Deep Learning Approach to Floor Area and Building Material Stocks Estimation Using Aerial & Street View Image

Akihiro Okuyama[a]

[a]*Princeton University, Department of Civil and Environmental Engineering*

Abstract

Urban centers contribute substantially to global greenhouse gas emissions, and with ongoing urbanization, the demand for construction materials is set to rise. This paper addresses the challenge of quantifying building material stocks (MS) in urban landscapes, a critical step in mitigating the environmental footprint of urban development. Traditional methods of estimating MS often falter due to the lack of granular building data. We propose a novel solution by employing deep learning to derive MS estimates from readily available aerial and street-view imagery. Our methodology involves the development of two deep learning models that adeptly classify building types and predict floor areas, respectively. The models demonstrate exceptional performance, with building type classification accuracy reaching 84.71% and floor area predictions achieving a mere 1.86% error. These predictions facilitate an MS estimation of concrete and total building materials with errors as low as 2.07% and 0.29%, respectively. The successful application of these models illustrates a scalable and effective approach to MS estimation, thereby aiding numerous cities in planning for a sustainable, circular economy where conventional methods are impractical.

Keywords – Remote Sensing, Google Street View, Deep Learning, Computer Vision, Material stocks

∗Corresponding author

✉ ao3526@princeton.edu (A. Okuyama)

1 INTRODUCTION

The construction and building sector significantly impacts the environment, responsible for 36% of the world's energy consumption and 39% of its CO_2 emissions, as highlighted by (Abergel, Dean, and Dulac 2017). Annually, the global production of concrete and cement reaches 17.7 Gt and 4.1 Gt, respectively, making cement the second most utilized substance after water (Monteiro, Miller, and Horvath 2017) and resulting in the emission of 3.1 Gt of CO_2. Cement production alone contributes to 9-10% of worldwide CO_2 emissions related to energy (Cao et al. 2021), a figure that could escalate to 26% of all anthropogenic CO_2 emissions by 2050 if the current production methods persist (Beyond Zero Emissions 2017). The challenge in reducing the carbon footprint of cement lies in its production process, particularly the high-temperature (up to 1450 °C) calcination process, which converts $CaCO_3$ to CaO, accounting for 60-65% of the CO_2 emissions during production (Antunes et al. 2022).

Intergovernmental Panel on Climate Change (IPCC) reports that urban areas are responsible for 67-72% of global greenhouse gas emissions in 2020 (Abergel, Dean, and Dulac 2017). Moreover, the IPCC projects that by 2050, urban areas will expand by up to 211% relative to their size in 2015, inevitably boosting the demand for construction materials. Therefore, there is a pressing need to measure and manage cement consumption effectively. As urbanization accelerates, cities face the dual challenge of meeting infrastructural demand while adhering to global sustainability targets. Accurate estimation of cement consumption can enable policymakers, urban planners, and industry stakeholders to make informed decisions to achieve sustainable urban development.

The buildings and infrastructures in cities are essential component of cities to support human life. Due to the urbanization of the recent century, MS per capita has increased by 23-fold over the 20th century globally (Krausmann et al. 2017). In the last century, more than half of the resources are used for urban expansion and their infrastructure renewal. Furthermore, the global speed of accumulation of material stocks will be faster in the future. We have accumulated 600Gt of building MS from 1970 to 2010, but the study estimates to add 800Gt to stock from 2010 to 2030 (Fishman, Schandl, and Tanikawa 2016). By quantifying the building MS, we can assess the environmental impact associated with extraction, processing, use, and disposal of these materials. This helps to calculate carbon footprint of buildings and identifies opportunities to reduce greenhouse gas emissions through more sustainable alternative materials and construction practice. Understanding MS helps city planners to practice better resource management and promote a circular economy.

A retrospective understanding of building MS and consumption in cities is essential to reducing the amount of cement we will consume in the future. Many researchers have used MS analysis to estimate building MS in cities worldwide (Lanau and Liu 2020; Mao et al. 2020; Göswein et al. 2019). In the MS analysis, the MS is calculated by multiplying floor area and material intensity, representing the amount of material used per floor area. Two approaches are widely used to estimate material use and their embodied environmental impact in cities: top-down and bottom-up (Tanikawa, Guo, and Fishman 2022). The top-down approaches utilize materials flow statistics, while bottom-up approaches use inventory data of end-use objects such as floor area to determine the material's environmental impact in cities (Reyna and Chester 2015; Fishman et al. 2014). These conventional methods highly depend on statistical or inventory data, which are not widely available across cities.

Emerging remote sensing technology and machine learning techniques can overcome this limitation. With the recent development of remote sensing technologies in recent years, high-resolution satellite and aerial imagery are available across the globe. Advancements in remote sensing technologies and computer vision techniques have informed many features of cities (Hipp et al. 2022; Gebru et al. 2017; Jean et al. 2016; Yeh et al. 2020). For example, satellite and aerial images are used to map socioeconomic inequalities and human activity volumes (Abitbol and Karsai 2020; Xing et al. 2020). In addition to remote sensing images, Google Street View images are also used to understand building features. Some research used street view images to identify the façade material of buildings and count external features of buildings (Raghu, Bucher, and De Wolf 2023; Arbabi et al. 2022). However, using computer vision through remote sensing and street view imagery to estimate building MS is limited. Since this image data is widely available across cities, it can advance MS estimation with machine learning technology.

Floor area is a crucial metric for assessing building MS, as material intensity calculations are typically based on this measure. Traditionally, regional-scale floor area estimation relied on top-down and bottom-up survey approaches (Arehart et al. 2021). The top-down method calculates floor area per capita at the national level and then downscales this to specific regions. Conversely, the bottom-up approach starts with estimating floor area per capita in smaller regions and aggregates this data to larger scales. These conventional methods, however, often overlook unused buildings and rely on a limited dataset from surveys. Recently, there has been a shift towards integrating machine learning and remote sensing technologies to develop more accurate methodologies for estimating urban floor areas (Barbour et al. 2019; X. Zhang et al. 2019; Ji and Tang 2020; 2022; Liu et al. 2021). Nonetheless, many of these innovative approaches have yet to achieve an accuracy level that is practical for real-world application. Additionally, while there

have been advancements in estimating the footprint and height of individual buildings with considerable accuracy, as demonstrated by Microsoft using remote sensing data (Microsoft), the development of a robust methodology for precise floor area estimation of each building remains an area needing further refinement. In this study, we develop two machine learning models. One classifies the building types, and the other predicts the floor area of individual buildings (parcels). These predicted variables and material intensity data enable us to estimate the building MS of each building. Then we aggregate them to estimate city-wide building MS by using aerial and street view imagery.

The quantification of material usage in buildings is significantly influenced by the type of building, specifically single-family houses (SFHs), multi-family houses (MFHs), and non-residential houses (Dodoo 2019; Soonsawad, Martinez, and Schandl 2022; Mollaei, Ibrahim, and Habib 2021). Acknowledging the building type in individual structures is vital for accurate estimations of building material usage within urban contexts. Overlooking this aspect can lead to substantial discrepancies between estimated figures and actual data. While previous city-scale studies with conventional approach of MS might have included this differentiation (Lanau and Liu 2020), obtaining detailed information about building types poses a challenge, as such data is not uniformly accessible across different cities. For instance, several studies employing machine learning and remote sensing to predict urban MS have not accounted for building type variations (Bao et al. 2023). Although previous research has successfully used machine learning models to classify land usage, such as residential versus non-residential areas, from remote sensing data (Zhao et al. 2023; Zhou et al. 2023; Bergado, Persello, and Stein 2020; Huang, Zhao, and Song 2018; Li and Stein 2020; X. Zhang et al. 2019; C. Zhang et al. 2018; Zhou et al. 2020), there is a notable lack in models that differentiate SFHs from MFHs using remote sensing imagery.

Addressing this gap, our study introduces a deep learning model capable of distinguishing between SFHs and MFHs using remote sensing and street view images. This innovation aims to enhance the precision of material use estimations at the city scale.

This study pioneers the integration of machine learning, street view, and remote sensing technologies to classify residential building types, accurately estimate floor area, and assess building MS at the individual house level. By employing advanced machine learning models, we significantly enhance the precision of estimating building MS, surpassing the estimations provided by prior research. Specifically, one of our deep learning models adeptly differentiates between SFHs and MFHs using both aerial and street view imagery. Concurrently, another model forecasts the building floor area for each tax parcel. Subsequently, we calculate MS based on the predicted building types and floor area. Leveraging computer vision techniques alongside remote sensing and street view data, our methodology refines the comprehension of material usage in the construction sector. Aerial imagery elucidates the geographical footprint of structures, while street view images furnish detailed insights into building heights and façade materials. Consequently, our approach offers a more accurate and reliable assessment of building MS and their environmental repercussions. Our proposed model is capable of executing parcel-level MS estimations using publicly accessible imagery datasets. This model empowers urban planners, policymakers, and researchers to proactively foresee and tackle the challenges associated with sustainable urban expansion, thereby facilitating the development of greener and more resilient urban landscapes. This research highlights the critical role of innovative data utilization amidst escalating urbanization while demonstrating the transformative potential of deep learning in reforming traditional practices within the construction industry.

2 METHODOLOGY

Our study area is Saint Paul (St. Paul), the capital city in Minnesota, U.S. St. Paul holds a distinguished place in the American Midwest. It is adjacent to its twin city, Minneapolis; St. Paul forms one-half of the Twin Cities metropolitan area, the largest conurbation in the state. The area of the city is 145.5 km^2 and hosts 307,193 population in 2021. St. Paul is a metropolitan city expected to experience stable development in the future, with less than a 10% population increase from 2018 to 2040. The city launched the city development plan, Climate Action and Resilience Plan, in 2019, which aims to achieve carbon neutrality in city operations by 2030 and citywide by 2050. This plan includes a goal of reducing carbon emissions by 50% by 2030 with the effort to reduce carbon emissions across every sector in the city. The construction sector is one of the key sectors to achieve this goal and its consequent opportunities for urban mining, circular economy, and waste management. The finding of our research in St. Paul can have practical applications to other cities in the U.S. and even globally that face similar situations beyond the region itself.

This study analyzes the in-use and non-use buildings in all 84,770 tax parcels in St. Paul. The sample in this study includes both residential and non-residential buildings. The Minnesota government provides the building floor area, building type, and other relative information for each tax parcel (Minnesota Geospatial Information). We collect this tax parcel data for 2021 and use this data to create the ground truth of our deep learning model. We develop two deep learning models in this study. One classifies building type and another predicts floor area. Both models use two remote sensing imagery data: aerial images and street view images. We use the NAIP (National Agriculture Imagery Program) image as the aerial image and the Google Street View

image for corresponding parcel areas as our input data (United States Geological Survey 2017; Google Map Platform). We obtained the NAIP image of St. Paul taken in 2021 from Google Earth Engine and the most recent street view image by using Google Maps Static API. Our deep learning model has dual input channels, which process two distinct types of images (aerial and street view) to extract relevant features for predicting building type and floor area change. We used the NAIP image to make the deep learning model learn the building patterns and their footprint from the sky. We employed Street view to let our model learn features from the ground-level information.

Our research employed a Convolutional Neural Network (CNN) as the foundational technology for our deep learning framework (Figure 1 & 2). CNNs, a class of deep neural networks, are adept at processing data, such as images, with a grid-like topology. This capability stems from their unique architecture, which involves the use of learnable weights and biases to assign significance to different aspects or objects within an image, thereby facilitating the distinction between diverse visual elements. The design of CNNs draws inspiration from the human brain's visual processing system, incorporating convolutional layers that sift through input data to identify useful information and pooling layers that streamline this information by reducing its dimensionality. This approach enables the effective recognition and classification of images. The utilization of machine learning and deep learning methodologies proves to be exceedingly efficient in managing the vast datasets commonly encountered in urban and geographical research. The automated feature extraction and processing capabilities of machine learning and CNNs significantly diminish the time and resources required, offering a considerable improvement over conventional technique. CNNs, in particular, are celebrated for their proficiency in recognition tasks and their ability to discern critical features from images autonomously. Another notable benefit of these models is their scalability; once trained, they can be adapted for use across different

geographical contexts or scaled up to cover more extensive areas. Opting to refine a pre-training classification model, such as ResNet-18 (He et al. 2015), through hyperparameter adjustments using our specific dataset has been demonstrated to be more efficacious than constructing a new deep learning model from the ground up.

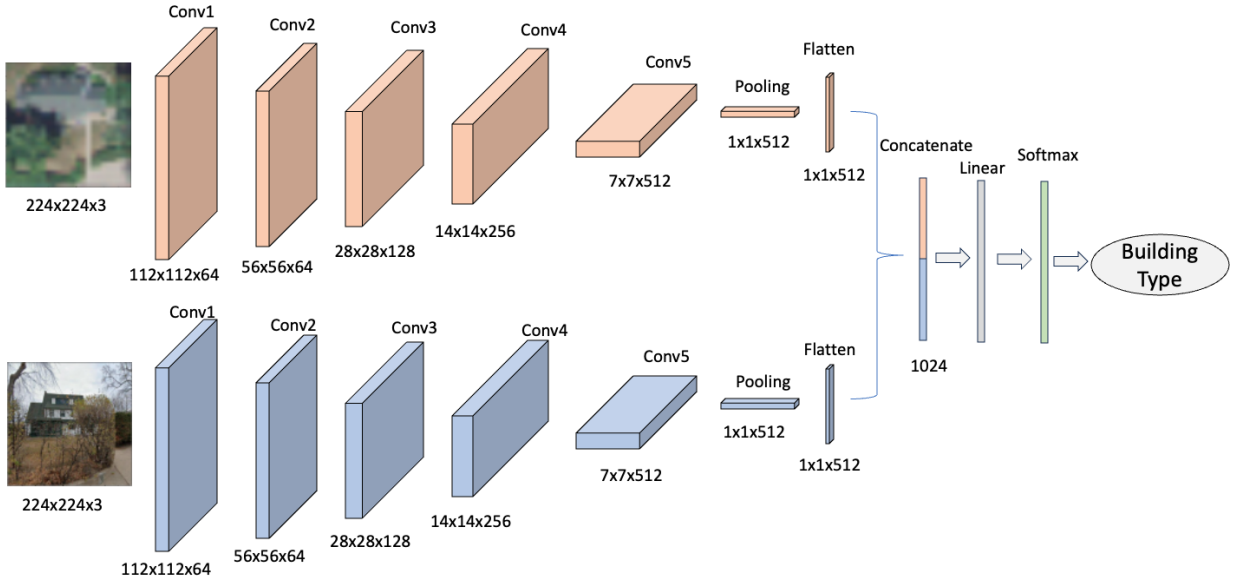

Figure 1: Architecture of the Convolutional Neural Network for Building Type Classification

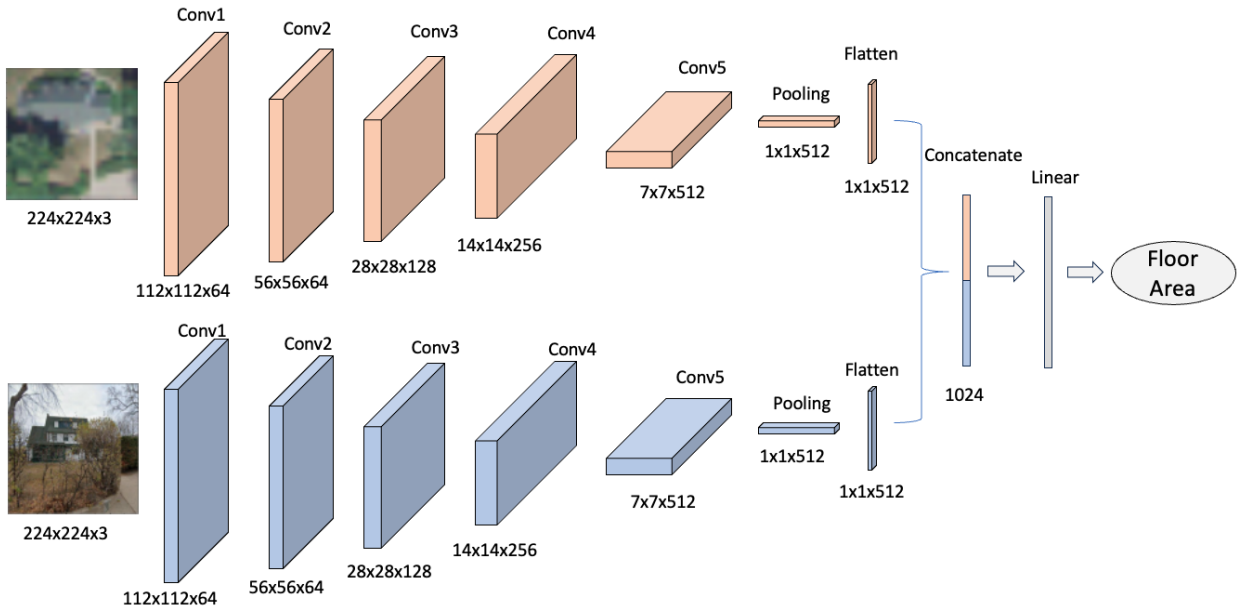

Figure 2: Architecture of the Convolutional Neural Network for Floor Area Prediction

To enhance our deep learning framework, we also integrated Vision Transformers (ViT), a novel approach that applies the transformer architecture, previously successful in natural language processing, to image recognition tasks (Dosovitskiy et al. 2021) (Figure 3 & 4). Unlike CNNs, which process images through local convolutions, ViTs divide an image into patches and treat these patches as sequences, similar to words in a sentence. This methodology allows ViTs to capture global dependencies within an image, providing a comprehensive understanding of the visual context beyond what local convolutions can achieve. ViTs start by linearly embedding each of the image patches, followed by the addition of positional embeddings to retain the order of the patches. The transformer then processes this sequence of embeddings through self-attention mechanisms, which enable the model to weigh the importance of each patch relative to others for a given task. This ability to focus on relevant parts of an image dynamically is a key advantage of ViTs, allowing for more nuanced and context-aware interpretations of visual data. The incorporation of ViTs into our framework complements the strengths of CNNs by providing a different perspective on image analysis. While CNNs excel at capturing local patterns and textures, ViTs offer superior capabilities in understanding the global structure and relationships within images. This combination ensures a robust and versatile model capable of handling a wide range of visual recognition tasks with high accuracy.

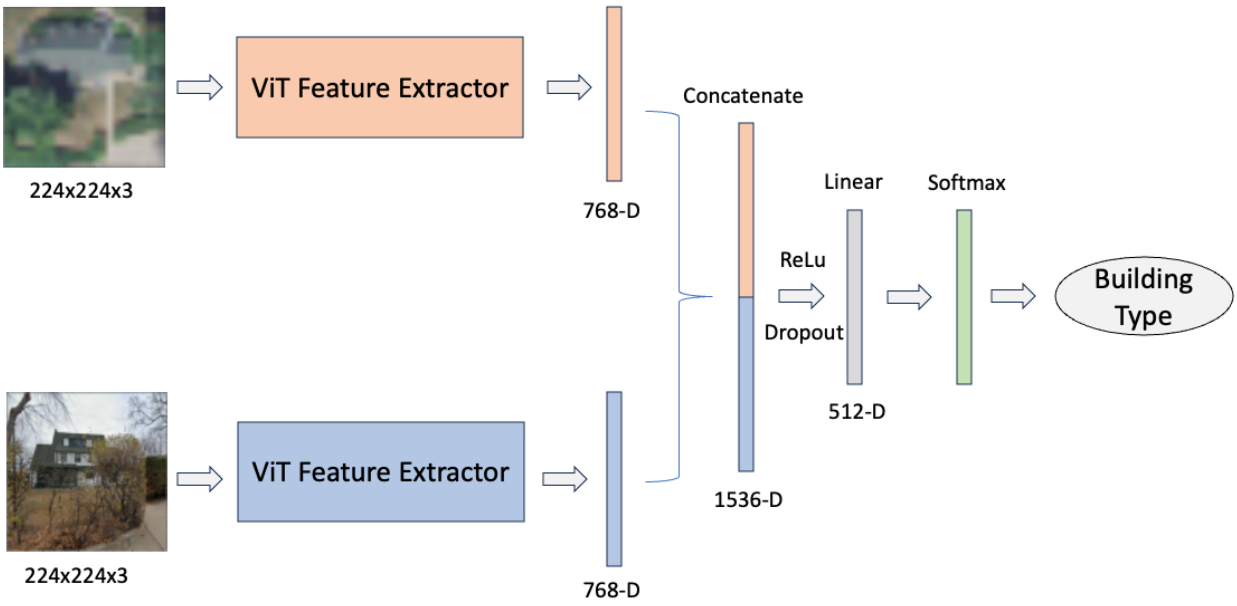

Figure 3: Architecture of the Vision Transformers for Building Type Classification. 768-D indicates 768-dimensional feature.

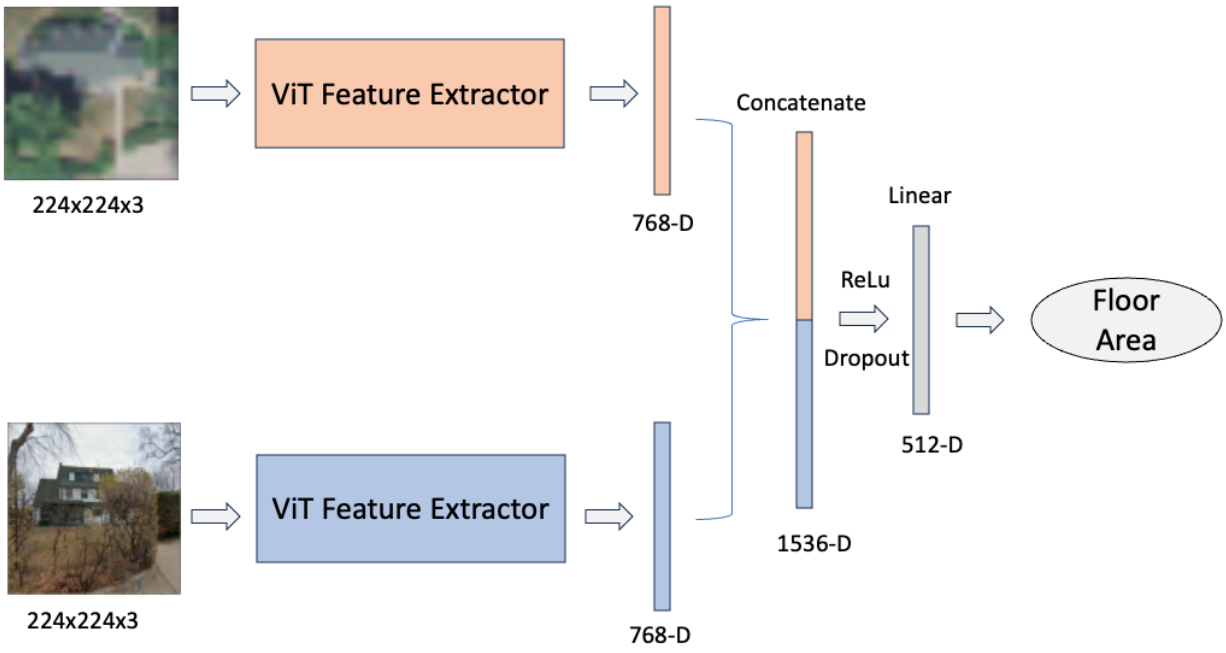

Figure 4: Architecture of the Vision Transformers for Floor Area Prediction. 768-D indicates 768-dimensional feature

Our studies compared two deep learning method with different sample splits ratio. Then we calculated the MS by using the predictive model with highest accuracy. The training data was augmented using techniques such as cropping, flipping, and rotation to improve generalization. Our building type classification model was trained by using cross-entropy loss and the floor area prediction model was trained by mean squared error. The models used stochastic gradient descent to optimize weights and an early stopping method to prevent overfitting. We used three different ratio to split the data into training, validation, and test, 60:20:20, 80:10:10, and 90:5:5. Once we trained the model, we calculated the accuracy with test dataset.

2.1 Building Type Classification Model

This paper develops a deep learning model to classify building types from remote sensing data. The original GIS-based tax parcel data in St. Paul includes information on building type with 34 categories. Based on original building type information, we grouped all buildings into three categories: SFHs, MFHs, and non-residential buildings (Table A 1). Our sample data comprises 57,924 SFHs, 16,951 MFHs, and 9,895 non-residential buildings. Since our data is imbalanced, we split the sample into training, validation, and test sets using the stratified split method to ensure the distribution of building type categories is the same for all sets. We calculated the accuracy, recall, precision, and F1 score of test data to see the reliability of the created model. Precision, recall, and F1 score are finer-grained ideas of how well the model can classify targets than just looking at overall accuracy. Precision shows how many positive predictions are correct. Recall, also called sensitivity, measures how many of the positive predictions are correct over all the positive cases. F1 score is the measure of combining precision and recall, often described as harmonic means of these two. The idea is a single metric that weights the two ratios in a balanced way. Thus, both measures must be high to have a high F1 score. These measures are presented

with values between 0 and 1; a higher value shows the better performance of the model. The following are equations for calculating these measurements (equations 1 & 2). Higher values for each measurement indicate a more accurate model.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

True Positives (TP): Number of samples correctly predicted as "positive."
False Positives (FP): Number of samples wrongly predicted as "positive."
True Negatives (TN): Number of samples correctly predicted as "negative."
False Negatives (FN): Number of samples wrongly predicted as "negative."

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (2)$$

2.2 Floor Area Change Prediction Model

We also created another deep learning model to predict floor area in 2021. The collected municipality tax parcel data has floor area information for each tax parcel. This information in

squared meters is the ground truth of our deep learning model. We used the same sample split with building type classification model to ensure that we trained and tested the model with same samples. To evaluate model accuracy, we used mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), r-squared (R^2) based on predicted floor area for each building. Furthermore, since our main focus was to predict building MS at city-wide scale, we aimed to develop the model that can accurately predict total floor area of multiple buildings. Therefore, we evaluated accuracy of prediction model based on sum of floor area of test data by using total error in addition to other accuracy metrics. We used the following equations to calculate each accuracy measurement (equations 3, 4, 5, & 6).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

$$MAE\ [m^2] = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (5)$$

$$Total\ Error\ [\%] = \frac{|\sum_{i=1}^{n}(y_i - \hat{y}_i)|}{\sum_{i=1}^{n} y_i} * 100 \quad (6)$$

2.3 Material stocks Estimation

We used the classified building type and predicted floor area of test data to estimate building MS at city-scale. The building MS was calculated based on the equation (7). We looked into eight building materials: concrete, cement, wood, brick, gypsum, aggregates, asphalt, and steel by using material intensity data presented in the previous paper (Mollaei, Ibrahim, and Habib 2021; Hottle et al. 2022) (Table 1). We calculated the ground truth of MS based on the ground truth of building type and floor area. Predicted MS were estimated based on the predicted values from our deep learning model. We evaluated our deep learning based building MS estimation by checking the error percentage of each material (equation 8).

$$MS_{m,i} = \sum_{m,i} (MI_{m,i} * FA_{i}) \tag{7}$$

$$Error\ [\%] = \frac{\left|MS_m - \widehat{MS_m}\right|}{MS_m} * 100 \tag{8}$$

Where $MS_{m,i}$ is the material stocks of building material *m* in building type *i*; $MI_{m,i}$ is the material intensity of building material *m* in building type *i*; and FA_i is the floor area of building type *i*.

Table 1: Building Material Intensity adapted from (Mollaei, Ibrahim, and Habib 2021; Hottle et al. 2022) . The unit is [kg/m^2]. SFHs: single-family houses, MFHs: multi-family houses, Non-R: non-residential houses

Material	Concrete	Cement	Wood	Brick	Gypsum	Aggregates	Asphalt	Steel	Total
SFHs	477	43	104	136	100	318	8		1143
MFHs	342	31	109	143	34	295	4	59	985
Non-R	451	41	6	57	5	54	18	87	679

3. RESULT

3.1 Building Type Classification Model

In the evaluation of building type classification models, We find that our CNN models exhibited superior performance over the ViT models. Specifically, the CNN models achieved a peak accuracy of 85.00%, contrasting with 30.97% accuracy exhibited by the best performing ViT model. This discrepancy is reflected across multiple training-validation-test splits, as detailed in Table 2. A notable characteristic of ViT models is that they are more conservative in predicting positive classes; they prefer to be more confident about their positive predictions since they have higher precision values than recall values. Conversely, the CNN models demonstrated a commendable balance in performance, as evidenced by high F1 scores. This indicates proficient identification of positive instances while effectively minimizing false positives. The confusion matrix further corroborates the CNN models' capability to discriminate among three distinct building types (Figure 5). However, it was observed that there was a propensity to misclassify non-residential buildings and multi-family houses as single-family houses. This indicates areas for potential refinement in future develpment of the model.

Table 2: Accuracy Comparison of Building Type Classification Model

Model	Accuracy [%]	Precision	Recall	F1 Score
CNN 60:20:20 ratio	84.71	0.845	0.847	0.836
CNN 80:10:10 ratio	84.84	0.842	0.848	0.842
CNN 90:5:5 ratio	85.00	0.846	0.850	0.843
ViT 60:20:20 ratio	30.97	0.568	0.310	0.359
ViT 80:10:10 ratio	12.78	0.266	0.128	0.052
ViT 90:5:5 ratio	20.45	0.453	0.205	0.137

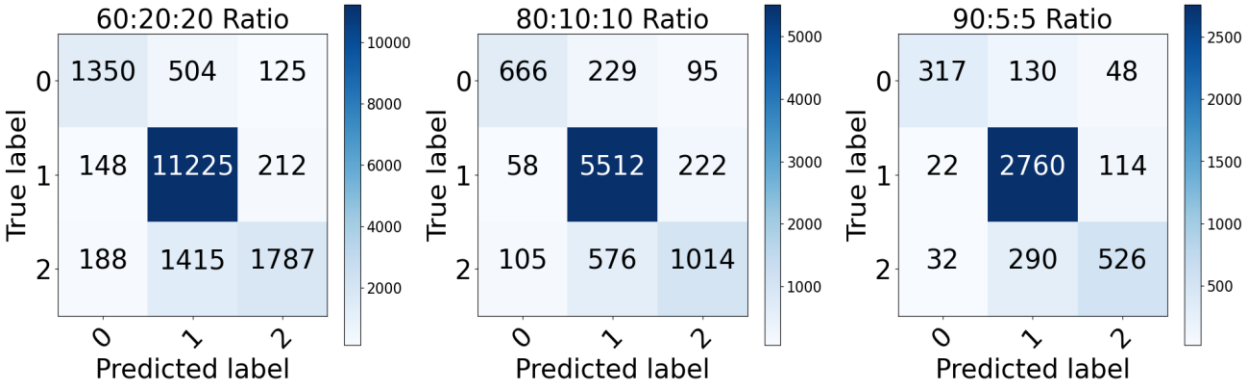

Figure 5: Confusion Matrix of Classification Model. 0: Non-residential, 1: SFH, 2:MFH.

3.2 Floor Area Change Prediction

In contrasting two deep learning methodologies across three data split ratios, ViTs exhibited lower total error when compared to CNNs (Table 3). Specifically, the ViT model with an 80:10:10 simple split ratio achieved a minimum total error percentage of 0.32% on a test set with 8,477 samples. This finding is somewhat counterintuitive, given that other predictive performance metrics suggested that the CNN methods generally predicted floor area with greater accuracy than their ViT counterparts. For instance, R^2 values of CNN methods hovered around

0.40 for two of the data splits, indicating a moderate positive correlation with the observed data. Meanwhile, all ViT models displayed R^2 values around 0.00, signifying a lack of predictive strength.

Table 3: Accuracy Comparision of Floor Area Prediction Model

Model	MSE	RMSE	MAE	R^2	Total Error [%]
CNN 60:20:20 ratio	0.74	0.86	0.61	0.38	1.86
CNN 80:10:10 ratio	0.71	0.84	0.60	0.40	5.05
CNN 90:5:5 ratio	1.05	1.02	0.75	0.10	20.21
ViT 60:20:20 ratio	1.22	1.10	0.78	-0.02	9.28
ViT 80:10:10 ratio	1.20	1.09	0.78	-0.00	0.32
ViT 90:5:5 ratio	1.16	1.08	0.77	0.00	1.54

The overall performance of CNN models in predicting floor area surpassesed that of the ViT models when considering the comprehensive suite of metrics, excluding total error. Such a trend was analogous to the one observed in building type classification, where CNN models outperformed in accuracy. Among the CNN variants, the 60:20:20 ratio displays the lowest total error, making it the preferred model for city-scale estimation of floor area. Concurrently, the accuracy of the CNN building type classification models was consistent across different sample splits, prompting the selection of the CNN model with the 60:20:20 ratio as the most reliable for building type classification.

3.3 Material stocks Estimation

Utilizing the optimal CNN model configuration with a 60:20:20 data split ratio for both building type classification and floor area prediction, we estimated the MS for the city of St. Paul. Our sample,

encompassing 16,954 tax parcels, served as the test data for our study. The model's estimation of MS was reasonably accurate, exhibiting low error rates when juxtaposed with ground truth values, as delineated in Table 4. Exceptional accuracy was observed in the prediction of MS for all building material types—excluding asphalt and steel—with errors for remaining materials under the 2.25% threshold. Concrete and cement MS predictions were particularly precise, with a mere 2.07% error. Furthermore, the aggregate error for all building materials was significantly low at 0.29%.

Table 4: Material Stocks Estimation Results [10^3 ton]

Material	Concrete	Cement	Wood	Brick	Gypsum	Aggregates	Asphalt	Steel	Total
Truth	878.79	79.09	202.84	267.21	168.20	606.42	14.77	24.48	2162.72
Prediction	896.99	80.73	198.31	265.38	165.36	596.38	16.01	30.50	2168.93
Error [%]	2.07	2.07	2.23	0.68	1.68	1.66	8.37	24.57	0.29

4. DISCUSSION

The findings of this study underscore the efficacy of deep learning methodologies in predicting MS at an urban scale by leveraging publicly accessible datasets. The CNN models showcased a high degree of precision, successfully classifying various building types with an accuracy exceeding 84%. This performance is particularly commendable given the complexity of urban landscapes and the diversity of architectural styles captured in aerial and street-view imagery. Moreover, the CNN models demonstrated remarkable capability in estimating floor area, with a total error of 1.86%. This level of precision in predicting floor area is critical as it directly influences the accuracy of MS calculations. When it came to predicting the quantity of a variety types of building materials, the deep learning model achieved a low error rate of 0.29% for a total weight of material, suggesting that the model not only grasps the nuances of material usage

patterns across different building types but also aligns closely with the ground truth of urban construction practices. These outcomes collectively suggest that our deep learning model can serve as a robust tool for predicting city-wide MS estimation using data sources that are both widely accessible and freely available. This methodology stands as a testament to the potential of AI-driven approaches to significantly contribute to the fields of urban development and sustainability, paving the way for more informed and data-driven decision-making in urban planning.

To elucidate the decision-making processes of our CNN model, class activation maps (CAMs) play an instrumental role by revealing which regions of the input imagery command the model's attention during the prediction phase. To shed light on the internal workings of the model, we employed Grad-CAM to generate CAMs for a selection of random test samples, following the methodology proposed by (Selvaraju et al. 2020). As depicted in Figure 6, the CAMs illustrate that our CNN model proficiently identifies the architectural features of buildings within both aerial and street-view images, corroborating the model's configuration detailed in the results section. The utility of CAMs extends beyond mere validation of correct predictions; they provide a visual exposition of the model’s focal points. For instance, the CAMs revealed instances where the model's attention was inadvertently drawn to non-building objects, such as vehicles and foliage in street-view images. These distractions, prevalent in urban settings, have the potential to divert the model’s focus from the building, consequently leading to inaccurate predictions of the floor area or building type. The presence of such objects underscores the need for enhanced feature discrimination within the model to refine its predictive accuracy further.

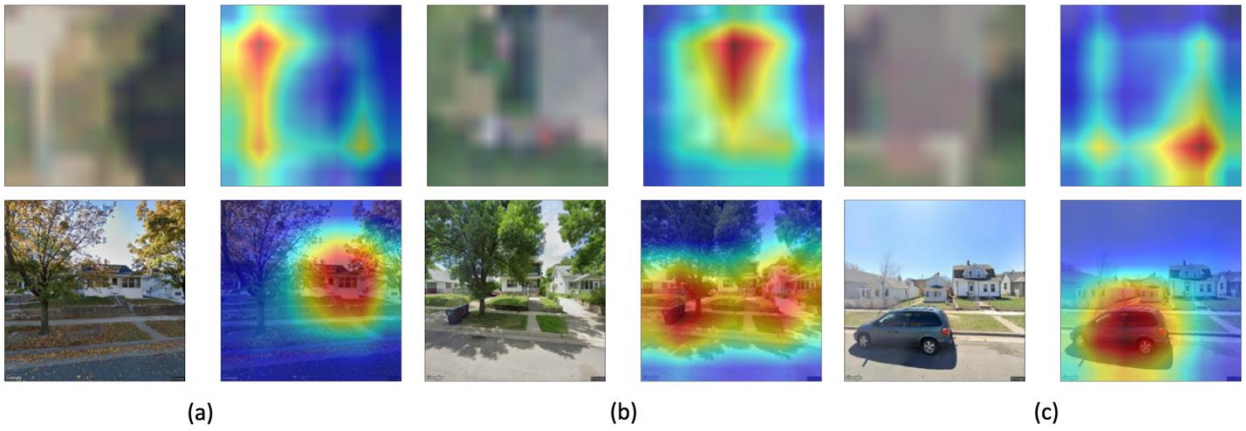

Figure 6: Class Activation Maps for three random samples from the test set. (a) is SFH, and the model correctly classified it as SFH. (b) is MFH, but the model classified it as SFH. (c) is a Non-residential building, but the model classified it as SFH.

In Figures 7 & 8, I juxtapose the actual and predicted total Material Stocks (MS) across the urban landscape of St. Paul to examine the spatial accuracy of our deep learning model. Despite minor discrepancies, the predictive mapping of our model aligns remarkably well with the actual data. While localized deviations exist, they typically remain within the mapping categorical threshold, underscoring the model's spatial precision in estimating MS. This comparison is visually represented through point data corresponding to individual parcel MS, with panel (a) illustrating ground truth and panel (b) depicting predictions (Figure 7). The MS of each tax parcel is aggregated to 200m by 200m grid level by summing up the MS of parcels within the same grid (Figure 8).

Figure 7: Total Material Stocks Map for Test Set in St. Paul. Point markers represent MS of corresponding parcels. The ground truth MS is presented in (a) and the predicted MS is shown in (b).

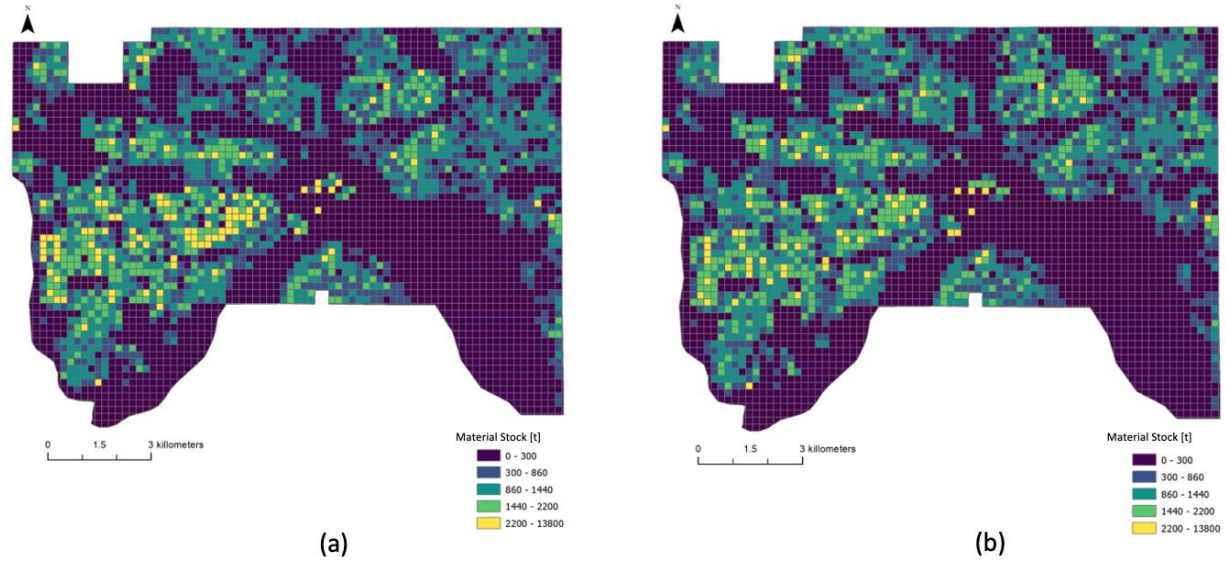

Figure 8: Aggregated Total Material Stocks Map for Test Set in St. Paul. Each pixel is 200m by 200m. The ground truth MS is presented in (a), and the predicted MS is shown in (b).

In this paper, the CNN method outperforms the ViT method for both the building type classification model and the floor area prediction model. The architecture of CNNs incorporates strong inductive biases that are well-suited to image data. These biases include translation invariance, the idea that same object can be recongnized regardless its position in the image. Another bias is locality, which the CNNs make assumption that nearby pixels are more relevant to each other. On the other hand, ViTs have fewer inductive biases regarding the spatial structure of images. They treat the image more like a sequence of tokens and rely on the self-attention mechanism to learn relationships between these tokens. This allows ViTs to potentially capture more global and complex patterns without being constrained by the predefined kernel shapes or local receptive fields of CNNs. However, this also requires a large number of data to learn from, as they must implicitly learn the spatial hierarchies that CNNs are explicitly designed to capture. Therefore, CNNs perform better to classify building type and predict floor area in our study.

The disparity in MS estimation accuracy, particularly concerning steel, necessitates further examination. While our model effectively predicts the MS for various building materials with commendable precision, the steel estimates deviate significantly from the ground truth, resulting in an error of 24.57%. This substantial error is primarily attributed to the challenges our building type classification model faces in accurately identifying MFHs and non-residential buildings. According to the material intensity MI data utilized in this study, steel usage is exclusive to MFHs and non-residential structures. Hence, the capability to precisely recognize these building categories is paramount for accurate steel MS prediction. Unfortunately, the limited representation of MFHs and non-residential buildings in our dataset undermines the model's performance, leading to less reliable steel MS estimations. This issue highlights the importance of having a well-distributed, representative sample across all building types for training purposes. Enhancing the

model's ability to distinguish these specific structures will be essential to improve the accuracy of steel MS predictions and, by extension, the model's overall utility in urban planning and resource management.

5 CONCLUSION

This research delves into the viability of deploying deep learning models to ascertain MS at a comprehensive city-wide level using datasets that are both publicly accessible and prevalent across urban environments. We have crafted two sophisticated deep learning models—the building type classification and floor area prediction models. These models, particularly the CNNs, have proven adept at classifying building types and estimating floor areas within urban settings. Leveraging the predictive outputs of these models, we have formulated a methodology to estimate MS with a remarkable degree of accuracy, even in scenarios where traditional MS estimation techniques may be impractical.

The implications of this study are manifold, extending the frontier of machine learning and deep learning applications within the realms of building construction and urban development. The insights provided from this research offer valuable guidance to policymakers, municipal authorities, and urban planners as they strategize the design of future cities in alignment with the principles of a circular economy on a global scale. Nonetheless, there remains an avenue for improvement in the classification accuracy for multi-family homes and non-residential buildings. To enhance the sophistication and reliability of our approach, the incorporation of additional data dimensions, such as building footprints or heights, is suggested for future research endeavors. Through such enhancements, the predictive capabilities of the models can be refined, thereby bolstering the strategic planning of urban infrastructure and resource allocation.

REFERENCES

Abergel, Thibaut, Brian Dean, and John Dulac. 2017. “Towards a Zero-Emission, Efficient, and Resilient Buildings and Construction Sector.” UN Environment and International Energy Agency.

Abitbol, Jacob Levy, and Márton Karsai. 2020. “Interpretable Socioeconomic Status Inference from Aerial Imagery through Urban Patterns.” *Nature Machine Intelligence* 2 (11): 684–92. https://doi.org/10.1038/s42256-020-00243-5.

Antunes, Mónica, Rodrigo Lino Santos, João Pereira, Paulo Rocha, Ricardo Bayão Horta, and Rogério Colaço. 2022. “Alternative Clinker Technologies for Reducing Carbon Emissions in Cement Industry: A Critical Review.” *Materials* 15 (1): 209. https://doi.org/10.3390/ma15010209.

Arbabi, Hadi, Maud Lanau, Xinyi Li, Gregory Meyers, Menglin Dai, Martin Mayfield, and Danielle Densley Tingley. 2022. “A Scalable Data Collection, Characterization, and Accounting Framework for Urban Material Stocks.” *Journal of Industrial Ecology* 26 (1): 58–71. https://doi.org/10.1111/jiec.13198.

Arehart, Jay H., Francesco Pomponi, Bernardino D’Amico, and Wil V. III Srubar. 2021. “A New Estimate of Building Floor Space in North America.” *Environmental Science & Technology* 55 (8): 5161–70. https://doi.org/10.1021/acs.est.0c05081.

Bao, Yi, Zhou Huang, Han Wang, Ganmin Yin, Xiao Zhou, and Yong Gao. 2023. “High-Resolution Quantification of Building Stock Using Multi-Source Remote Sensing Imagery and Deep Learning.” *Journal of Industrial Ecology* 27 (1): 350–61. https://doi.org/10.1111/jiec.13356.

Barbour, Edward, Carlos Cerezo Davila, Siddharth Gupta, Christoph Reinhart, Jasleen Kaur, and Marta C. González. 2019. “Planning for Sustainable Cities by Estimating Building Occupancy with Mobile Phones.” *Nature Communications* 10 (1): 3736. https://doi.org/10.1038/s41467-019-11685-w.

Bergado, J. R., C. Persello, and A. Stein. 2020. “LAND USE CLASSIFICATION USING DEEP MULTITASK NETWORKS.” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B3-2020 (August): 17–21. https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-17-2020.

Beyond Zero Emissions. 2017. “Zero Carbon Industry Plan: Rethinking Cement.” Beyond Zero Emissions. Australia. https://apo.org.au/node/103031.

Cao, Zhi, Eric Masanet, Anupam Tiwari, and Sahil Akolawala. 2021. “Decarbonizing Concrete: Deep Decarbonization Pathways for the Cement and Concrete Cycle in the United States, India, and China.”

Dodoo, Ambrose. 2019. “Lifecycle Impacts of Structural Frame Materials for Multi-Storey Building Systems.” *Journal of Sustainable Architecture and Civil Engineering* 24 (1): 17–28. https://doi.org/10.5755/j01.sace.24.1.23229.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv. https://doi.org/10.48550/arXiv.2010.11929.

Fishman, Tomer, Heinz Schandl, and Hiroki Tanikawa. 2016. “Stochastic Analysis and Forecasts of the Patterns of Speed, Acceleration, and Levels of Material Stock Accumulation in

Society." *Environmental Science & Technology* 50 (7): 3729–37. https://doi.org/10.1021/acs.est.5b05790.
Fishman, Tomer, Heinz Schandl, Hiroki Tanikawa, Paul Walker, and Fridolin Krausmann. 2014. "Accounting for the Material Stock of Nations." *Journal of Industrial Ecology* 18 (3): 407–20. https://doi.org/10.1111/jiec.12114.
Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. "Using Deep Learning and Google Street View to Estimate the Demographic Makeup of Neighborhoods across the United States." *Proceedings of the National Academy of Sciences* 114 (50): 13108–13. https://doi.org/10.1073/pnas.1700035114.
Google Map Platform. n.d. "Street View Static API." Accessed November 3, 2023. https://developers.google.com/maps/documentation/streetview/overview.
Göswein, Verena, José Dinis Silvestre, Guillaume Habert, and Fausto Freire. 2019. "Dynamic Assessment of Construction Materials in Urban Building Stocks: A Critical Review." *Environmental Science & Technology* 53 (17): 9992–10006. https://doi.org/10.1021/acs.est.9b01952.
He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Deep Residual Learning for Image Recognition." arXiv. https://doi.org/10.48550/arXiv.1512.03385.
Hipp, John R., Sugie Lee, Donghwan Ki, and Jae Hong Kim. 2022. "Measuring the Built Environment with Google Street View and Machine Learning: Consequences for Crime on Street Segments." *Journal of Quantitative Criminology* 38 (3): 537–65. https://doi.org/10.1007/s10940-021-09506-9.
Hottle, Troy, Troy R. Hawkins, Caitlin Chiquelin, Bryan Lange, Ben Young, Pingping Sun, Amgad Elgowainy, and Michael Wang. 2022. "Environmental Life-Cycle Assessment of Concrete Produced in the United States." *Journal of Cleaner Production* 363 (August): 131834. https://doi.org/10.1016/j.jclepro.2022.131834.
Huang, Bo, Bei Zhao, and Yimeng Song. 2018. "Urban Land-Use Mapping Using a Deep Convolutional Neural Network with High Spatial Resolution Multispectral Remote Sensing Imagery." *Remote Sensing of Environment* 214 (September): 73–86. https://doi.org/10.1016/j.rse.2018.04.050.
Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–94. https://doi.org/10.1126/science.aaf7894.
Ji, Chao, and Hong Tang. 2020. "Number of Building Stories Estimation from Monocular Satellite Image Using a Modified Mask R-CNN." *Remote Sensing* 12 (22): 3833. https://doi.org/10.3390/rs12223833.
———. 2022. "Gross Floor Area Estimation from Monocular Optical Image Using the NoS R-CNN." *Remote Sensing* 14 (7): 1567. https://doi.org/10.3390/rs14071567.
Krausmann, Fridolin, Dominik Wiedenhofer, Christian Lauk, Willi Haas, Hiroki Tanikawa, Tomer Fishman, Alessio Miatto, Heinz Schandl, and Helmut Haberl. 2017. "Global Socioeconomic Material Stocks Rise 23-Fold over the 20th Century and Require Half of Annual Resource Use." *Proceedings of the National Academy of Sciences* 114 (8): 1880–85. https://doi.org/10.1073/pnas.1613773114.
Lanau, Maud, and Gang Liu. 2020. "Developing an Urban Resource Cadaster for Circular Economy: A Case of Odense, Denmark." *Environmental Science & Technology* 54 (7): 4675–85. https://doi.org/10.1021/acs.est.9b07749.

Li, Mengmeng, and Alfred Stein. 2020. “Mapping Land Use from High Resolution Satellite Images by Exploiting the Spatial Arrangement of Land Cover Objects.” *Remote Sensing* 12 (24): 4158. https://doi.org/10.3390/rs12244158.

Liu, Miao, Jun Ma, Rui Zhou, Chunlin Li, Dikang Li, and Yuanman Hu. 2021. “High-Resolution Mapping of Mainland China’s Urban Floor Area.” *Landscape and Urban Planning* 214 (October): 104187. https://doi.org/10.1016/j.landurbplan.2021.104187.

Mao, Ruichang, Yi Bao, Zhou Huang, Qiance Liu, and Gang Liu. 2020. “High-Resolution Mapping of the Urban Built Environment Stocks in Beijing.” *Environmental Science & Technology* 54 (9): 5345–55. https://doi.org/10.1021/acs.est.9b07229.

Microsoft. 'Global Building Footprints'.github repository. https://github.com/MICROSOFT/GLOBALMLBUILDINGFOOTPRINTS

Minnesota Geospatial Information, Office. n.d. “Minnesota Geospatial Commons.” Accessed May 22, 2023. https://gisdata.mn.gov/.

Mollaei, Aida, Nadine Ibrahim, and Komal Habib. 2021. “Estimating the Construction Material Stocks in Two Canadian Cities: A Case Study of Kitchener and Waterloo.” *Journal of Cleaner Production* 280 (January): 124501. https://doi.org/10.1016/j.jclepro.2020.124501.

Monteiro, Paulo J. M., Sabbie A. Miller, and Arpad Horvath. 2017. “Towards Sustainable Concrete.” *Nature Materials* 16 (7): 698–99. https://doi.org/10.1038/nmat4930.

Raghu, Deepika, Martin Juan José Bucher, and Catherine De Wolf. 2023. “Towards a ‘Resource Cadastre’ for a Circular Economy – Urban-Scale Building Material Detection Using Street View Imagery and Computer Vision.” *Resources, Conservation and Recycling* 198 (November): 107140. https://doi.org/10.1016/j.resconrec.2023.107140.

Reyna, Janet L., and Mikhail V. Chester. 2015. “The Growth of Urban Building Stock: Unintended Lock-in and Embedded Environmental Effects.” *Journal of Industrial Ecology* 19 (4): 524–37. https://doi.org/10.1111/jiec.12211.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” *International Journal of Computer Vision* 128 (2): 336–59. https://doi.org/10.1007/s11263-019-01228-7.

Soonsawad, Natthanij, Raymundo Marcos Martinez, and Heinz Schandl. 2022. “Material Demand, and Environmental and Climate Implications of Australia’s Building Stock: Current Status and Outlook to 2060.” *Resources, Conservation and Recycling* 180 (May): 106143. https://doi.org/10.1016/j.resconrec.2021.106143.

Tanikawa, Hiroki, Jing Guo, and Tomer Fishman. 2022. “Spatial-Temporal Views on Urban Construction Material Flow and Stock towards Sustainability.” In *Routledge Handbook of the Extractive Industries and Sustainable Development*. Routledge.

United States Geological Survey. 2017. “National Agriculture Imagery Program (NAIP).” USGS EROS Archive - Aerial Photography - National Agriculture Imagery Program (NAIP). 2017. https://www.usgs.gov/centers/eros/science/usgs-eros-archive-aerial-photography-national-agriculture-imagery-program-naip?qt-science_center_objects=0#qt-science_center_objects.

Xing, Xiaoyue, Zhou Huang, Ximeng Cheng, Di Zhu, Chaogui Kang, Fan Zhang, and Yu Liu. 2020. “Mapping Human Activity Volumes Through Remote Sensing Imagery.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13: 5652–68. https://doi.org/10.1109/JSTARS.2020.3023730.

Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. “Using Publicly Available Satellite Imagery and Deep Learning to Understand Economic Well-Being in Africa.” *Nature Communications* 11 (1): 2583. https://doi.org/10.1038/s41467-020-16185-w.

Zhang, Ce, Isabel Sargent, Xin Pan, Huapeng Li, Andy Gardiner, Jonathon Hare, and Peter M. Atkinson. 2018. “An Object-Based Convolutional Neural Network (OCNN) for Urban Land Use Classification.” *Remote Sensing of Environment* 216 (October): 57–70. https://doi.org/10.1016/j.rse.2018.06.034.

Zhang, Xiaoyong, Zhengchao Chen, Yuemin Yue, Xiangkun Qi, and Charlie H. Zhang. 2019. “Fusion of Remote Sensing and Internet Data to Calculate Urban Floor Area Ratio.” *Sustainability* 11 (12): 3382. https://doi.org/10.3390/su11123382.

Zhao, Yijiang, Xiao Tang, Zhuhua Liao, Yizhi Liu, Min Liu, and Jian Lin. 2023. “Multi-Type Features Embedded Deep Learning Framework for Residential Building Prediction.” *ISPRS International Journal of Geo-Information* 12 (9): 356. https://doi.org/10.3390/ijgi12090356.

Zhou, Wen, Dongping Ming, Xianwei Lv, Keqi Zhou, Hanqing Bao, and Zhaoli Hong. 2020. “SO–CNN Based Urban Functional Zone Fine Division with VHR Remote Sensing Image.” *Remote Sensing of Environment* 236 (January): 111458. https://doi.org/10.1016/j.rse.2019.111458.

Zhou, Wen, Claudio Persello, Mengmeng Li, and Alfred Stein. 2023. “Building Use and Mixed-Use Classification with a Transformer-Based Network Fusing Satellite Images and Geospatial Textual Information.” *Remote Sensing of Environment* 297 (November): 113767. https://doi.org/10.1016/j.rse.2023.113767.

APPENDIX

Table A 1: Categorization of Building Types

Category	Building Type in Original Data
Single-Family Houses	"SINGLE FAMILY DWELLING, PLATT" "SINGLE FAMILY W/ACCESSORY UNI" "TWIN HOME" "TOWNHOME - DETACHED UNIT" "BED & BREAKFAST" "RESIDENTIAL, OTHER"
Multi-Family Houses	"TWO FAMILY DWELLING - SIDE/SI" "TWO FAMILY DWELLING - UP/DWN" "APARTMENTS 4-6 RENTAL UNITS" "APARTMENTS 7-19 RENTAL UNITS" "APARTMENTS 20-49 RENTAL UNITS" "APARTMENTS 50-99 RENTAL UNITS" "APT OR COMPLEX 100+ UNITS" "CONDO" "TOWNHOME-INNER UNIT" "TOWNHOME - OUTER UNIT" "TOWNHOME - TICO" "CONDO GARAGE" "CONDO STORAGE UNIT" "RESIDENTIAL CO-OP" "APARTMENT VACANT LAND" "APARTMENT MISC IMPROVEMENT" "TOWNHOME - VACANT LOT" "TOWNHOME - GARAGE ONLY" "TOWNHOME - NON-TAX OUTLOT" "CONDO VACANT LAND" "ASSISTED LIVING APT COMPLEX" "NURSING HOME & PRIVATE HOSPIT" "FRATERNITY/SORORITY HOUSE" "2ND RESID 4+ UNITS, CLASS APT" "THREE FAMILY DWELLING, PLATTE" "TWO RESIDENCES ON ONE PARCEL"
Non-Residential Buildings	" " (This is empty rows, but they indicate non-residential buildings in this dataset) "RESIDENTIAL, VACANT LAND, LOT"