different contribution with different scenarios. Specifically, S2, as an optical source, is more effective in delineating boundaries and reducing overall errors when assessing all pixels. In contrast, when focusing solely on building pixels, S1 data, which provides richer altitude information through SAR, outperforms S2. Overall, the model achieves the RMSE in both scenarios when it incorporates data from both S1 and S2. This outcome underscores the synergistic and complementary nature of SAR and optical images, suggesting that their integration is advantageous for enhancing model learning.

	S1+S1	S2+S2	S1+S2
RMSE (All)	1.152	1.126	1.078
RMSE (Bldg)	4.233	4.489	3.823

Table 2. RMSE of all pixels / building-only with different input using Dual-Vit model.

4.2. Performance on Building Loss

4.2.1 Architecture and Backbone Experimentation

Similar to the approach detailed in Section 4.1.1, we also experiment with various backbones and architecture based on building loss, as shown in Table.3. The results indicate that the Dual-VGG16 model outperforms others, al-

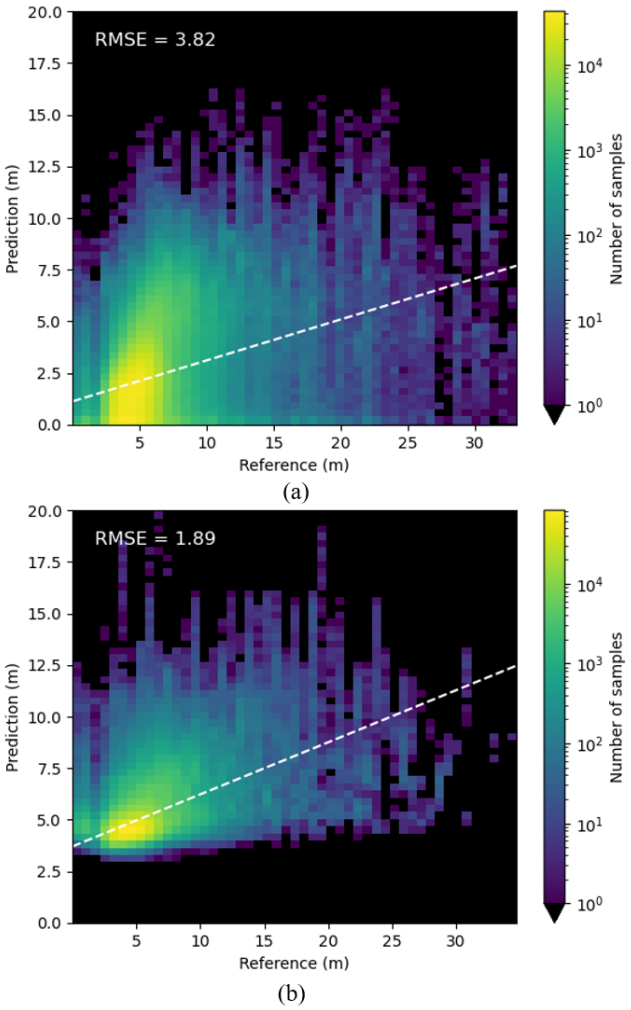


Figure 3. The scatterplot between referenced building heights and predicted heights for building-only pixel task where while dash light represents linear regression of overall prediction: (a) Comprehensive loss; (b) Building Loss

beit with nuanced differences in structure. Notably, using building loss for training leads to a considerable enhancement in performance, evidenced by a 55.4% reduction in RMSE for building-only pixels, compared to the best RMSE achieved with comprehensive loss. This underscores the value of building loss in specialized applications where building footprint data can be leveraged for improved accuracy.

4.2.2 Comparison: Comprehensive and Building Loss for Building Pixels

As discussed in Section.4.2.1, for the task that only evaluates the building pixels, using the building loss will result in better prediction accuracy compared to the comprehensive loss. This section furthers the analysis to visualize the actual difference with the same Dual-ViT architecture.

	Dual ViT	Dual ResNet50	Dual Vgg16	Single Vgg16
RMSE (Bldg)	1.790	1.733	1.697	1.706

Table 3. RMSE of building-only pixels with different encoder selections.

The scatterplot in Fig.3 illustrates the performance of building height predictions using two different loss functions: comprehensive loss and building loss. The plot for comprehensive loss indicates considerable variability in the predictions, largely influenced by the inclusion of zero values, which correspond to non-building pixels. This has led to a tendency for the model to underpredict building heights, as evidenced by a dense cluster of data points around the 5m mark on the reference heights, with the predictions averaging around 2.5m.

In contrast, the building loss scatterplot demonstrates a more focused grouping of predictions around the 5m reference height, suggesting a higher accuracy in height prediction when non-building pixels are not a factor. Notably, there's a distinct cutoff in the predictions at approximately 3m, which is a positive indicator for the model's utility in building-related tasks. This sign aligns with domain knowledge since the majority of buildings exceed one story, typically around 3m in height, suggesting that the model is effectively distinguishing between building and non-building elements and can reliably identify structures that are at least one story tall.

In Fig.4, we showcase two illustrative examples to qualitatively evaluate the model's performance. Following the exclusion of non-building pixels, the model utilizing building loss clearly shows a marked improvement in predicting building heights, both in the accuracy of location and the predicted values, as compared to the model trained with comprehensive loss. The latter struggles to differentiate between small buildings and surrounding terrain without the aid of building footprints. Despite these advancements, both models exhibit limitations in accurately predicting the heights of unusually tall structures, which can be attributed to the insufficient resolution of 10-meter imagery that fails to provide the necessary detail.

4.3. Domain Adaptation

To validate the robustness of our model, we tested our model, trained on imagery from Texas, on three independent domains. These three datasets were preprocessed using the same pipeline as our training data, and contained Sentinel-1/2 imagergy from the states of Arizona and Pennsylvania, and the city of Los Angeles, CA (LA). Arizona

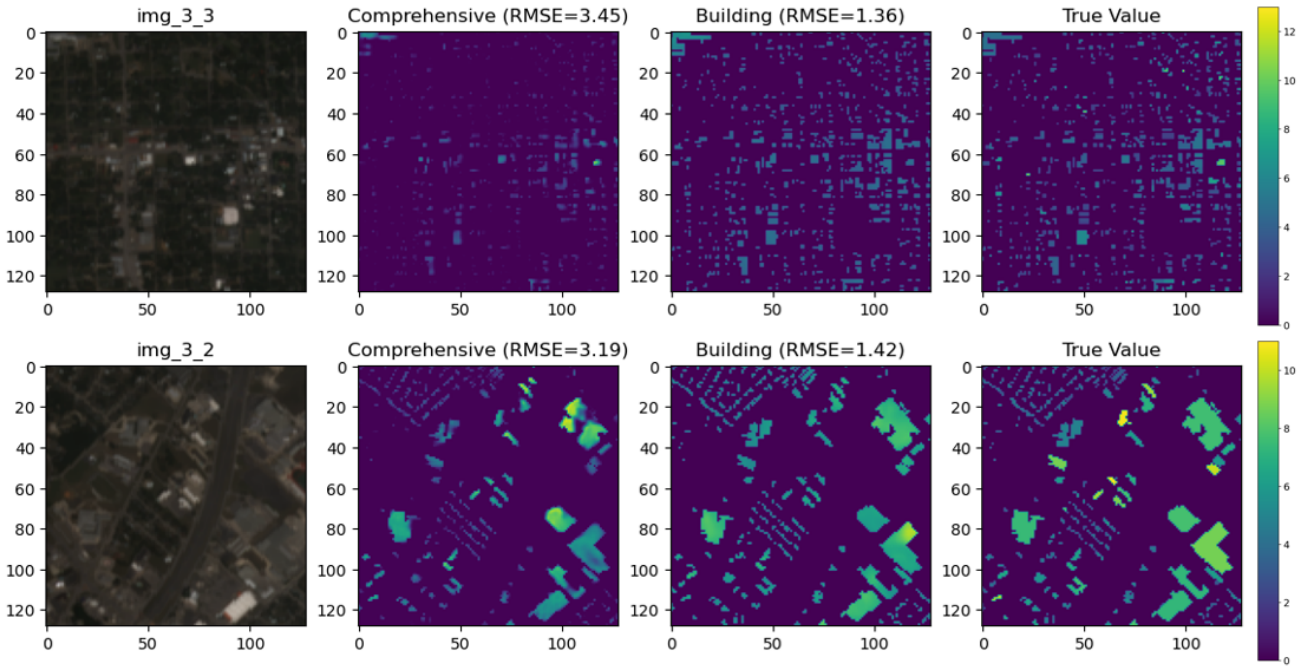


Figure 4. Two Result Samples for building pixels task and non-building pixels has been filtered with building footprint masks. From left to right, Optical Image, Comprehensive Loss Prediction, Building Loss Prediction, Referenced Building Height

and Pennsylvania have different terrain than Texas, and so these datasets served to test our model's ability to generalize to new geographies. With more tall buildings than Houston or San Antonio, Los Angeles is a significantly larger city than those in Texas, and thus was intended to test the model's performance in an unfamiliar and challenging environment. To explore domain adaptation potential across downstream applications, we tested models trained on both comprehensive and building loss on these independent domains.

4.3.1 Domain Adaptation with Comprehensive Loss

Using a Texas-trained dual-ViT model with comprehensive loss and the previously defined hyperparameters, we found the results shown in Table 4.

Overall, the model appears to generalize well to environments with unfamiliar terrain – in Arizona and Pennsylvania, it achieved all-pixel RMSE on par with its validation loss in Texas. The differences in RMSE between these states may mostly be attributable to the relative frequency of building pixels in the datasets.

To that end, the high all-pixel RMSE observed on the Los Angeles dataset is likely due to the high density of buildings, as well as the increased presence of taller buildings. This is supported by the building-pixel RMSE results, where the model shows a smaller decrease in performance on LA. Domain transfer experiments in the literature [7] reflect the challenge of testing on independent domains with higher proportions of tall buildings, but emphasize the importance of including such domains in training.

RMSE/Domain	TX	AZ	PA	LA
RMSE (All)	1.079	1.182	1.020	1.798
RMSE (Building)	3.823	3.671	3.585	4.239

Table 4. RMSE of dual-ViT model tested on independent domains.

4.3.2 Domain Adaptation with Building Loss

To explore the robustness of our building loss-trained models, we tested our two best performing building loss models on these independent domains, achieving the results displayed in Table 5. We observe that the single-VGG16 model outperforms the dual-VGG16 model on two out of three independent domains, achieving markedly better performance on the Arizona dataset in particular. Once again, we do not believe that unfamiliar terrain has affected model performance, as variations in RMSE can be accounted for by variation in building height distributions.

In Figure 5, we plot predicted versus ground truth values for single and dual-VGG16 models tested on Arizona. The single-VGG16 model generates a smaller range of pre-

Backbone/Domain	TX	AZ	PA	LA
Dual-VGG16	1.6973	2.1243	1.5926	1.8447
Single-VGG16	1.7061	1.7935	1.5382	1.9019

Table 5. Building-pixel RMSE of models tested on independent domains.

dictions, likely allowing it to outperform the dual-encoder model on this dataset dominated by buildings less than 15m in height. This is further supported by the dual-VGG16's slightly superior performance on the LA dataset which contains more tall buildings.

While we feel confident in the ability of both comprehensive and building loss trained models to generalize to unfamiliar terrain in the United States, further work could be done to improve the models' ability to adapt to domains with varied building density and height distributions.

5. Conclusion

In this study, we have developed the Dual-ViT model, utilizing UNet architecture for the estimation of building heights, incorporating both Sentinel-1 SAR and Sentinel-2 Optical data. The model demonstrates consistent prediction performance with RMSEs of 1.078m for all pixels and 3.823m for building-only pixels tasks. Another significant advancement in our study is the adoption of a building loss function, which utilizes building footprints, resulting in a 55.4% RMSE reduction for building-only pixels compared to the previously used comprehensive loss. Additionally, the proposed model has been tested for domain adaptation, showing promising robustness in this new application. Our findings also indicate a strong correlation between prediction accuracy and building density. Future work will focus on optimizing the architecture to enhance building loss application and creating a more universally applicable model to adeptly handle a variety of scenarios.

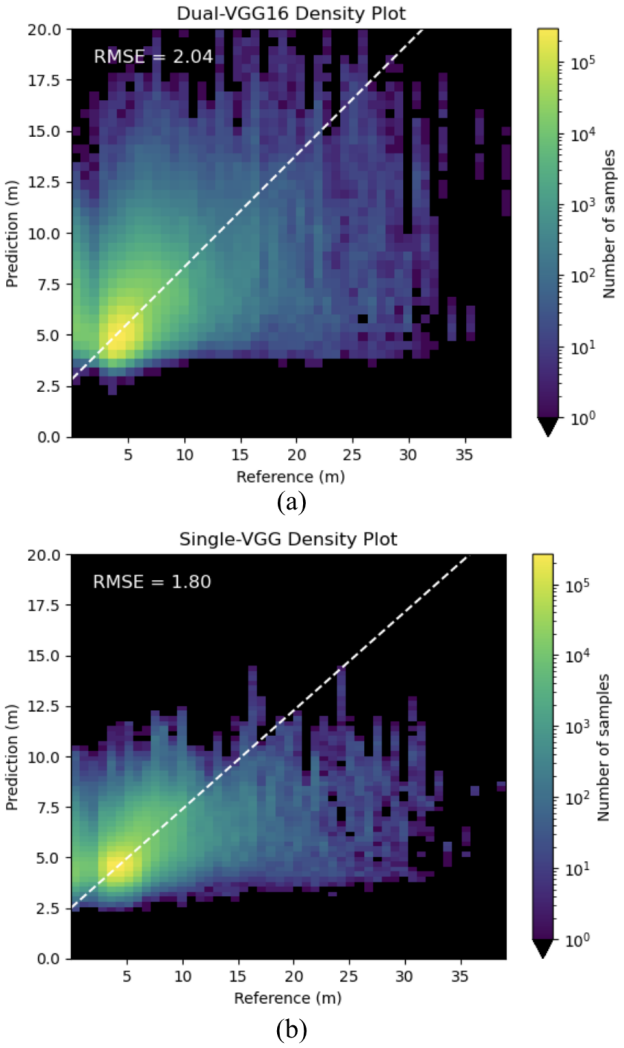


Figure 5. The scatterplot between referenced building heights and predicted heights for building-only pixel task in Arizona: (a) Dual-VGG16 backbone; (b) Single-VGG16 backbone.

References

[1] Google earth engine api. https://earthengine.google.com/. Accessed: 2023. 2

[2] Mohammed Alahmadi, Peter Atkinson, and David Martin. Estimating the spatial distribution of the population of riyadh, saudi arabia using remotely sensed built land cover and height data. *Computers, Environment and Urban Systems*, 41:167–176, 2013. 1

[3] Rainald Borck. Will skyscrapers save the planet? building height limits and urban greenhouse gas emissions. *Regional Science and Urban Economics*, 58:13–25, 2016. 1

[4] Bowen Cai, Zhenfeng Shao, Xiao Huang, Xuechao Zhou, and Shenghui Fang. Deep learning-based building height mapping using sentinel-1 and sentinel-2 data. *International Journal of Applied Earth Observation and Geoinformation*, 122:103399, 2023. 2, 3

[5] Yinxia Cao and Xin 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. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 3

[7] David Frantz, Franz Schug, Akpona Okujeni, Claudio Navacchi, Wolfgang Wagner, Sebastian van der Linden, and Patrick Hostert. National-scale mapping of building height using sentinel-1 and sentinel-2 time series. *Remote Sensing*

of Environment, 252:112128, 2021. 2, 6

[8] Sebastian Hafner, Yifang Ban, and Andrea Nascetti. Unsupervised domain adaptation for global urban extraction using sentinel-1 sar and sentinel-2 msi data. *Remote Sensing of Environment*, 280:113192, 2022. 3

[9] Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019. 3

[10] Mengmeng Li, Elco Koks, Hannes Taubenböck, and Jasper van Vliet. Continental-scale mapping and analysis of 3d building structure. *Remote Sensing of Environment*, 245:111859, 2020. 2

[11] Xuecao Li, Yuyu Zhou, Peng Gong, Karen C. Seto, and Nicholas Clinton. Developing a method to estimate building height from sentinel-1 data. *Remote Sensing of Environment*, 240:111705, 2020. 2

[12] Xiao Ma, Guang Zheng, Xu Chi, Long Yang, Qiang Geng, Jiarui Li, and Yifan Qiao. Mapping fine-scale building heights in urban agglomeration with spaceborne lidar. *Remote Sensing of Environment*, 285:113392, 2023. 2

[13] Microsoft. Global ml building footprints. https://github.com/microsoft/GlobalMLBuildingFootprints, 2023. 1, 2

[14] Eirik Resch, Rolf André Bohne, Trond Kvamsdal, and Jardar Lohne. Impact of urban density and building height on energy use in cities. *Energy Procedia*, 96:800–814, 2016. 1

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3

[16] Ritu Yadav, Andrea Nascetti, and Yifang Ban. A cnn regression model to estimate buildings height maps using sentinel-1 sar and sentinel-2 msi time series, 2023. 2, 3, 4