Seeing Is Believing: Understanding and Expanding Multifamily Building Retrofit Markets Through Large-Scale Computer Vision Analysis of Building Characteristics

Conner Geery, Nathan Hinkle, and Allyson Dugan; Cadmus Group
James Geppner; New York State Energy Research and Development Authority[1]

ABSTRACT

Decarbonizing millions of buildings equitably, effectively, and affordably demands urgent innovation and financial resources, but market data for retrofit solutions is largely unavailable. To address this gap, our team extracted details of over 130,000 multifamily buildings—nearly every one in New York State—by using machine learning to develop a novel process to analyze over 650,000 aerial images. By benchmarking our results against data collected in 391 in-person site visits, we were able to assess the model's performance compared to human observation across a range of building styles, sizes, densities, and vintages.

The resulting dataset contains dozens of attributes for each multifamily building. Energy efficiency and clean energy programs, manufacturers, providers, and others can search and filter these data to determine the distribution and precise locations of retrofit candidates, build capacity based on accurate information about the size of the market, and prioritize investments in technologies that address market gaps.

This paper describes our challenges with and solutions for systematically acquiring multiple aerial images of each building; distinguishing the targeted buildings from others nearby; training and deploying instance segmentation models to identify building features; and analyzing the outputs to calculate building configurations and dimensions, window-to-wall ratios, rooftop equipment quantities, and other metrics. We also explore how the results of this study can be used to direct investment in building retrofit programs, the performance and limitations of our technique and how to apply it in other jurisdictions, and topics for future research.

Introduction

The residential and commercial construction industry must make substantial changes to achieve the rate of building decarbonization necessary to meet the objectives established by the Paris Agreement and the New York State (NYS) Climate Leadership and Community Protection Act. Upgrading or retrofitting over 100 million buildings across the United States in the next 30 years will require the industry to create standardized offerings that appeal to sizable populations of building owners (Harris 2021) by defining customer segments and designing products or offerings that deliver significant and evident value to those customers. For the construction industry, manufacturers need to identify buildings that are a good match for their current or future product offerings, but data at this level of granularity is not currently available.

The process of identifying and delivering viable new offerings involves identifying the size of the market for a packaged solution, forecasting demand, making decisions about resource

[1] Any opinions expressed, explicitly or implicitly, are those of the authors and do not necessarily represent those of the New York State Energy Research and Development Authority. All results and analyses described were developed by Cadmus and its subcontractors.

allocation, developing a price estimate for the fully installed solution, estimating the energy and costs that the solution will save, and then engaging proactively with building owners to discuss the bundled measures. For example, an exterior insulated panel manufacturer may want to know how many masonry buildings in a region have a similar volume, window-to-wall ratio (WWR), and height to estimate how many are a match for a specific retrofit solution. Decisions about whether to increase capacity, develop a packaged solution, partner with an installer, or begin an installer training (Harris 2021) will impact the cost and availability of the selected solution in that area. Existing data on multifamily buildings is often sparse and unreliable, with basic attributes typically generalized at the property level and many useful variables such as WWR and roof configuration rarely if ever reported. Manufacturers face a high degree of uncertainty about their investments and will be less likely to bring new solutions to market without the ability to systematically identify buildings that meet specific criteria.

Access to detailed information about buildings and equipment allows installers to group buildings that may benefit from the same solution and to perform a general remote site review before going to the location. For example, information about ease of access to the property and features like fire escapes or rooftop equipment may streamline selecting and installing appropriate solutions. Remote screening reduces the total cost of solution deployment, enabling new offerings to be more cost-effective and reach more customers.

The New York State Energy Research and Development Authority (NYSERDA) contracted with Cadmus to conduct its first statewide baseline study of multifamily buildings, including a combination of interviews and on-site inspections to collect detailed data on building and equipment specifications and operations. Data processing in these studies involves statistical sampling and weighting methods to produce reliable estimates of overall equipment saturations and other valuable metrics at an overall population level and for individual strata. However, such studies generally rely on proxy variables such as total floor area and are limited in their ability to identify individual building locations and attributes. In addition to conducting a traditional baseline study, Cadmus developed a novel approach to locate and analyze externally visible attributes of nearly every individual multifamily building across the state. The resulting dataset is more comprehensive than a traditional baseline study and is accessible via a public dashboard for companies to remotely match a particular solution to a particular building—a task that cannot be performed at a large scale today but that will enable companies to target building decarbonization investments at the rate and scale necessary to achieve ambitious climate targets.

Methodology

The primary goal for the computer vision analysis Cadmus performed was to develop a dataset of building characteristics for the full population of multifamily buildings in NYS that NYSERDA and market actors could use to determine market sizing information for new offerings or identify potential retrofit candidates. The final dataset consisted of building characteristics derived from traditional data sources and through computer vision analysis of aerial images. Because most existing data sources report at the parcel level, we used advanced property analytics to disaggregate this data to the building level.

Population Data

Cadmus used Res-Intel's Benchmark.AI Multifamily Characterization Toolset to develop an inventory and initial metrics for the full population of multifamily buildings in NYS based on

traditional public and private data sources (Nelson and Johnson 2022). Using county and municipal parcel data, census data, and commercial property data from this toolset, combined with a dataset of low- and moderate-income multifamily buildings previously produced for NYSERDA, the team identified multifamily sites and their attributes. This process sourced building footprint data from county and city GIS systems, OpenStreetMap, and Bing Maps to locate individual buildings, then extracted building heights from the NYS GIS program office's Light Detection and Ranging (LiDAR) data, which is generated using laser measurements from an aircraft (NOAA 2024).

The team then disaggregated these parcel-level attributes, distributing and recalculating those reported as a single value for an entire property into distinct values for individual buildings on that property. Because parcel data reports a single value for address, number of floors, and units even for properties that comprise more than one building, our analysis matched all building footprints to the parcel, geocoded individual addresses and estimated floors for each building, and distributed reported residential units among the buildings proportionally by approximate total floor area, resulting in more-granular building-level population data.

Aerial Imagery

Before starting our analysis, we conducted a thorough review of public and commercial imagery sources, considering factors such as coverage, image resolution, API and bulk access options, measurement features, and cost to ensure the selection of comprehensive and high-resolution imagery for our study area. While oblique (side-view) and 3D imagery exists from free sources such as Google Maps and Bing Maps, licensing restrictions preclude using those sources for automated analysis. Other public domain aerial image sources such as those provided by state and federal governments typically offer only low-resolution orthogonal (top-down) images. Thus, Cadmus opted to use a commercial image vendor with existing statewide coverage in New York and an image access and measurement API for bulk analysis.

We used the central latitude and longitude of each building as the search parameter for the vendor's API to identify and download sets of high-resolution images like those in Figure 1, consisting of one orthogonal and four oblique images for the area around each building in our population. Analyzing oblique imagery from multiple directions in addition to orthogonal imagery captures the most complete perspective of a building (Wilson and Williams 2019). We also retrieved additional image metadata such as the original capture date, ground surface resolution, and positional reference data from the vendor's image retrieval API.

Figure 1: Example of multifamily buildings seen in orthogonal (left) vs. oblique (center and right) aerial imagery captured via low-altitude fixed-wing aircraft. *Imagery source:* Pictometry International Corp. Copyright © 2023. All Rights Reserved.

Computer Vision Model Training

Computer vision uses human-labeled example images to train a machine learning model to recognize similar objects within new images (V7 Labs 2024). Before training could begin, Cadmus had to select an appropriate model type for classifying each feature. We considered two types of models—object detection and instance segmentation, depicted in Figure 2. Object detection provides a simple bounding box around each instance of an object, while segmentation provides a pixel-level mask around each instance or area of the image. Object detection is simpler and much more efficient to train, but segmentation is more precise—especially for complex shapes. We ruled out image classification, an even simpler method that applies a single category to an entire image, because a single image may contain numerous buildings and features. Based on our evaluation, we determined that instance segmentation was the best fit for identifying irregularly shaped features like building roofs, walls, and windows, while object detection was suitable for identifying small, isolated items like exterior HVAC units.

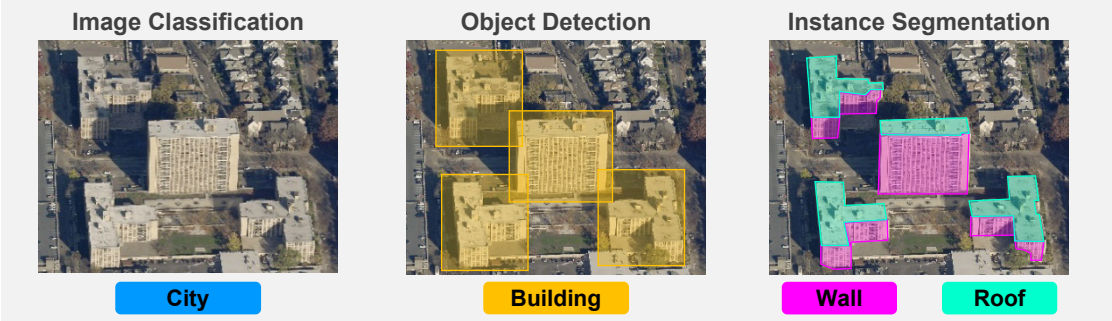

Figure 2: Example of computer vision identification methods. *Imagery source:* Pictometry International Corp. Copyright © 2023. All Rights Reserved.

Cadmus created two training datasets: one for oblique images and one for orthogonal images. We developed the labeling schema, or classes, based on common visually identifiable building characteristics for the NYS multifamily building population. Our final labeling schema and quantity of training images for both datasets is summarized in Table 1.

Table 1. Labeling schema classes, model types, and training data

Dataset	Training file count	Class	Method
Oblique	265	Building	Segmentation
		Roof	Segmentation
		Wall	Segmentation
		Window	Segmentation
		Window AC	Object detection
		Minisplit	Object detection
Orthogonal	330	Flat roof	Segmentation
		Shingle roof	Segmentation
		Metal roof	Segmentation
		Rooftop patio	Segmentation
		Solar panel	Segmentation
		Skylight	Segmentation
		RTU	Object detection
		Split system	Object detection

A team of human annotators labeled the training images, drawing a polygon around each individual instance of the targeted features and assigning it to the corresponding class label. Figure 3 shows an example of a training image with hundreds of manually annotated features alongside an example of the automated outputs for an image processed through the final model.

Figure 3: Example of model training image (left) and model output image (right).
Imagery source: Pictometry International Corp. Copyright © 2023. All Rights Reserved.

We followed an iterative approach, shown in Figure 4, to train computer vision models that would perform well across the diverse building stock for this project. Team members with expertise in building systems worked together to annotate the initial sample of images and establish clear criteria for additional annotators to distinguish building features and assign them to the correct class. As we developed the models, we assessed their performance on test images and analyzed errors to identify patterns and challenges. We iteratively augmented the training datasets and retrained the models with additional instances for classes that did not consistently identify the targeted features, addressing weaknesses identified in each evaluation until the models demonstrated the performance described in the Evaluation of Results section. The cloud-based software we used required a minimum of 10 images and 100 instances to begin training a new model, but in practice with our dataset we needed at least 1,000 instances for the model to correctly identify features on the majority of the test images. The greater the visual variation of the overall images and the specific features being analyzed, the more example instances were required for the model to learn to reliably distinguish them. Cadmus trained the model using at least five images from different zip codes within each unique combination of three building sizes, five vintages, and four regions across the diverse building population to ensure that each relevant class was adequately represented.

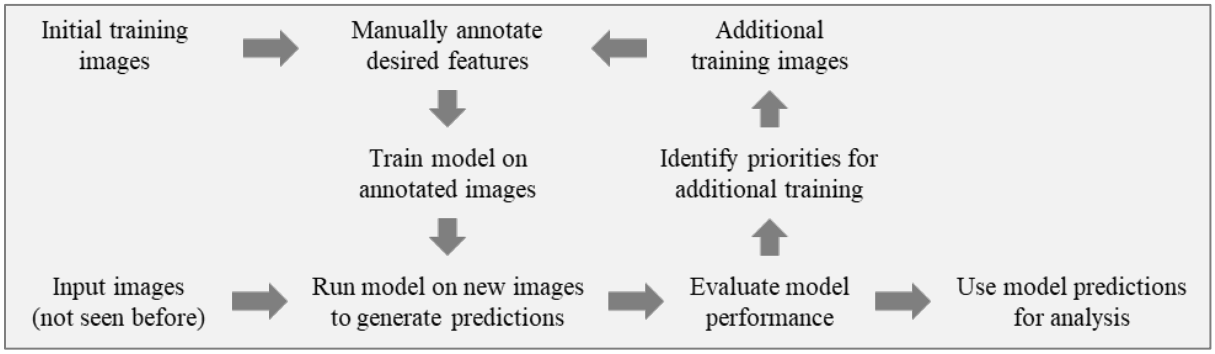

Figure 4. Computer vision model development process.

Model Deployment

To ensure the correct identification and assessment of multifamily structures, especially in dense urban areas, we developed methods for distinguishing targeted buildings from others nearby. Available aerial imagery was often produced at a high range and high resolution, providing data on entire neighborhoods in which we wanted to focus on only one or two buildings.

Using the building footprints from our population data development process, we cropped the high-resolution images to include only the clusters of multifamily buildings within our population. Using existing footprint data sources was expedient, ensured consistency with other datasets, and allowed us to correctly separate buildings with adjacent walls, which are common in urban areas like New York City (NYC). For future research in areas without existing footprint data, solutions are available to extract footprints from medium- and low-resolution orthogonal images (Singh et al. 2022), although this process requires additional computation and tends to produce less-accurate footprints compared to cadastral sources.

Scaling our methodology involved running the final models on 123,000 orthogonal images and 488,000 oblique images, with each image often containing multiple buildings of interest. To reduce the processing load of the full population image dataset, we used a batching process and cropped images to show only buildings of interest, giving our models less data to interpret and limiting the scope of required annotation.

Output Analysis

Cadmus analyzed the image annotations the computer vision model produced to extract useful building metrics including window-to-wall ratio, building height, primary roof material, and count of externally visible HVAC equipment. A summary of key metrics in the final dataset is shown in Table 2. We detail the process of filtering annotations from an entire image to only those relevant to an individual building in the Challenges and Solutions section.

Table 2. Sources of building characteristic data in the final dataset

Characteristic	Building population	Image analysis
Unique Building Identifier (UBID)	✓	
County, address, and zip code	✓	
Assessor's parcel number or BBL/BIN	✓	
Building centroid (latitude/longitude) and footprint	✓	
Year of construction	✓	
Footprint area and perimeter	✓	✓
Building height and floors	✓	✓
Envelope/facade square footage		✓
Window-to-wall ratio		✓
Roof construction and configuration		✓
Presence of rooftop equipment (e.g., HVAC units, solar PV, skylights)		✓
Number of residential units	✓	✓
Estimated interior square footage	✓	✓

Window-to-wall ratio. To determine the WWR for a building, we calculated the total area of all window and wall annotations from each available oblique image orientation. We calculated areas directly in pixel units for each orientation to allow for differences in image resolution and scale ratio, and then we divided the resulting window area by wall area to calculate the dimensionless ratio without requiring conversion to feet.

Building height. We estimated the building height from each oblique image by taking vertical cross sections of all wall annotations associated with the targeted building and identifying the longest cross-section line, as shown in Figure 5. We converted the height from pixels to feet using the vertical scale ratio provided by the image vendor API for that section of the image.

Figure 5: Examples of building height measurement using vertical cross-section detection.
Imagery source: Pictometry International Corp. Copyright © 2023. All Rights Reserved.

We compared the estimated height from oblique images to the building height from LiDAR data (if available) and selected the most reliable height to report from the available estimates. Our testing showed that when LiDAR height and image analysis height are close, the LiDAR height is more precise because of the nature of how LiDAR measurement works. In those instances, we reported the LiDAR height as the final height value. Otherwise, we selected the most probable height by looking at the building's reported number of floors and at clusters of similar image-based height estimates from the building's multiple images.

Number of floors. Similarly, to determine the number of floors in a building, we developed selection criteria to choose the most reliable value considering data points from the image analysis and the population dataset. The population dataset reports the number of floors as a single value for an entire parcel, which may contain multiple buildings of different heights. When a building's final height estimate aligned with the reported number of building floors for the parcel, we used the reported number of floors from the parcel data as our final value. When the estimated building height did not align with the parcel's reported number of floors, we predicted the number of floors from the final building height estimate by assuming an average floor-to-floor height of 11.6 feet. We derived this assumption from a distribution analysis of buildings from the initial population dataset where only one building is on the parcel, LiDAR data is available, and the average height per floor is in a plausible range for the number of floors.

Total square footage. From the predicted number of floors, we calculated the total multifamily square footage as the building's footprint area multiplied by the number of residential floors and

adjusted the predicted number of floors if the building was reported to be a mixed commercial and residential use building, assuming that the first floor of a mixed-use building is commercial. Based on the building's calculated multifamily floor area, we estimated its number of residential units by allocating the reported number of residential units on the parcel among all multifamily buildings on the parcel proportionally to each building's fraction of the total estimated multifamily floor area for all buildings on the parcel.

Roof material and components. We analyzed the orthogonal image of each building to determine its primary roof material, the presence of solar panels and skylights, and the count of rooftop HVAC units. In cases where the segmentation model yielded overlapping roof-type annotations, we prioritized the annotation for each pixel with the highest confidence score reported by the model. We then determined the primary roof material based on the class with the highest coverage of the building's footprint area.

Exterior HVAC systems. We determined the number of visible exterior HVAC system objects for the building by counting split systems and packaged unit bounding boxes that overlapped with the building footprint on the orthogonal images and counting window AC bounding boxes contained within a wall segment attributed to the building in oblique images.

Challenges and Solutions

Artificial intelligence (AI) instance segmentation is well documented and can be implemented with off-the-shelf software and cloud services. Producing usable computer vision models that consistently identify the desired segmentation classes was relatively straightforward. However, preparing the appropriate images and associating the model outputs with the correct buildings presented numerous technical challenges, and analysis at a statewide scale demanded solutions that could reliably be run autonomously across a diverse population.

Conversion Between Image and Spatial Coordinates

Figure 6 demonstrates images for which our model has annotated similar features on multiple adjacent visible buildings that must be differentiated for analysis. This requires precise conversion between image pixel coordinates and physical world coordinates.

Figure 6: Examples of oblique images with multiple adjacent buildings visible. Attributing image features to the correct building requires converting between pixel and spatial coordinates. *Imagery source:* Pictometry International Corp. Copyright © 2023. All Rights Reserved.

Such conversions require calibration of the intrinsic lens characteristics and the exact position and rotation of the camera used to take each photo (Zhang 2000); however, such details are generally not available for commercial imagery. Instead, vendors conduct these calculations internally using proprietary data and share results via an API. The imagery vendor we contracted for this project provided us with basic image metadata via its API, including the latitude and longitude coordinates of each corner of the image. Direct conversion of multiple points between the image and world coordinates was not available in the API service we used.

We converted ground-level spatial coordinates to and from corresponding image pixel coordinates by representing each with a cartesian coordinate system and developing a transformation matrix to convert between them, enabling us to identify the region of the image to analyze and to relate image analysis results back to a physical location. We used affine transformation for orthogonal images. The oblique images had a nonlinear scale, meaning equidistant world coordinates farther from the camera position were closer together in pixel coordinates. Using the four corner world coordinates from the image vendor API and the image's pixel dimensions, we created a perspective transformation matrix using a projected coordinate system to provide the correct nonlinear projection. The resulting matrix can be applied to any singular coordinate point or to all points in a complex polygon such as a building footprint or image annotation for seamless conversion between coordinate systems.

The perspective transformation approach was effective in converting between coordinate systems in most cases; however, precision sometimes declined for images covering a large physical area, particularly in areas with significant variation in terrain elevation. The location data provided by the image vendor’s API is calibrated to the actual surface elevation using a digital elevation model, introducing a third dimension to the underlying data that is not exposed via the API and cannot easily be recreated. The impact is most significant when a particularly steep terrain feature is near a single corner of the image resulting in an uneven distance scale along an image axis. We did not identify a suitable solution to this issue with the data available to us in this project.

Vertical Projection of Building Footprints onto Oblique Images

We successfully applied the perspective transformation to translate building footprints into oblique image coordinates. Cropping the image and selecting the relevant annotations required vertical translation of the ground-level footprint. We used the vendor-provided measurement API to calculate the number of pixels corresponding to the presumed height of the building based on estimated height in feet and then translated the footprint vertically.

A challenge emerged when we reviewed the results of this method: because the image perspective was rotated based on the angle of the camera in flight, projections assuming a completely vertical image of a building did not properly align with the targeted building on images with a tilt. We were able to develop a vendor-specific solution using image metadata to calculate a correction rotation, shown in Figure 7, which we applied to the image during the cropping process.

Figure 7: An original image (left) is first rotated to correct for sensor angle, next building footprints are projected onto the image (center), then translated to form a bounding polygon around each building. *Imagery source:* Pictometry International Corp. Copyright © 2023. All Rights Reserved.

After we corrected the rotation, the vertical projections consistently located the desired buildings within a large image correctly and efficiently, enabling automated cropping and feature extraction. This solution was successful for most of our images; however, it is not generalizable to other image sources. We researched other existing automated image alignment tools and algorithms, such as those designed to identify sideways images in smartphone photo galleries, and identified RotNet as a potential solution (Sáez 2017). However, testing showed that RotNet's pre-trained models were not effective with oblique aerial imagery, likely because the training dataset consists primarily of street-level photographs. We recommend further research into training RotNet with aerial imagery, which was not possible within the scope of this project.

Identification and Segmentation of Overlapping and Obstructing Buildings

Although the vertically projected building footprints were successful in identifying the area of an oblique image where a building should be located, the accuracy of this mapping varied with the precision of the underlying footprint GIS data, image metadata, and the initial building height estimate. Projected building footprints are generally sufficient to identify freestanding buildings but are often inadequate to distinguish adjacent buildings in dense urban areas. Target buildings were often obstructed on oblique images by adjacent buildings, vegetation, terrain, or other elements in the photo. To resolve these issues, we added to our oblique instance segmentation model a building outline class, which we trained to segment each distinct building even when it touched adjacent buildings. In the analysis process we overlay the projected building footprint with the building outline annotations to predict the area of the image containing relevant imagery of the selected building.

This approach was effective in most areas of the state, although some buildings were fully obstructed in some orientations by adjacent buildings in dense urban areas including NYC. For images within NYC only, we used official building footprints with accurate height measurements from the city GIS office. We vertically projected building footprints of the targeted building and all adjacent buildings in the image and excluded annotations in parts of the image where the targeted building footprint was obstructed by adjacent buildings. Figure 8 provides an example of this process, which reduced the inclusion of annotations of obstructing buildings. Similar data was not readily available elsewhere in the state; this limited us to targeting annotations based on the building outline segmentation model, which occasionally resulted in detecting a portion of adjacent obstructing buildings. Future studies could address this issue by developing obstruction masks directly from LiDAR data where available.

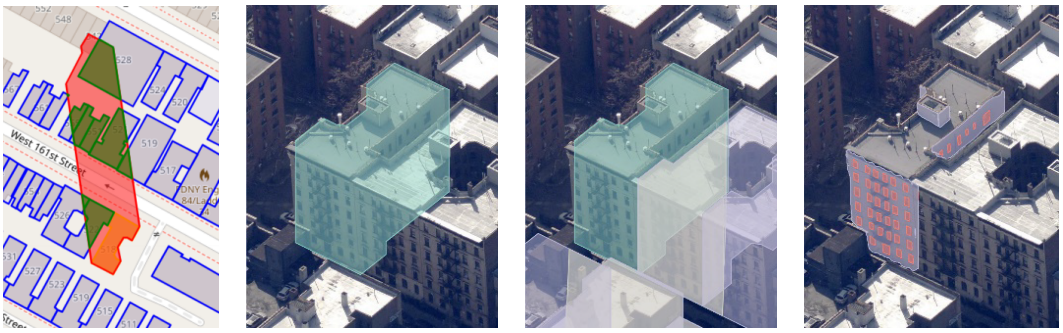

Figure 8: Example of selecting only relevant annotations (right) by identifying nearby building footprints (left), projecting the outline of the targeted building onto the image (center-left), then projecting adjacent buildings onto the image (center-right). *Imagery source:* Pictometry International Corp. Copyright © 2023. All Rights Reserved. *Basemap source:* NYC DoITT, OpenStreetMap.

Correction for Shifted Orthogonal Images

Orthogonal imagery providers typically orthorectify images to align imagery with the underlying topography, making it possible to take measurements directly off the image. However, orthorectification only ensures that features at ground level are true to scale. The primary application for orthogonal images in this project is detection of rooftop materials and equipment, which must be attributed to each building's footprint geometry. The farther a building is from the center of the image the more likely there is to be perspective distortion, causing a portion of the side of the building to be visible and the rooftop position to be transposed from the ground-level footprint of the building. The higher a rooftop is, the greater the offset may be. Imagery captured by lower-flying aircraft contains more detail but exhibits greater distortion of elevated features compared to higher-altitude captures. The imagery used for this project was captured at varying resolutions and altitudes, resulting in inconsistent offsets that made attributing rooftop features to the correct building footprint challenging.

Cadmus developed and applied a translation factor algorithm to systematically align the building footprints with the roof positions on each image. We calculated a loss function comparing the area of the building footprint polygons to the area of the roof annotations. The closer the building polygons are to the detected rooftop area on the image, the lower the output is of the loss function. We applied an optimization algorithm to minimize the loss function by adjusting x and y offset input variables, shifting the footprints until they best overlapped with the rooftop annotations. We used these aligned footprints, demonstrated in Figure 9, for all subsequent analyses of the image. We found a mean required shift distance of 35 pixels with a standard deviation of 37 pixels, but some images required a shift of over 400 pixels to align the footprints to the rooftop positions. This solution depends upon existing building footprint data, but a proposed alternate approach uses a machine learning model to segment roof area from the sides of a structure (Chen et al. 2021), which could enable simultaneous roof and ground-level footprint extraction.

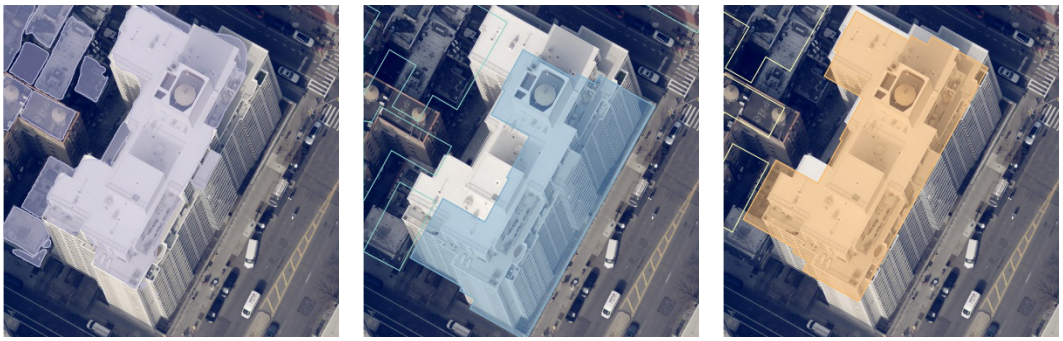

Figure 9: Minimizing overlap between rooftop annotations (left) and initial ground-level building footprints (center) shifts the building footprints to align with roof level (right). *Imagery source:* Pictometry International Corp. Copyright © 2023. All Rights Reserved.

Temporal Alignment of Disparate Data Sources

Aggregating and aligning data from disparate sources with varying refresh cycles led to assorted issues. Some datasets such as county property assessor data are typically updated annually while others are updated on an as-needed basis. LiDAR data in some areas of NYS had not been updated since before 2014 at the time of this study. In addition to the variable data update timelines, buildings can exist in various stages of completion that are not documented in property records. We observed new buildings under construction in recent aerial images where property data indicated a decades-old year of construction, and we found properties with a recent year of construction with either empty plots of land or old buildings. We flagged instances where the year of construction was reported to be more recent than the image capture date and found that less than 0.2% of the images analyzed predated the reported year of construction.

We did not track the update dates for all primary data sources we used in the population analysis. For future projects we recommend tracking the last-updated date for all data sources used throughout the analysis, including all aggregated data sources for the property analytics process.

Photogrammetric 3D Models as an Alternative to 2D Image Analysis

Early in the project the team evaluated using photogrammetric 3D models in lieu of directly analyzing 2D images. A georeferenced digital 3D model confers many advantages, including the ability to easily distinguish separate buildings, simplify measurements of heights and other physical dimensions, and attribute detected material types and building characteristics to specific exposures of the building. Prior research has demonstrated the effectiveness of developing photogrammetric 3D models for energy audits on individual buildings (Singh et al. 2022); however, we did not find this approach to be cost-effective at a statewide scale. Although large technology companies like Google and Microsoft have constructed photogrammetric 3D models of many large cities including NYC for display in their consumer mapping applications, these models cannot be licensed for automated research use. This is likely to change in the coming years as advances in automated photogrammetric approaches and computing power increase availability of these solutions and may eventually offer a superior alternative to processing 2D oblique images.

Evaluation of Results

We found that our process reliably estimated the core metrics of interest for many multifamily buildings, though it is worth noting that all machine learning algorithms produce imperfect estimates. This computer vision analysis was specifically constrained by the visibility of building features and overall quality of the aerial images. For our evaluation, we used real-world data from 391 site visits and manually reviewed the results to verify the true value for each compared metric.

Analysis of Quantitative Metrics

We developed a statistical analysis for quantitative values present in both datasets, including the number of floors, gross building square footage, and WWR, summarized in Table 3 across several strata. Other metrics including number of residential units, total building height, and building façade area were not directly comparable to the collected site visit data.

Table 3: Summary of estimated mean across full building inventory (N=122,604), and average difference between verified values and corresponding image analysis values from 391 site visits

Category	Stratum	Building inventory mean and average difference in sample		
		Building floors	Gross floor area	WWR
Building Size	1-3 stories	2.5 ▼ 0.11	10,500 ▼ 1,200	0.104 ▼ 0.005
	4-7 stories	5.0 ▲ 0.05	24,900 ▲ 1,400	0.122 ▼ 0.011
	8+ stories	13.8 ▲ 0.08	180,900 ▲ 13,800	0.165 ▲ 0.026
Density	Urban	4.4 ▼ 0.05	29,700 ▲ 2,700	0.118 ▼ 0.001
	Rural	2.4 ▲ 0.09	13,800 ▲ 1,100	0.078 ▼ 0.008
Commercial space present	Mixed use	5.6 ▲ 0.38	45,700 ▲ 24,000	0.121 ▲ 0.017
	Residential	4.1 ▼ 0.11	24,800 ▲ 60	0.115 ▼ 0.004
Height source	LiDAR	4.4 ▼ 0.08	27,500 ▲ 1,300	N/A
	Image analysis	4.2 ▲ 0.59	43,400 ▲ 10,600	N/A
Overall		4.4 ▼ 0.05	29,100 ▲ 2,600	0.116 ▼ 0.001

▼ = image analysis underestimates compared to site visits, ▲ = image analysis overestimates compared to site visits

Our process underestimated WWR by 0.001 on average and was most reliable for buildings that had minimal obstruction for all oblique orientations. Higher-resolution images with higher-confidence window annotations also improved the reliability of the WWR estimates. If the obstructed side of a building had fewer or more windows than the other sides, this distorted the building's reported WWR. Greater discrepancies also existed in urban areas due to window and wall annotations for an obstructing building being mistakenly associated with the targeted building. We believe the WWR estimates are sufficiently accurate to support screening buildings for cladding retrofit solutions.

A building's calculated number of floors is an input for calculating gross floor area and number of residential units, making it an important metric for the reliability of results. Our analysis was more likely to underestimate the number of floors in low-rise buildings and to overestimate them in mid- and high-rise buildings. The predicted number of floors matched our site visit observations for over 80% of the sampled buildings, with a mean difference of only

0.05 floors. This alignment is attributable to our process (described in the "Output Analysis" section above), which recalculates estimated floors only if the reported number of floors for the parcel are implausible given the building height. Relying on reported floors alone results in significant discrepancies for parcels with multiple buildings of varying height, but image-based floor estimates are often incorrect by one or two levels due to the wide variation in floor-to-floor height.

Our process slightly underestimated gross floor area on average in low-rise buildings, largely due to partial basements, whereas in mid- and high-rise buildings it significantly overestimated the area, primarily because we assumed the same building footprint area for all floors, which is often not the case. Developing techniques for extracting per-floor footprints or full 3D models of buildings would enable more-precise gross floor and façade area estimates.

Buildings with LiDAR data yielded more-accurate estimates for number of floors (underestimated by 0.08 floors on average) than image-based height methods (overestimated by 0.59 floors on average). LiDAR data is broadly available throughout the United States and yields more-precise results than measuring building height from oblique imagery; however, it tends to be updated infrequently whereas high-resolution aerial imagery is captured annually in most population centers. Having both the LiDAR and image-based height estimates created a more robust and reliable dataset from which to estimate building height than relying on either source alone. Future studies primarily requiring building geometry and not focused on newer construction may prefer to exclusively use LiDAR data, whereas applications focusing on recent buildings or estimating other characteristics like WWR and building materials would be best served by leveraging both when available. Recent research to develop predictive models capable of directly estimating the number of floors from imagery, 3D models, and demographic data rather than estimating geometrically from measured height has also shown potential to improve floor level predictions (Roy et al. 2023).

Analysis of Categorical Metrics

We compared site visit categorical data to determine the image analysis detection rate for rooftop configuration, solar panels, skylights, and visible exterior HVAC systems. The image analysis correctly identified the presence or absence of HVAC units for the majority of the 391 sites evaluated. Table 4 presents the model's relative performance for various system types.

Table 4: Summary of model performance for categorical building characteristics and equipment identification comparison between image analysis and site visits

Building characteristic	Overall building inventory mean	Site visit comparison results		
		Correct result	False positive	False negative
Rooftop material	70% flat	99%	1%	N/A
Solar panel	6% detected	96%	4%	0%
Skylights	45% detected	90%	7%	3%
Split system AC	17% detected	98%	1%	1%
Window AC	41% detected	91%	1%	8%
Packaged RTUs	4% detected	Insufficient site visits with these systems present to evaluate		
Minisplit heat pumps	3% detected			

Object detection worked best on HVAC units in higher resolution images, as these units can be quite small and difficult to detect. In some cases, particularly with split systems, the image analysis correctly identified HVAC units but did not attribute them to the building due to their distance from the building or ambiguity about which building they served.

Conclusions

The computer vision analysis Cadmus conducted provides valuable insights into NYS's multifamily building retrofit market and has demonstrated a viable approach for developing accurate building inventory data at a large scale. By extracting details of nearly every multifamily building across an entire state and making them accessible through an interactive data exploration dashboard, we have produced a dataset that can be used to size the market for novel retrofit solutions; identify specific candidate buildings; and support virtual audits and remote information gathering to prepare retrofit proposals that are more relevant, accurate, and effective. Our approach was particularly successful at disaggregating widely available property-level attributes to individual buildings and estimating building dimensions and structural configuration, including roof type and WWR. These metrics are directly relevant to targeting shell retrofit measures and generating building simulation model geometries, and they are useful proxies for energy efficiency factors that influence interior measures. For example, knowing the WWR and the ratio of façade area to floor area can help identify good candidates for daylight harvesting lighting controls, and understanding the total number of dwelling units in a building can help size the number of heat pump water heaters required to serve its residents. With this public resource with robust location data, researchers and market actors could overlay additional internal or public information including demographics, energy intensity, sales, and program participation data to produce novel market insights well beyond the scope of our project.

With further research and training, computer vision building analysis can have even broader applications. Additional training data and higher resolution oblique imagery would improve detection of HVAC system details and could make it possible to approximate system capacity based on overall system size and the diameter and quantity of condenser fans. Similarly, further training of our model to better identify solar panel system details could offer value to the solar industry, where computer vision is already being deployed in the single-family residential solar industry to generate customer leads and expedite shading analysis.

Computer vision holds significant promise for driving decarbonization efforts by providing manufacturers and stakeholders with data to make informed decisions at scale, while traditional site visits for acquiring building characteristic data remain important to understand the inner workings of a population. The feasibility of conducting similar studies in other jurisdictions depends largely upon the data sources available. While the specific scripts, training data, and models developed for this project are tailored to the building stock and data available in New York circa 2020 to 2022, the challenges and solutions we describe are broadly applicable. The growing proliferation of open-source machine learning toolkits and models and training datasets makes it possible for small teams of data scientists to quickly produce usable outputs; in fact, at a statewide scale we found data acquisition and management to be more challenging overall than model development. Various federal agencies produce free orthographic imagery and LiDAR data across the United States, but oblique imagery and more recent LiDAR and orthographic imagery are typically collected on behalf of individual states, counties, or municipalities. While pricing, access methods, and data formats can vary significantly between jurisdictions, some

form of oblique imagery and LiDAR data is available for most population centers nationwide. We expect that growing interest in combining remote sensing with machine learning may also drive new government initiatives or commercial offerings to streamline data aggregation, allowing future building science researchers to focus on developing useful models.

References

Chen, Q., Y. Zhang, X. Li, and P. Tao. 2021. "Extracting Rectified Building Footprints from Traditional Orthophotos: A New Workflow." *Sensors* 22 (1): 207–7. doi.org/10.3390/s22010207.

Harris, C. 2021. "Opaque Envelopes: Pathway to Building Energy Efficiency and Demand Flexibility: Key to a Low-Carbon, Sustainable Future." *OSTI OAI (U.S. Department of Energy Office of Scientific and Technical Information)*, September. doi.org/10.2172/1821413

Nelson, Hal and Hunter Johnson. 2022. "Data-Driven, Equity-Centered Energy Efficiency for Multifamily Complexes." In *Summer Study on Energy Efficiency in Buildings*. ACEEE. https://aceee2022.conferencespot.org/event-data/indexes.

Roy, Ellie, Maarten Pronk, Giorgio Agugiaro, and Hugo Ledoux. 2022. "Inferring the Number of Floors for Residential Buildings." *International Journal of Geographical Information Science* 37 (4): 938–62. https://doi.org/10.1080/13658816.2022.2160454.

Sáez, D. 2017. "Correcting Image Orientation Using Convolutional Neural Networks." Github.io. d4nst.github.io/2017/01/12/image-orientation/.

Singh, R., S. Fernandes, A. Prakash, P. Mathew, J. Granderson, C. Snaith, R. Pusapati, J. Prakash, A. Zakhor, R. Upadhyay, O. Gonen, H. Bergmann. 2022. "Scaling Building Energy Audits through Machine Learning Methods on Novel Drone Image Data." In *Summer Study on Energy Efficiency in Buildings*. ACEEE. eta-publications.lbl.gov/sites/default/files/scaling_building_energy_audits.pdf.

NOAA. 2024. "What Is LiDAR?" Noaa.gov. oceanservice.noaa.gov/facts/lidar.html.

V7 Labs. 2024. "What Is Computer Vision? [Basic Tasks & Techniques]." V7labs.com. https://www.v7labs.com/blog/what-is-computer-vision.

Wilson, J., and C. Williams. 2019. "Aerial Photography Ortho & Oblique Imagery." ncdor.gov/documents/files/wilson-aerial-photography-business-personal-property/open.

Zhang, Z. 2000. "A Flexible New Technique for Camera Calibration". In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. microsoft.com/en-us/research/publication/a-flexible-new-technique-for-camera-calibration/.