Deep Learning-Driven Green Building Facade Segmentation: Enhancing Sustainable Urban Development Through U-Net Architectures, Edge Detection, and Attention Mechanisms

Shruti Semwal
CSE Department, School Of Computing, DIT University

Garima Verma
School of Computing, DIT University

https://doi.org/10.5109/7407641

出版情報：Evergreen. 13 (1), pp.141-155, 2026-03. Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Japan
バージョン：

Deep Learning-Driven Green Building Facade Segmentation: Enhancing Sustainable Urban Development Through U-Net Architectures, Edge Detection, and Attention Mechanisms

Shruti Semwal[1], Garima Verma[2,*]

[1]CSE Department, School of Computing, DIT University, INDIA
[2]School of Computing, DIT University, INDIA

[*]Author to whom correspondence should be addressed
E-mail: garimaverma.research@gmail.com

(Received February 05, 2025; Revised September 09, 2025; Accepted February 17, 2026)

Abstract: This work aims to examine the function of deep learning (DL) in improving building facade segmentation to facilitate sustainable urban planning and green building (GB) assessment. The study used a large collection of open-source façade images with 51,731 architectural elements labeled. To address class imbalance, six methods for data augmentation were used. The study utilized a progressive model augmentation technique. First, the original dataset was used to train basic U-Net architecture. Second, Canny Edge Detection (CED) was included to improve the presentation of boundaries and structures. Third, an attention mechanism was added to improve feature selection and contextual learning. Lastly, two improvements were made to the architecture: a learnable edge branch for adaptive boundary modeling and a boundary-aware hybrid loss function to improve contour accuracy. Experimental results show that the proposed new framework performs better at segmentation, achieving an overall accuracy of 0.982 and improved border consistency. The combination of edge-guided learning and boundary optimization yields much better facade delineation than a regular U-Net-based model. Proposed identifying architectural parts makes facade assessment more reliable, which supports energy-efficient design analysis, sustainable retrofitting plans, and environmentally friendly urban development. The suggested approach provides a robust computational resource for architects, urban planners, and sustainability researchers assessing green infrastructure.

Keywords: Building segmentation; Canny edge detection; Deep learning; Green buildings; Sustainability; U-Net

1. Introduction

The building industry is starting to understand how important it is to use sustainable building methods to help the environment and make buildings last longer. GB has become an important way to reduce the negative effects that buildings have on the environment. These buildings put energy efficiency, resource conservation, and the use of eco-friendly materials and technologies first. These are all important for making cities more sustainable and protecting the environment[1)]. Their goal is to make the indoor environment better (IEQ), make people happier with their homes, and make cities more sustainable by using things like green roofs, facades, and new construction technology[2)]. GBs improve air quality, reduce the urban heat island effect, and save energy by adopting less common ways to heat and cool spaces[3)].

Studies have also become more focused on how GB design affects human health and IEQ. It is clear that green buildings not only make the air inside better, but they also make the health of the people who live there better. These buildings try to make the inside of the buildings healthier by using new technology and eco-friendly design features. It also improves the health and productivity of the people who live there. This part of the work shows that GBs have two benefits: they are excellent for the environment and make people far more comfortable and healthy[4)]. Green facades enhance urban environments, combat climate change, and improve aesthetics by incorporating plants into building designs, thereby promoting sustainable environmental management[5)].

Achieving sustainability goals requires greening of existing buildings, especially when renovating older buildings to improve their environmental performance. Adding green elements to existing buildings may significantly decrease energy use and carbon emissions. These initiatives are greatly aided by frameworks and

Cite: S. Semwal, G. Verma, "Deep Learning-Driven Green Building Facade Segmentation: Enhancing Sustainable Urban Development Through U-Net Architectures, Edge Detection, and Attention Mechanisms". Evergreen, 13 (01) 141-155 (2026). https://doi.org/10.5109/7407641.

evaluation programs for maximizing the green potential of older structures. Encouraging sustainable urban growth and achieving environmental goals requires thoroughly greening existing buildings[6]. Building sustainability is assessed using Green Building Rating Systems (GBRS), prioritizing social, economic, and environmental factors. Nevertheless, a balanced approach is necessary since present evaluations reveal asymmetry, with environmental concerns frequently overshadowing social and economic aspects[7].

Construction, operation, and maintenance are only a few of the life cycle stages that must be carefully considered when designing GB to rectify this imbalance. The optimization of these processes is greatly aided by intelligent technologies like artificial intelligence (AI) and the Internet of Things (IoT), which provide data-driven approaches for environmental monitoring, energy management, and adaptive control systems. These developments promote human engagement with building surroundings, lower energy usage, and improve thermal comfort[8]. DL algorithms are one such recent development that has significantly enhanced image processing and segmentation, especially semantic segmentation. Modern DL approaches, which provide improved capabilities for analyzing complicated pictures, are gradually replacing more conventional ways[8]. This progress also reaches the architectural field where building façade elements like doors, walls, and windows can now be automatically recognized using DL frameworks. These improvements make use of the hierarchical structures seen in DL models, combining the channels and spatial attention processes to improve identification performance[9]. Furthermore, a significant advancement in the field of architectural picture analysis has been made with the employment of creative loss functions designed specifically for these kinds of jobs, which have enhanced the detection and categorization of façade features[10]. Another improvement was made by developing multi-view U-Net models. These models use pre-trained convolutional neural networks (CNN) to extract features from multiple perspectives, achieving higher segmentation accuracy and finer-grained results[11]. High-resolution remote sensing imaging enhances urban green space delineation accuracy and efficiency by combining indicators and advanced processing techniques, addressing complex landscapes and vegetation borders in urban monitoring[12].

In parallel, advances in U-Net models and their improved variants have been studied recently for accurate building segmentation. These models have come a long way, but there are still issues, especially with intricate architectural designs. These results highlight the continued need for improved dataset quality and additional model optimization to improve segmentation performance and successfully handle complex scenarios[13]. Furthermore, new approaches based on the U-Net architecture have been developed because of recent developments in segmentation techniques, and they greatly improve object recognition in pictures. These advancements, fueled by advances in computer vision (CV), have enhanced uses for augmented reality, robotic navigation, and autonomous driving. Object classification and scene comprehension have been further enhanced using complex feature extraction techniques, such as different edge detection and feature extraction algorithms [14]. The efficacy of merging CED with other methods like adaptive thresholding, fuzzy clustering, and wavelet coefficient modification greatly improves segmentation accuracy and efficiency in managing complicated fracture patterns[15].

CNNs have proven successful in semantic segmentation tasks. U-Net outperforms other CNN models in accuracy, making it useful for classification, but still requires further work on interpretability and performance optimization[16]. Modern segmentation models include sophisticated methods to improve contextual awareness and feature extraction while maximizing computing capacity. These enhancements have proven successful across a range of testing scenarios, indicating possible advantages for applications with limited resources[17]. There are several significant obstacles to façade segmentation, such as the requirement for more accurate classification of different types of façades, enhancements to data quality, and the incorporation of GB principles into façade design. The suggested study focuses on improving preprocessing methods to enhance façade component distinction to overcome these difficulties. The main goal of this research work is to improve façade segmentation and aid novel contribution in the construction of sustainable and energy-efficient buildings.

The main author’s contributions are as follows:

- This study proposes a progressive DL framework for building facade segmentation aimed at supporting GB evaluation and sustainable urban planning.
- A structured multi-stage architectural enhancement strategy is introduced, starting from a baseline U-Net model, followed by edge-enhanced segmentation using CED, and further improved through the integration of an attention mechanism to strengthen contextual feature learning.
- A novel refinement is introduced through the incorporation of a learnable edge branch that enables adaptive boundary modelling, combined with a boundary-aware hybrid loss function to improve contour precision and structural consistency.
- A comprehensive dataset comprising 51,731 annotated architectural elements is curated and balanced using six augmentation techniques to address class imbalance and improve model generalization.

Cite: S. Semwal, G. Verma, "Deep Learning-Driven Green Building Facade Segmentation: Enhancing Sustainable Urban Development Through U-Net Architectures, Edge Detection, and Attention Mechanisms". Evergreen, 13 (01) 141-155 (2026). https://doi.org/10.5109/7407641.

- Extensive comparative experiments are conducted across multiple model variants and the proposed enhanced framework, demonstrating significant improvements in segmentation accuracy and boundary delineation performance.

Through the accomplishment of these goals, the study hopes to further façade segmentation methods and encourage the growth of more environmentally and energy-conscious construction practices.

There are other parts in the remaining portion of the document. In Section 2, relevant research is reviewed, and their shortcomings are highlighted. The methods for creating a pre-processed dataset are described in depth in Section 3. The suggested model, together with the techniques used, is presented in Section 4. An overview of the workflow and the performance measures that were employed for assessment is given in Section 5. The work is finally ended in Section 6, which summarizes the results and suggests possible avenues for further research.

2. Related Work

To learn about the present body of research, a thorough examination of relevant sources was conducted. This section summarizes the scholarly publications regarding the pertinent research.

As a sustainable architectural solution to urban heat, climate change, and environmental quality, GBs have gained popularity. Green facades can lower temperatures, increase energy efficiency, and improve urban aesthetics by adding plants to buildings[5)]. They can reduce air conditioning demand, improve air quality, boost biodiversity, and reduce urban heat islands. However, they have drawbacks like maintenance, installation costs, and climate-related issues. Further research and data are needed to fully understand the impact of green facades on urban temperatures and energy savings. According to the study[8)], implementing intelligent systems like automated controls and smart grids can maximize building performance by lowering expenses and energy consumption while enhancing the surroundings and occupant comfort. Adoption might, however, be hampered by issues including data scarcity, algorithm stability, and hefty upfront expenses. The study pointed out that there weren't enough in-depth case studies on how these technologies were really used in the real world and recommended more investigation into cost-benefit analysis and overcoming integration obstacles.

In a study[13)], the IST Building Dataset, created using Pleiades satellite pictures, covered 34 km² and had 2100 tiles with 7604 classified structures. U-Net and U-Net++ architectures were trained and tested using the dataset. Intersection over Union (IoU) metrics were used to assess segmentation performance. Post-processing methods like noise reduction and boundary smoothing improved the findings. U-Net design performed better in segmentation performance, but Slum and other building types had lower IoU ratings. U-Net++ model produced longer training durations but did not significantly outperform U-Net. Another research[18)] looked at Jordan's construction industry's knowledge of, use of, and obstacles to sustainable building methods. It was discovered that although stakeholders were aware of the principles of green construction, there was little application of these concepts because of a lack of government regulation, financial constraints, and a lack of awareness, particularly among (medium-sized enterprises) SMEs. The study suggests that more government support, including rules and incentives, is needed to encourage GB adoption. Education and public awareness are crucial. However, due to a small sample size and international norms, future research should focus on Jordanian issues.

It is often acknowledged that mixed land use and urban greenery have a major role in determining the standard of living in metropolitan areas[19)]. A study in Switzerland found that urban greenery positively impacts life satisfaction, particularly for people over 65. However, younger individuals prefer mixed-use areas over green spaces. The study suggests age-specific urban planning and green space distribution to optimize life satisfaction across age groups, but concerns remain about generalizability and the need for longitudinal analysis. Recent advancements in semantic image segmentation, particularly with Deep Neural Networks (DNNs), have significantly improved efficiency and accuracy in sectors like medical imaging and intelligent transportation systems. Traditional techniques like edge detection and thresholding have also contributed to this progress[20)]. The study categorized semantic image segmentation into DNN-based and traditional methods, revealing DNN-based techniques outperform conventional ones. Combining Conditional Random Fields (CRF) with Fully Convolutional Networks (FCNs) improves spatial consistency. However, issues like overfitting, data dependence, and computing complexity persist, and research gaps include inadequate unsupervised techniques investigation.

Building facade segmentation using DL techniques improves urban planning and administration by enhancing precision in architectural details, integrating attention mechanisms and unlabeled data in training procedures. The study[21)] improved building facade segmentation using DL techniques by adding a multi-headed attention mechanism to the PointNet++ network. The model produced 88.2% accuracy and 77.1% mean intersection-over-union (mIoU), a 1% increase in accuracy and 3.3% improvement in mIoU. However, the study's limited dataset and increasing complexity suggest further research and optimization are needed for real-world applications. A study[22)] presented an automated DL technique for window segmentation using a modified SOLOv2 algorithm. The system, which combined squeeze-and-excitation (SE) and bi-directional feature pyramid network (BiFPN), outperformed conventional techniques with a mean absolute error of 2.9% and a 93% mean average precision (mAP). However, the model's complexity grew, requiring further work.

Parallelly, the article[23)] evaluated various façade detection algorithms, including Faster R-CNN and YOLO, using original facade images. The most accurate model was YOLOv4, while YOLOv5, with 85% accuracy and 0.79 mAP, was the quickest. To better handle the complexities of urban environment, the encoder-decoder nature of the U-Net architecture has made it popular for use in semantic segmentation tasks, such as the study of building facades. Researchers used the U-Net model for residential building facade segmentation to support GB initiatives[24)]. The model successfully split building facades, detecting windows, walls, and doors. However, the model's accuracy was affected by the intricacy of input photos and the dependence on high-quality photos. The study suggested U-Net could help with sustainable urban planning but suggested real-time data handling and integration of other data sources are needed for better segmentation accuracy. While conventional approaches are generally resource-intensive and highly reliant on on-site measurements, to overcome these drawbacks, researchers[25)] developed a computational method to estimate the Green Area Factor (GAF) by combining DL techniques with satellite remote sensing. This method has a high relationship between estimated and real GAF levels and accurately detects and categorizes green features using satellite images and DL. It is less expensive than traditional on-site measurements and suitable for various urban situations. However, it faces performance variations, dependence on ground truth data, and high processing costs.

Meanwhile, in another study[26)], the Concatenated Residual Attention U-Net (CRAUNet) model improved semantic segmentation of urban green areas by integrating residual structures with channel attention processes. Tested on high-resolution satellite photos of Shenzhen City, it showed significant improvements in visual quality and accuracy. However, its complexity and lack of cross-dataset validation limit its applicability in various urban settings and requires further research for real-time surveillance applications. Window line recognition is crucial for urban planning and 3D building modelling. An enhanced stacked hourglass network improved 3D reconstruction accuracy by extracting local and global characteristics from building façade photos[27)]. However, it relied on well-annotated data and high-quality photos, and more study is needed to prove its generalizability and extend its application to other structural aspects.

Furthermore, a study[28)] improved virtual building models by identifying windows in façade photos, using a sliding window detector and cascaded classifier. The system was tested on the Düsseldorf subway project, but faced challenges in accurate opening ratios, false positives, and store window detection. More research is needed to improve accuracy across different facade types. Also, the study[29)] proposed a DL-based method for semantic categorization to segment façade photos down to the pixel level. Three designs, MULTIFACSEGNET, SEPARABLE, and COMPATIBILITY, were introduced

Table 1: Review summary for some most relevant studies

Year and Study	Technology Used	Data	Parameters	Challenges
2024,[13)]	Unet with boundary smoothing and noise reduction	IST Building Dataset, created using Pleiades satellite pictures	Accuracy	No data balancing and only binary classification
2023, [21)]	Pointnet++ with unlabelled data processing	A small building façade dataset constructed using on-board LiDAR point clouds.	Accuracy	Small dataset is used for the study, after labelling balancing may be applied for more appropriate results
2023, [22)]	Building Facades for windows segmentation	SOLOV2 with bidirectional feature pyramid and noise reduction	Accuracy	Single type of class may limit the scope of research
2021, [24)]	Satellite images of residential street view building facades	FCN, CNN	Accuracy, F1-Score	Data balancing may affect the results towards better prediction.
2024, [25)]	Satellite orthophotos of urban areas	Deep neural network for semantic segmentation	Accuracy	Conventional on-site measurements remain more resource-intensive, suggesting that computational estimation may require occasional ground-truth verification.
2023, [27)]	Building facades	Stacked hourglass network with point line extraction module	Accuracy	Study is limited for only one class detection.

to address issues with thin façade features, multi-label assignment per pixel, and data distortion. The models showed increased accuracy, particularly in recognizing important façade components like windows. A recent study[30] employed DL-based semantic segmentation to create the GreenRoof collection of aerial photos and masks, identifying suitable rooftops for greening across 15 German cities. The methodology prioritized areas based on vegetation cover, thermal environment, and building density, revealing that over 20% of roof space was amenable to retrofitting. This research presented a scalable, multi-city framework for urban sustainability planning, distinguishing itself from earlier single-city or LiDAR-dependent approaches. Table 1 shows the comparison of some most relevant existing studies based on some parameters such as- model used, dataset source, etc.

3. Methods and Materials

3.1. Description of Dataset

The study employed a well curated open source dataset that was gathered from several sources and places to facilitate facade segmentation tasks. It has 606 photos in total, tagged with 51,731 elements in 12 different categories which are background, facade, window, door, cornice, sill, balcony, deco, molding, pillar and shop[31]. Subsets of CMP-World, ECP-World, ZuBuD, and Prague were used to source the photos, which varied in terms of resolution, location, and acquisition date. The 213 images in the CMP-Prague subset were captured in 2007 in Prague, Czech Republic, with a Canon G2 camera. Each image has a resolution of around 6 MPx. The CMP-World subset consists of 99 photos taken with various cameras at resolutions of about 6 MPx in locations such as Bratislava, Buenos Aires, and London between 2007 and 2009. The 177 images in the ZuBuD subset were captured in 2003 at a resolution of around 0.3 MPx while in Zurich, Switzerland. Finally, the ECP-World subset comprises 177 photographs collected in 2010 from locales such as Barcelona, Greece, and the USA, with a resolution of around 0.6 MPx. Rectangular bounding boxes are used to annotate each image, designating items like windows, doors, facades, and ornamental components. The dataset is split into two categories: CMP-base, which concentrates on planar facades with more regularity, and CMP-extended, which includes photos of irregular, non-planar facades or severe occlusions. Figure 1 shows the sample images of dataset.

Fig. 1: Sample images from the dataset

3.2. Canny Edge Detection

The dataset images were subjected to the CED method, which is a popular technique in CV for identifying edges. It was created by John F. Canny in 1986 and a widely used edge detection method. There are four main phases in the process: First, as shown by Equation 1, the input image is smoothed using a Gaussian filter to minimize noise. The Sobel operator then uses the kernels given in Equation 2 and 3 to compute the gradients S_x in the *x*-direction and S_y in the *y*-direction to find edges. Equation 1 displays the equation for picture smoothing, with γ standing for the standard deviation:

$$C(x,y)=\frac{1}{2\pi\gamma^2}e^{-\frac{x^2+y^2}{2\gamma^2}} \tag{1}$$

$$S_x=\begin{pmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{pmatrix} \tag{2}$$

$$S_y=\begin{pmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{pmatrix} \tag{3}$$

The remaining pixels are then eliminated and non-maximum suppression is used to keep local maxima value in the gradient direction. Equation 4 and Equation 5, respectively, are used to determine the gradient's magnitude and direction. A double-thresholding approach is used to remove noise-induced edges by deleting pixels below a threshold and keeping pixels above a higher threshold. Lastly, hysteresis tracking creates the final edge map by using two thresholds, d_h and d_l, to categorize pixels as strong, weak, or non-edges[32]. Equation 4 illustrates the gradient's magnitude, which reflects edge intensity:

$$C_m(x,y)=\sqrt{S_x{}^2+S_y{}^2} \tag{4}$$

Equation 5 provides the gradient's direction, which indicates the edge orientation:

$$\psi(x,y)=\tan^{-1}\left(\frac{S_y}{S_x}\right) \tag{5}$$

3.3. Pre-Processing

3.3.1. Binary Mask Splitting

In image segmentation tasks, binary mask splitting (BMS) is an essential process. It comprises segmenting an image into areas according to predetermined standards (such as object vs background) and then giving each pixel a label (0 or 1) to indicate the category to which it belongs[33]. The primary four mathematical operations are used in this approach to discern between various visual components. BMS involves four operations-

- Thresholding – An intensity value of a pixel is compared with a threshold, and based on the results, the pixel is assigned to either one class or other. For a pixel at position (i, j) in the image Equation 6 defines the thresholding.

$$BM(i, j) = \begin{cases} 1, \text{if}\, (I(i,j)) \geq T \\ 0, \text{if}\, (I(i,j)) < T \end{cases} \qquad (6)$$

Where, $I(i, j)$ is an intensity of the pixel at position (i, j), $BM(i, j)$ is binary mask at same position and T is a threshold value. if value is >=T, means it is a part of image and <T, means part of background.

- Connected component labeling – After thresholding, identification of connected regions in binary mask is done. This step groups pixels that are connected into separate regions. Two pixels $p1(i_1, j_1)$ and $p2(i_2, j_2)$ are connected if $\sqrt{(i_1 - i_2)^2 + (j_1 - j_2)^2} \leq d$, where d is a distance threshold.
- Morphological operation – this is applied to refine the binary mask by removing noise and smoothing object boundaries. Equation 7 and 8 show the erosion and dilation operation under morphology.

$$I_1 \ominus B = \{z \mid (B)_Z \subseteq I_1\} \qquad (7)$$

$$I_1 \oplus B = \{z \mid (B)_Z \cap I_1 \neq \phi\} \qquad (8)$$

Where, I_1 is binary image and B is the structuring element.

- Binary mask multiplications – finally binary mask is multiplied with original image or another mask to isolate the object of interest from the background. Equation 9 shows the binary masked image.

$$I_{masked}(i, j) = I(i, j) \bullet BM(i, j) \qquad (9)$$

3.3.2. Data Balancing

The data balance is addressed by implementing strategies for bespoke data augmentation. Various methods such as rotation, flipping, scaling, and normalizing help to increase the resilience of the model as discussed below. Research has demonstrated that data augmentation is essential for getting around dataset constraints, enhancing model accuracy, and boosting performance in a variety of DL applications[32]. Six data augmentation techniques are applied to augment as well as balance the dataset- resizing, rotation, translation, flipping, Gaussian noise and contrast. The model uses real-time augmentation during training on dataset batches, improving its generalization and broadening the training set variety. This approach overcomes class imbalance difficulties and enhances model performance.

3.4. Attention Mechanism

It is a way for models to focus on the most important features of the input data. The technique is inspired by the human visual system where human pay attention to only selected parts of the environment. By concentrating on the most relevant areas of a picture, these mechanisms help models become more accurate and efficient. From early uses in RNNs to more modern self-attention models, which serve as the basis for visual transformers, attention processes in DL have progressed. Temporal, spatial, channel, and branch attention are the four main attentional subtypes that are employed in visual activities[34,35]. There are mainly four steps involved in the attention mechanism in proposed model-

- Creation of attention map- for input let feature map $F \in R^{H \times W \times C}$ where H –height, W-width and C-number of channels. After this 1×1 convolution is used to reduce the number of channels in the map and new map is created ($F' \in R^{H \times W \times C}$), Equation 10 defines the convolution conversion.

$$F' = Conv_{1 \times 1}(F) \qquad (10)$$

- Sigmoid activation – now attention map is passed through a sigmoid activation function to constrain its values between 0 and 1. Equation 11 defines the sigmoid function and Equation 12 shows the attention map.

$$A(x) = \frac{1}{1 + e^{-x}} \qquad (11)$$

$$A = \eta(F') \in [0,1]^{H \times W \times 1} \qquad (12)$$

Where A, is an attention map that assigns a weight between 0 and 1 to each spatial location in the feature map.

- Element-wise multiplication – now the attention map A is multiplied elementwise with the unsampled feature map. Let $Fu \in R^{H \times W \times C}$ is unsampled feature map from the previous layer. Equation 13 defines the attention weighted feature map, generated from the element-wise multiplication ($\odot$) of unsampled map and attention map.

Cite: S. Semwal, G. Verma, "Deep Learning-Driven Green Building Facade Segmentation: Enhancing Sustainable Urban Development Through U-Net Architectures, Edge Detection, and Attention Mechanisms". Evergreen, 13 (01) 141-155 (2026). https://doi.org/10.5109/7407641.

$$Fa = Fu \odot A \qquad (13)$$

- Skip connection and feature concatenation – after applying attention to unsampled feature map, the resulting attention generated in the form of weighted feature map $Fs \in R^{H\times W\times C}$ from downsampling path. After this concat operation is applied (Equation 14) to keep all relevant information (original and attention) for final segmentation prediction.

$$F_{concat} = Concat(Fa, Fs) \in R^{H\times W\times C} \qquad (14)$$

4. Proposed Model

This section discusses the proposed progressive DL architecture for facade segmentation. It was created to support GB evaluation and long-term urban planning. The framework is built using a structured multi-layer architectural enhancement technique. It starts with a baseline U-Net and then adds edge priors, attention mechanisms, adaptive boundary modeling, and hybrid loss optimization one at a time. Figure 2 shows the Basic network pipeline and Figure 3 illustrates the layer architecture pipeline utilized to create U-net architecture with CED and attention mechanism. Figure 4 shows the final suggested model pipeline which is used to improvise the previous architecture. To systematically examine performance improvements, four model configurations are created. Model-I is the initial U-Net that was trained on the original facade dataset. Model-II improves the baseline by adding CED as an edge prior, which makes it easier to depict the structure of architectural boundaries. Model-III further improves segmentation performance by adding attention-based skip connections to enhance contextual feature learning and suppress irrelevant activations.

The final proposed model adds two novel components on top of these improvements: (i) learnable edge branch for adaptive boundary modeling, and (ii) a boundary-aware hybrid loss function that combines focal, Dice, and boundary losses to optimize both region-level accuracy and contour precision at the same time. This last setup completes the edge-guided attention U-Net system.

4.1. Learnable Edge Branch Formulation

To enable adaptive boundary modelling, a learnable edge branch is introduced. Unlike fixed CED edges, this branch predicts edge probabilities directly from decoder features. The predicted edge map enhances structural awareness by refining segmentation features through edge-guided fusion. Let $F \in R^{H\times W\times C}$ be the decoder feature map. The predicted edge map is computed as per Equation 15, where, W_e convolution kernel of edge branch and σ is sigmoid activation.

$$E_p = \sigma(W_e * F) \qquad (15)$$

The predicted edge map refines segmentation features using $F' = F \odot (1 + E_p)$, This enables adaptive boundary emphasis during decoding[36].

4.2. Hybrid Loss Function

The boundary loss penalizes misclassification near object contours by incorporating distance transform information. This improves structural precision and reduces boundary discontinuities. To jointly optimize region accuracy and boundary precision, a hybrid loss function is employed using Equation 22.

$$L_{total} = \lambda_1 L_{focal} + \lambda_2 L_{dice} + \lambda_3 L_{boundary} \qquad (16)$$

Where, dice loss $L_{dice} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i}$ boundary loss is $L_{boundary} = \frac{1}{N}\sum_i p_i . D(g_i)$ and focal loss $F_{Loss}(p_t) = -\alpha_t (1-p_t)^{\gamma} \log(p_t)$.

4.3. Optimizer

The Adam optimizer with a learning rate of 1e-4 was chosen for its improved convergence consistency. Gradient clipping with a clipvalue of 1.0 was also implemented to prevent exploding gradients and ensure stable optimization in deeper layers. This optimizer maintains two running averages for each parameter- the first moment (mean of gradients) and second moment (uncentered variance of gradients)[37]. Using a loss function defined in Equation 16, Adam updates the parameters iteratively using following steps-

- The computation of gradient defined by Equation 17

$$grad_t = \nabla(\theta_t) FLoss(\theta_t, Data_{train}) \qquad (17)$$

$grad_t$ - Gradient of loss function with respect to the parameters at the time step t.

- Exponential moving averages of gradient called as first moment defined by Equation 18, where mf_t is an estimate of first moment.

$$mf_t = \beta_1 * mf_{\{t-1\}} + (1-\beta_1) * grad_t \qquad (18)$$

- Exponential moving averages of squared gradient called as second moment defined by Equation 19, where vs_t is an estimate of second moment.

Cite: S. Semwal, G. Verma, "Deep Learning-Driven Green Building Facade Segmentation: Enhancing Sustainable Urban Development Through U-Net Architectures, Edge Detection, and Attention Mechanisms". Evergreen, 13 (01) 141-155 (2026). https://doi.org/10.5109/7407641.

$$vs_t = \beta_2 * vs_{\{t-1\}} + (1-\beta_2) * (grad_t)^2 \tag{19}$$

- Bias correction is required for the optimization since mf_t and vs_t are initialized to 0, therefore they are biased towards 0 during initial steps. To correct this Equation 20 and 21 are used for bias correction.

$$\overline{mf_t} = \frac{mf_t}{(1-\beta_1{}^t)} \tag{20}$$

$$\overline{vs_t} = \frac{vs_t}{(1-\beta_2{}^t)} \tag{21}$$

- Finally, the parameters are updated using Equation 22, where η is learning rate, ω is a small constant.

$$\theta_{\{t+1\}} = \theta_t - \frac{\eta}{(\sqrt{vs_t}) + \omega} \tag{22}$$

The Figure 5 illustrates the list of parameters used to define the models and algorithm-1 demonstrates all the important stages that were taken to define final model, which is final derivation of layered framework. The Figure 6 shows the new dataset sample received after CED operation.

Algorithm-1 Final Model

```
Input: CMP FACADE Dataset
Output: Segmented output with
classification of 5 classes
Begin
Loading the dataset
Pre-processing:
Resize images to (256 × 256).
Normalize pixel intensity values.
Apply CED with thresholds (100,
200).
Generate edge map
Concatenate edge map with RGB image
along channel dimension:
    Input shape → (256, 256, 4).
Convert annotations into multi-class
binary masks.
Apply data augmentation
Train validation Split
Model Creation
Initialize the U-Net Model, set the
input_shape to (256, 256, 4).
Build the Model Layers:
Convolutional Blocks: Create 5
encoding blocks (downsampling) and 4
upsampling blocks with skip-
connections.
Bottleneck Layer: It consists of the
deepest convolution layer with 1024
filters.
Create Upsampling and Attention
Block:
Define up_attention_block to combine
the upsampled features with the
corresponding encoder features using
skip connections and attention
mechanisms.
Construct the U-Net
Add a final convolution layer with
softmax activation to generate
segmentation masks.
Model Training and Evaluation
Define hybrid loss function using
Eq-16.
Train model for 132 epochs with
gradient clipping
Prediction:
After training and evaluation, do
predictions of masks on the test set
and visualize them to compare them
with the ground truth to assess the
model performance.
Labelling of top 5 features:
Convert the predicted labels into
human-readable form for 5 relevant
labels using a label-to-class
mapping.
End
```

5. Results and Discussions

The model is simulated using Python 3.x with Jupyter Notebook on a google colab platform with 16GB of RAM and a 4GB GPU32). The dataset utilized is free source and has 606 photographs in total. It has 51,731 tags in 12 different categories. The study uses layer framework and final model. The entire simulation procedure consisted of three phases: initially, layered framework was trained and validated with the original image dataset, and performance metrics were documented. In the second phase, the new pre-processed dataset created using CED followed by binary masking, and augmentation to train and validate layered framework. Finally, proposed modified boundary aware model was trained and validated with the new pre-processed dataset. An empirical approach established training epochs. Initial experiments showed that the model's performance levelled off at around 130 epochs, with no further improvements. So, all models were trained for 132 epochs to make sure they converged and did not overfit. phase-II outcomes found better than phase-1, but phase-III did the best of all. The 20% of the dataset used to validate the models. The loss and accuracy metrics were generated for Phase-I, training loss and accuracy value is found as 0.004 and 0.969 respectively, while validation loss and accuracy values are 0.1034 and 0.812 respectively. After training of phase-II, the metrics of training loss and accuracy values obtained are 0.1042 and 0.98 respectively, while validation loss and accuracy values are 0.1612 and 0.941. Now for the phase-III, training and validation was done and values of training loss and accuracy were found as 0.0074 and 0.981 respectively while validation loss and accuracy were found as 0.0235 and 0.982. From the values of the metrics it can be easily analysed that validation accuracy found in phase-1 was quite less but has improved in phase-II when model was trained on new dataset. In phase-3 final prosed model has outperformed in comparison to other phases, where validation accuracy has changed drastically to 0.982. The plots of loss and accuracy for all the phase are shown in

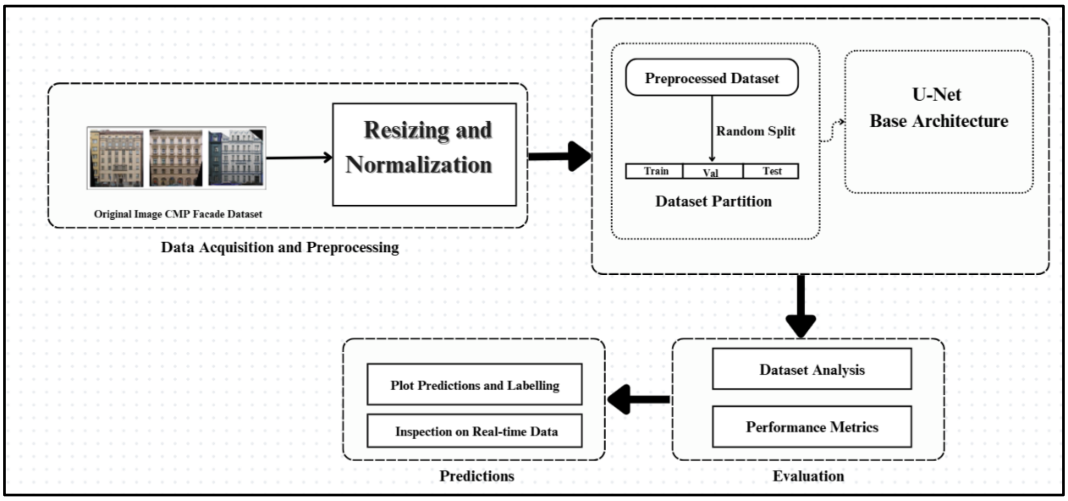

Original Image CMP Facade Dataset
Resizing and Normalization
Data Acquisition and Preprocessing
Preprocessed Dataset
Random Split
Train
Val
Test
Dataset Partition
U-Net
Base Architecture
Dataset Analysis
Performance Metrics
Evaluation
Plot Predictions and Labelling
Inspection on Real-time Data
Predictions

Fig. 2: Pipeline of Basic Network

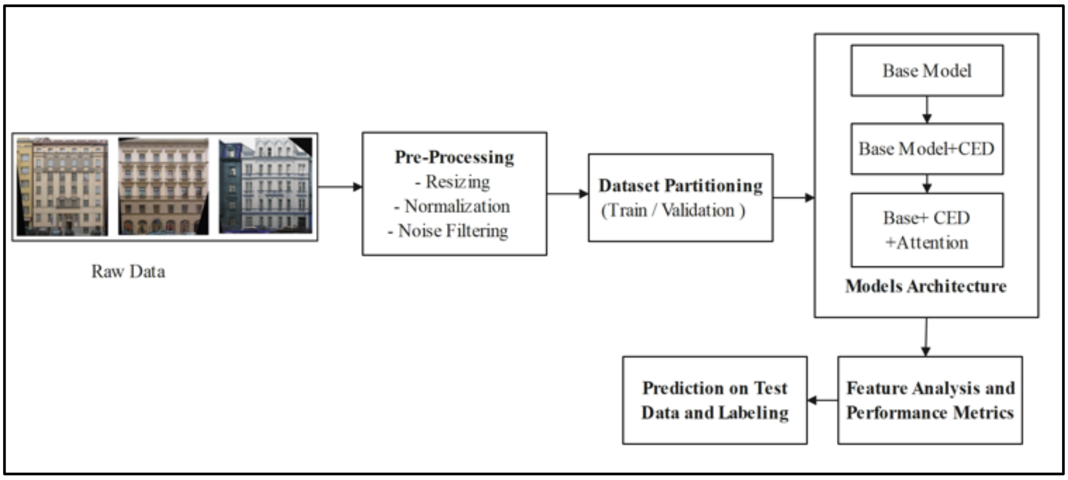
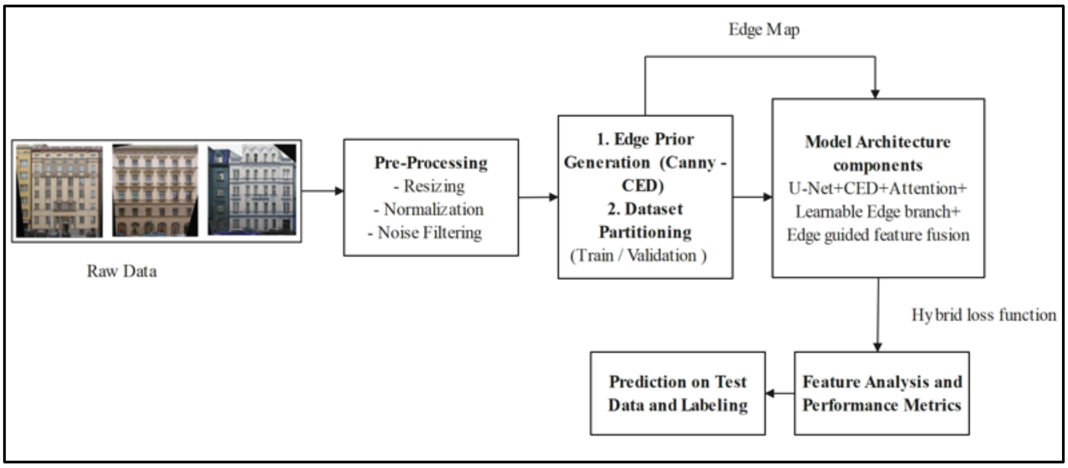
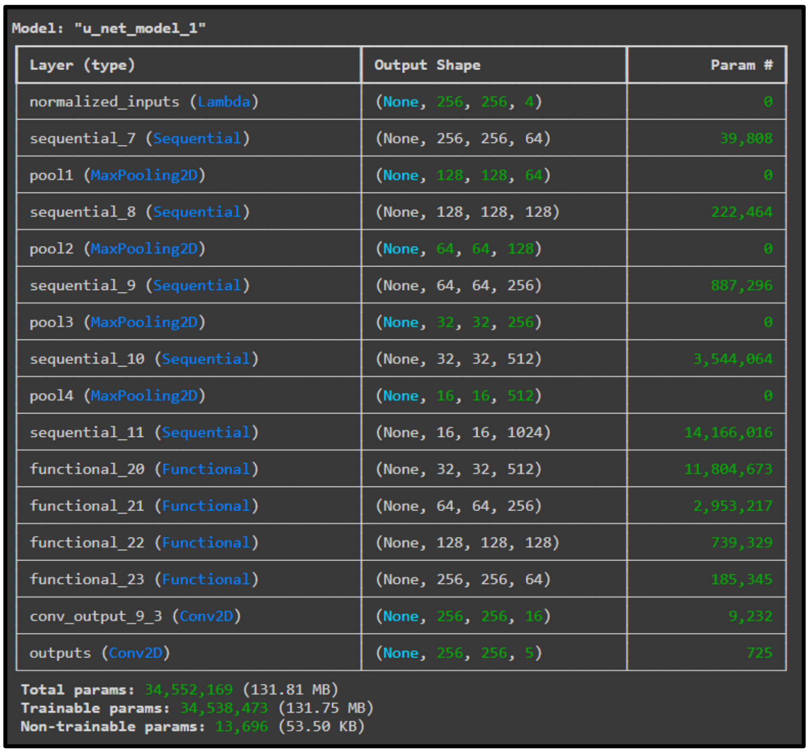


Fig. 3: Pipeline of Layered Framework



Cite: S. Semwal, G. Verma, "Deep Learning-Driven Green Building Facade Segmentation: Enhancing Sustainable Urban Development Through U-Net Architectures, Edge Detection, and Attention Mechanisms". Evergreen, 13 (01) 141-155 (2026). https://doi.org/10.5109/7407641.

Fig. 4: Pipeline of Proposed Boundary Aware Model



```
Model: "u_net_model_1"
┌──────────────────────────────────┬─────────────────────────┬────────────┐
│ Layer (type)                     │ Output Shape            │    Param # │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ normalized_inputs (Lambda)       │ (None, 256, 256, 4)     │          0 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ sequential_7 (Sequential)        │ (None, 256, 256, 64)    │     39,808 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ pool1 (MaxPooling2D)             │ (None, 128, 128, 64)    │          0 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ sequential_8 (Sequential)        │ (None, 128, 128, 128)   │    222,464 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ pool2 (MaxPooling2D)             │ (None, 64, 64, 128)     │          0 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ sequential_9 (Sequential)        │ (None, 64, 64, 256)     │    887,296 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ pool3 (MaxPooling2D)             │ (None, 32, 32, 256)     │          0 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ sequential_10 (Sequential)       │ (None, 32, 32, 512)     │  3,544,064 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ pool4 (MaxPooling2D)             │ (None, 16, 16, 512)     │          0 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ sequential_11 (Sequential)       │ (None, 16, 16, 1024)    │ 14,166,016 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ functional_20 (Functional)       │ (None, 32, 32, 512)     │ 11,804,673 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ functional_21 (Functional)       │ (None, 64, 64, 256)     │  2,953,217 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ functional_22 (Functional)       │ (None, 128, 128, 128)   │    739,329 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ functional_23 (Functional)       │ (None, 256, 256, 64)    │    185,345 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ conv_output_9_3 (Conv2D)         │ (None, 256, 256, 16)    │      9,232 │
├──────────────────────────────────┼─────────────────────────┼────────────┤
│ outputs (Conv2D)                 │ (None, 256, 256, 5)     │        725 │
└──────────────────────────────────┴─────────────────────────┴────────────┘
 Total params: 34,552,169 (131.81 MB)
 Trainable params: 34,538,473 (131.75 MB)
 Non-trainable params: 13,696 (53.50 KB)
```

Fig. 5: List of Parameters used to define the Model

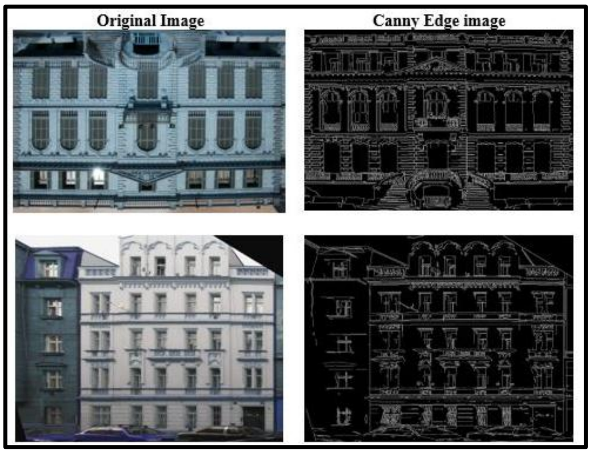


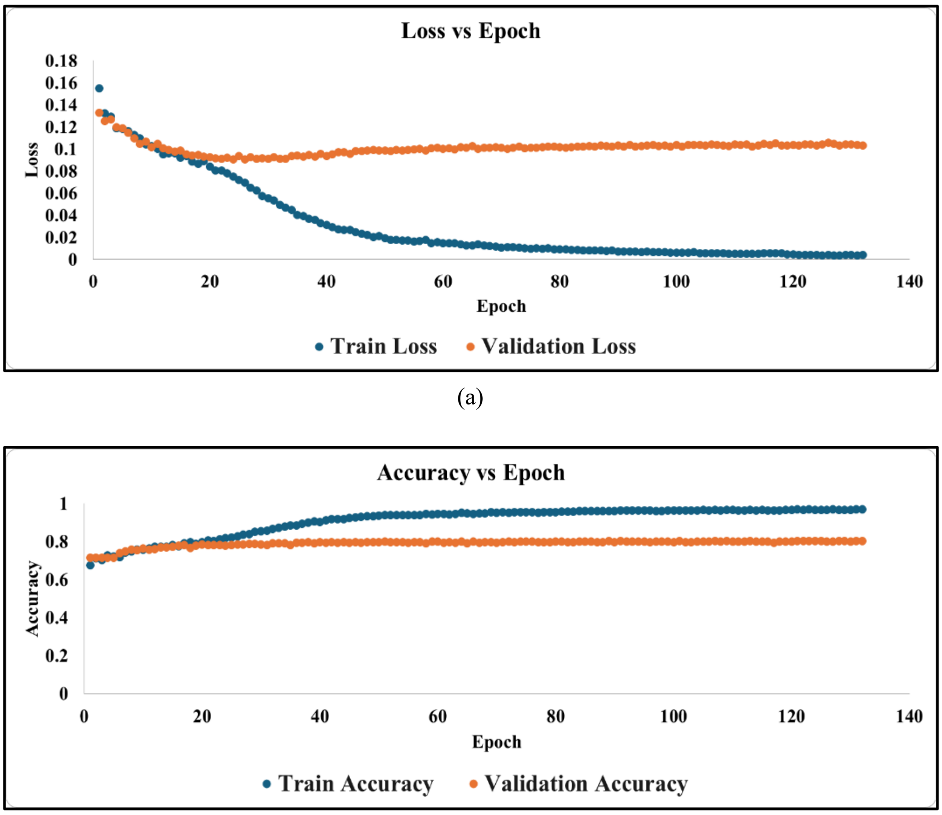
Fig. 6: Sample of images from new dataset after CED

Figure 7 (a), (b), Figure 8 (a), (b), and Figure 9 (a) and (b). There were 12 classes in the dataset to begin with, but only 5 were used in the study: windows, balcony, pillar, molding, and sill. The choice of these classes were used because they are the most relevant for GB assessments and they affect how long buildings last and how much energy they use. Windows let in more natural light, so you do not need as much artificial light and plants can grow better inside. Plants and trees can grow on balconies in cities, which makes the air pure and soothing. Pillar can support green walls and vertical gardens, which are significant features of GB. Molding makes the edges of plant layers stiff, which makes it easier to attach them to green walls or roofs. The sill keeps water from getting in, which helps the plants in the green areas around it stay healthy and is in line with the goals of green architecture. Figure 10 displays the results for different sample pictures. The window, balcony, pillar, molding, and sill are shown in blue, light green, red, dark orange, and light orange, respectively.

The models suggested in this study are also compared to other cutting-edge models to assess how well they operate. Table-2 compares the proposed models to other current models based on certain essential metrics, such as accuracy, F1-Score, mIoU, etc. This study prioritizes accuracy and F1-score for façade segmentation, surpassing the commonly applied mIoU segmentation metric. When it comes to class distribution and model generalization, accuracy and F1-score are better ways to determine how effectively the model classifies items in general and how well it balances accuracy and recall. The mIoU numbers are still used so that they can be compared to earlier research.

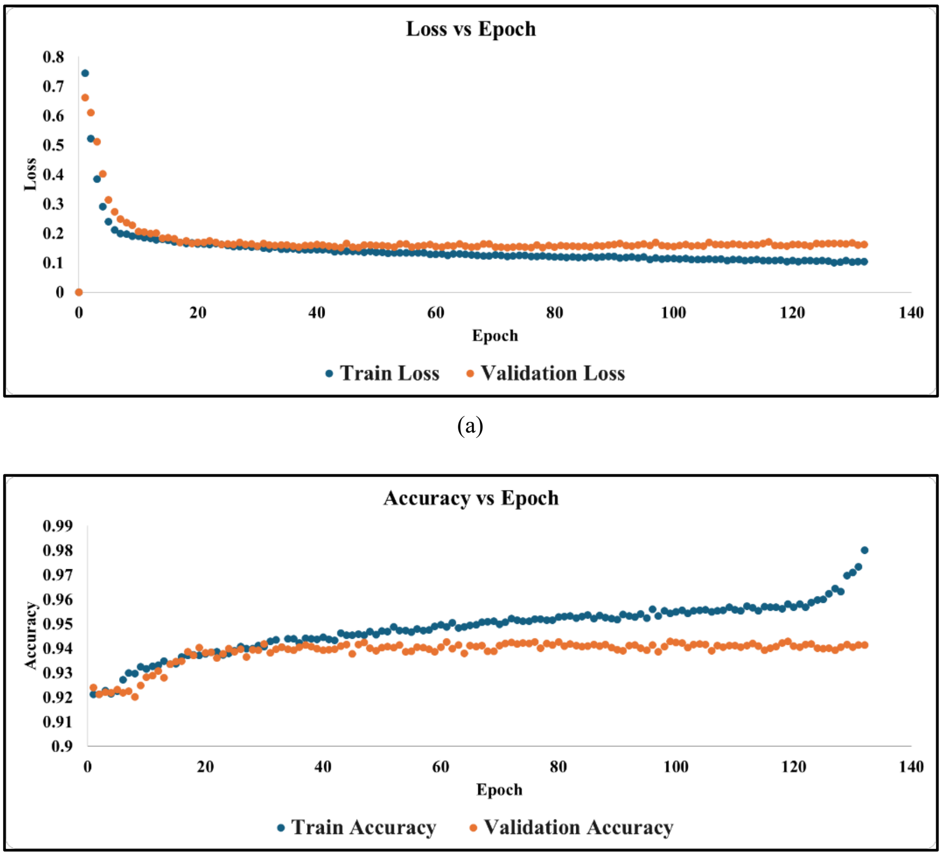
It is visible from the table that the proposed model did far better than other existing models when it came to accuracy and F1-score. Phase-1 has an accuracy of 80.1%, Phase-II has an accuracy of 94.1%, and Phase-III has the best accuracy at 98.2%. This is also the average F1-score, which suggests that the model is handling the balancing of classes well. After comparing the models, the best-performing proposed model tested using real-time photos obtained from local locations and some from the Google image database. The test dataset is then prepared by applying CED, binary mask, and augmentation. The

Phase-III applied on the images and predictions are done. Figure 11 shows the results of test set after labelling in the form of classes. Although the predicted classes from the



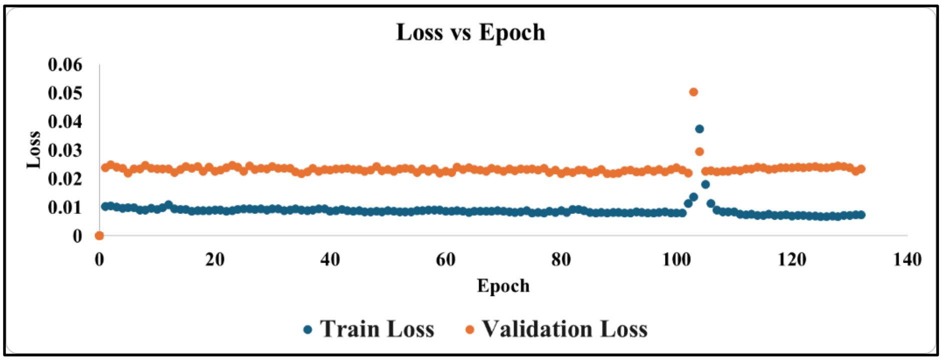
(b)

Fig. 7: (a) and (b). Plots of Phase-I Loss and accuracy

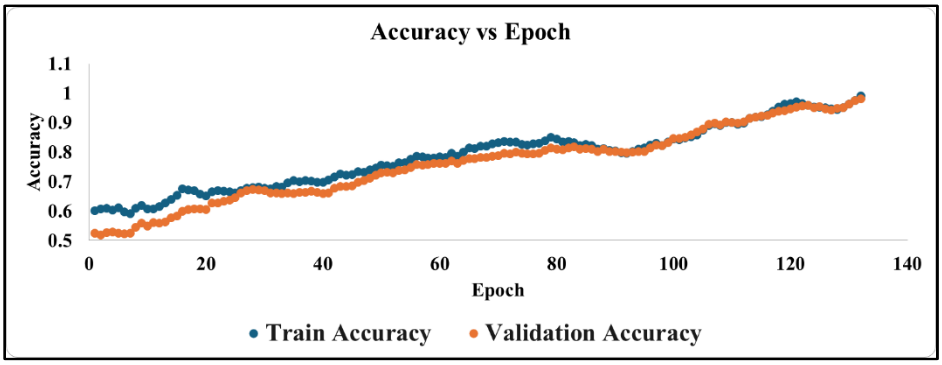


(b)

Fig. 8: (a) and (b). Plots of Phase-II Loss and accuracy



(a)



(b)

Fig. 9: (a) and (b). Plots of Phase-III Loss and accuracy

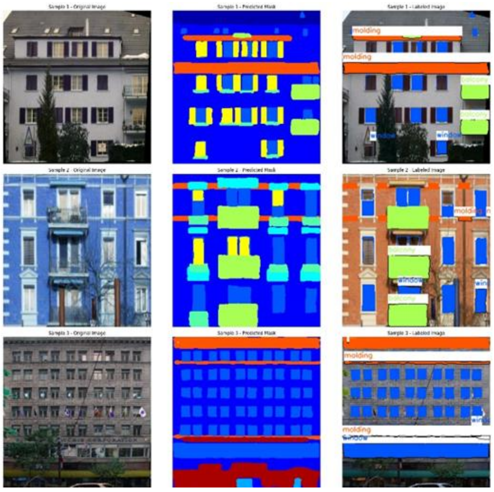

Sample 1 - Original Image
Sample 1 - Predicted Mask
Sample 1 - Labeled Image
molding
molding
balcony
balcony
window
window
Sample 2 - Original Image
Sample 2 - Predicted Mask
Sample 2 - Labeled Image
molding
balcony
window
Sample 3 - Original Image
Sample 3 - Predicted Mask
Sample 3 - Labeled Image
molding
molding

Fig. 10: Predicted masks labeling with sample images

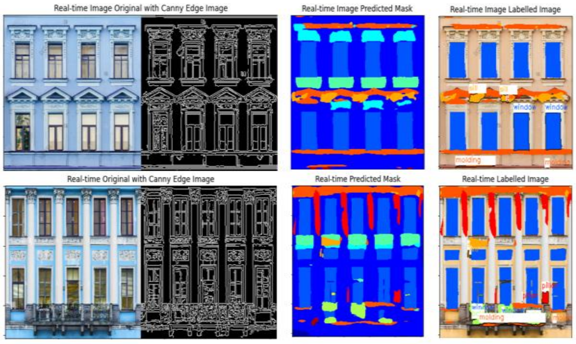

Real-time Image Original with Canny Edge Image
Real-time Image Predicted Mask
Real-time Image Labelled Image
window
window
molding
Real-time Original with Canny Edge Image
Real-time Predicted Mask
Real-time Labelled Image
pillar
pillar
window
window
molding

Fig. 11: Test results with prediction and labeling

Table 2: Comparison with some existing studies

Study	Method used	Data Balancing	Validation accuracy	F1 Score (average)	Validation mIoU
Neuhausen et al.[28]	**Soft Cascaded Classifier** with a sliding window detector	No	-	average detection (recall) rate : 85% with a false positive rate of 2%	-
Femiani et al. [29]	Multifacsegnet, separable, and compatibility architectures	Yes	0.94,0.96,0.95 respectively	0.73,0.76,0.77 respectively	-
Gu & Choo, [30]	SegNet Model with semantic segmentation	No	0.87	-	-
Proposed Phase-I	U-Net with Original images	No	0.802	0.81	0.78
Proposed Phase-II	U-Net with Canny edge detection, Binary mask and Data augmentation	Yes	0.941	0.934	0.81
Proposed Phase-III	U-Net with Canny edge detection, Binary mask, Data augmentation and attention mechanism	Yes	0.9817	0.981	0.87

model were 12. But later, attention mechanism labelling was done only for relevant five classes – windows, balcony, pillar, moulding and sill.

6. Conclusion

This study introduces a progressive-edge, attention-guided U-Net framework for facade segmentation, facilitating GB assessment and sustainable urban planning. The assessment adhered to a systematic three-phase approach. The original dataset was used to train and test the baseline U-Net model in Phase I. In Phase II, a pre-processed dataset created utilizing CED, binary masking, and augmentation methods to improve the representation of structural features. In Phase-III, the final suggested model was trained and tested. It included attention-based skip connections, a learnable edge branch, and a boundary-aware hybrid loss function. Phase-I had a validation accuracy of 0.812. After edge-based preprocessing, Phase-II's accuracy went up significantly to 0.941. The final suggested model in Phase-III achieved the highest validation accuracy of 0.982, along with better loss convergence and boundary consistency. The drop in validation loss from 0.1034 (Phase-I) to 0.0235 (Phase-III) shows that generalization and structural accuracy have improved. The results show that adding edge priors, contextual attention mechanisms, adaptive edge modeling, and hybrid boundary-aware optimization to the baseline architecture significantly improves facade segmentation performance. The structural enhancement technique works because the improvements keep becoming better over time. Accurate segmentation of architectural features enables reliable evaluation at the facade level, supporting energy-efficient retrofitting, assessing the long-term viability of structures, and making data-driven judgments in city planning. The proposed framework is a useful computational tool for architects, planners, and sustainability researchers, as it clarifies boundaries and strengthens class balance. Ultimately, this work shows how advanced deep learning methods can help analyze infrastructure in ways that are good for the environment and help cities grow in ways that are good for the long term.

References

1) L.F. Anzagira, D. Duah, and E. Badu. "A conceptual framework for the uptake of the green building concept in Ghana," Scientific African, 6, e00191, (2019). https://www.sciencedirect.com/science/article/pii/S2468227619307525

2) G. Nariman, M. Samari, and M. W. M. Shafiei. "Green buildings impacts on occupants' health and productivity," Journal of Applied Sciences Research 8(8), 4235-4241, (2012). https://www.cabidigitallibrary.org/doi/full/10.5555/20133032511

3) D. Oliveira Santos, T. Danily, F. António Leal Pacheco, and L. F. Sanches Fernandes. "A systematic analysis on the efficiency and sustainability of green facades and roofs." Science of The Total Environment, 173107, (2024). https://www.sciencedirect.com/science/article/pii/S0048969724032546.

4) J. G. Allen, P. MacNaughton, J. G. C. Laurent, S. S. Flanigan, E. S. Eitland, & J. D. Spengler, "Green buildings and health, " Current Environmental Health Reports, 2(3), 250–258, (2015). https://link.springer.com/article/10.1007/s40572-015-0063-y

5) S. M. Sheweka, & N. M. Mohamed, "Green facades as a new sustainable approach towards climate change," Energy Procedia, 18, 507-520, (2012). https://www.sciencedirect.com/science/article/pii/S1876610212008326

6) B. C. M. Leung, "Greening existing buildings [GEB] strategies," Energy Reports, 4, 159-206, (2018). https://www.sciencedirect.com/science/article/pii/S2352484717301841

7) M. Braulio-Gonzalo, A. Jorge-Ortiz, & M. D. Bovea, " How are indicators in Green Building Rating Systems addressing sustainability dimensions and life cycle frameworks in residential buildings?," Environmental Impact Assessment Review, 95, 106793, (2022). https://www.sciencedirect.com/science/article/pii/S0195925522000592

8) S. Yin, J. Wu, J. Zhao, M. Nogueira, & J. Lloret, "Green buildings: requirements, features, life cycle, and relevant intelligent technologies," Internet of Things and Cyber-Physical Systems, (2024). https://www.sciencedirect.com/science/article/pii/S2667345224000099

9) Y. Guo, Y. Liu, T. Georgiou, & M. S. Lew, "A review of semantic segmentation using deep neural networks," International journal of multimedia information retrieval, 7, 87-93, (2018). https://link.springer.com/article/10.1007/s13735-017-0141-z

10) G. Zhang, Y. Pan, & L. Zhang, "Deep learning for detecting building façade elements from images considering prior knowledge," Automation in Construction, 133, 104016, (2022). https://www.sciencedirect.com/science/article/abs/pii/S0926580521004672

11) S. El Hajjar, H. Kassem, F. Abdallah, & H. Omrani, "Enhancing building segmentation by deep multiview classification for advancing sustainable urban development," Journal of Building Engineering, 83, 108421, (2024). https://www.sciencedirect.com/science/article/abs/pii/S2352710223026049

12) Y. Cheng, W. Wang, Z. Ren, Y. Zhao, Y. Liao, Y. Ge, ... & C. Zhang, "Multi-scale Feature Fusion and Transformer Network for urban green space segmentation from high-resolution remote sensing images," International Journal of Applied Earth Observation and Geoinformation, 124, 103514, (2023). https://www.sciencedirect.com/science/article/pii/S1569843223003382

13) B. Amirgan, & A. Erener, "Semantic segmentation of satellite images with different building types using deep learning methods," Remote Sensing Applications: Society and Environment, 34, 101176, (2024). https://www.sciencedirect.com/science/article/abs/pii/S2352938524000405

14) N. A. Almujally, B. R. Chughtai, N. A. Mudawi, A. Alazeb, A. Algarni, H. A. Alzahrani, & J. Park, "UNet Based on Multi-Object Segmentation and Convolution Neural Network for Object Recognition," Computers, Materials & Continua, 80(1), (2024). https://www.researchgate.net/profile/Bisma-Riaz-Chughtai/publication/382381258

15) U. A. Nnolim, "Automated crack segmentation via saturation channel thresholding, area classification and fusion of modified level set segmentation with Canny edge detection," Heliyon, 6(12), (2020). https://www.cell.com/heliyon/fulltext/S2405-8440(20)32591-3

16) O. Pešek, L. Brodský, L. Halounová, M. Landa, & T. Bouček, "Convolutional neural networks for urban green areas semantic segmentation on Sentinel-2 data," Remote Sensing Applications: Society and Environment, 36, 101238, (2024). https://www.sciencedirect.com/science/article/pii/S2352938524001022

17) Y. Zuo, & W. Li, "An Improved UNet Lightweight Network for Semantic Segmentation of Weed Images in Corn Fields," Computers, Materials & Continua, 79(3), (2024). https://cdn.techscience.cn/files/cmc/2024/TSP_CMC-79-3/TSP_CMC_49805/TSP_CMC_49805.pdf

18) H. Jaradat, O. A. M. Alshboul, I. M. Obeidat, & M. K. Zoubi, "Green building, carbon emission, and environmental sustainability of construction industry in Jordan: Awareness, actions and barriers," Ain Shams Engineering Journal, 15 (2), 102441, (2024). https://www.sciencedirect.com/science/article/pii/S2090447923003301

19) S. Bahr, “The relationship between urban greenery, mixed land use and life satisfaction: An examination using remote sensing data and deep learning,” Landscape and Urban Planning, 251, 105174, (2024). https://www.sciencedirect.com/science/article/pii/S0169204624001737

20) X. Liu, Z. Deng, & Y. Yang, “Recent progress in semantic image segmentation,” Artificial Intelligence Review, 52, 1089-1106.6, (2019). https://link.springer.com/article/10.1007/s10462-018-9641-3

21) J. Yang, Y. Li, X. Pan, & J. Zang, “Research on Building Facade Feature Information Extraction,” Academic Journal of Computing & Information Science, 6(7), 72-83, (2023). https://www.francis-press.com/uploads/papers/vDHkbwkN5nZqqMvnckHYsBB1CnWTVtLjs6GXvjqu.pdf

22) Y. Lu, W. Wei, P. Li, T. Zhong, Y. Nong, & X. Shi, “A deep learning method for building façade parsing utilizing improved SOLOv2 instance segmentation,” Energy and Buildings, 295, 113275, (2023). https://www.sciencedirect.com/science/article/abs/pii/S0378778823005054

23) G. Sezen, M. Cakir, M. E. Atik, & Z. Duran, “Deep learning-based door and window detection from building façade,” The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences, 43, 315-320, (2022). https://isprs-archives.copernicus.org/articles/XLIII-B4-2022/315/2022/

24) M. Dai, W. O. Ward, G. Meyers, D. D. Tingley, & M. Mayfield, “Residential building facade segmentation in the urban environment,” Building and Environment, 199, 107921, (2021). https://www.sciencedirect.com/science/article/abs/pii/S0360132321003243

25) G. A. Rahaman, M. Längkvist, & A. Loutfi, “Deep learning based automated estimation of urban green space index from satellite image: A case study,” Urban Forestry & Urban Greening, 97, 128373, (2024). https://www.sciencedirect.com/science/article/pii/S1618866724001717

26) G. Men, G. He, & G. Wang, “Concatenated Residual Attention UNet for Semantic Segmentation of Urban Green Space,” Forests 2021, 12, 1441. Urban Forests and Landscape Ecology, 73, (2021). https://core.ac.uk/download/pdf/534901813.pdf#page=82

27) [27] F. Yang, Y. Zhang, D. Jiao, K. Xu, D. Wang & X. Wang, “Detecting window line using an improved stacked hourglass network based on new real-world building façade dataset,” Open Geosciences, 15 (1), 20220476, (2023). https://www.degruyter.com/document/doi/10.1515/geo-2022-0476/html.

28) M. Neuhausen, M. Obel, A. Martin, P. Mark, & M. König, “Window detection in facade images for risk assessment in tunneling,” Visualization in Engineering, 6, 1-16, (2018). https://link.springer.com/article/10.1186/s40327-018-0062-9

29) J. Femiani, W. R. Para, N. Mitra, & P. Wonka, “Facade segmentation in the wild,” (2018). arXiv preprint arXiv:1805.08634. https://arxiv.org/abs/1805.08634

30) Li Q, Taubenboeck H, Zhu XX. Identification of the potential for roof greening using remote sensing and deep learning. Cities. 2025 Apr 1;159:105782.

31) H. Gu, & S. Choo, “Method for Constructing a Façade Dataset through Deep Learning-Based Automatic Image Labeling,” Applied Sciences,. 12(15), 7570, (2022). https://www.mdpi.com/2076-3417/12/15/7570

32) G. Verma, “Leveraging smart image processing techniques for early detection of foot ulcers using a deep learning network. Polish Journal of Radiology,” 89, e368, (2024). https://pmc.ncbi.nlm.nih.gov/articles/PMC11321030/

33) TF Representations and Masking. Accessed on Aug (2024), https://source-separation.github.io/tutorial/basics/tf_and_masking.html.

34) M. H. Guo, T. X. Xu, J. J. Liu, Z. N. Liu, P. T. Jiang, T. J. Mu, ... & S. M. Hu, “Attention mechanisms in computer vision: A survey,” Computational visual media, 8 (3) 331-368, (2022). https://link.springer.com/article/10.1007/S41095-022-0271-Y

35) A. A. Umoh, C. N. Nwasike, O. A Tula, O. O. Adekoya, & J. O. Gidiagba, "A Review of Smart Green Building Technologies: Investigating The Integration And Impact of AI And IOT in Sustainable Building Designs," Computer Science & IT Research Journal, 5(1), 141-165, (2024). https://fepbl.com/index.php/csitrj/article/view/715

36) Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. InProceedings of the IEEE international conference on computer vision 2017 (pp. 2980-2988).

37) S. El Hajjar, H. Kassem, F. Abdallah, & H. Omrani, "Enhancing building segmentation by deep multiview classification for advancing sustainable urban development. Journal of Building Engineering," 83, 108421, (2024). https://www.sciencedirect.com/science/article/abs/pii/S2352710223026049