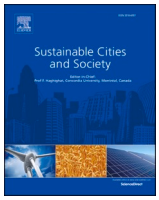
Sustainable Cities and Society 126 (2025) 106414

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs



AdvUNet3+: Segmentation of city-scale heterogeneous building façade materials from street view images

Jing Kai Daniel Tan [a,b], Rui Zhu [a,*], Jie Song [a], Zheng Qin [a], Yanqing Xu [c], Yumin Chen [d]

[a] Institute of High-Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, Singapore 138632, Singapore
[b] School of Information and IT, Temasek Polytechnic, 21 Tampines Ave 1, Singapore 529757, Singapore
[c] School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
[d] School of Resource and Environmental Sciences, Wuhan University, 129 Luoyu Road, Wuhan 430079, China

ARTICLE INFO

Keywords:
Deep learning
UNet3+
Façade material segmentation
Street view images
GIScience

ABSTRACT

City-scale three-dimensional (3D) building models associated with heterogeneous façade materials and albedos are crucial for investigating light pollution, analyzing urban heat island effects, and estimating solar potential distribution. Segmenting façade materials from street view images (SVIs) and identifying albedos is an effective approach to collect such thematic information in large urban areas, which, however, is challenged by complex street landscapes filled with confusable and small textures. Here, we propose a scalable framework, which develops a deep learning model, called AdvUNet3+, to accurately segment building façade materials from SVIs, and perform spatial analysis to project the identified façade material information onto 3D building models. We innovatively (i) incorporate multi-scale SVIs into the encoder backbone that balances the trade-off between high-detail and broader contextual awareness, and (ii) integrate the Dual Attention Module and Atrous Spatial Pyramid Pooling at the bottleneck that enhances spatial and channel relationships in feature maps and capture multi-scale contexts consistently. After transfer learning, we strategically integrated the multi-class and binary segmentations to facilitate façade identification, using our customized 800 SVIs with refined façade labels (i.e., paint, metal, brick, glass, clay, and rare material) in central Singapore. Although challenged by small, difficult, and non-mixed façade samples in the 360-degree panoramic SVIs, the mIoU remains at 52.9 %, with Precision, Recall, F1-Score, and Accuracy equaling 62.8 %, 77.2 %, 60.4 %, and 87.6 %, respectively. The successful segmentation of >24,000 SVIs in a small urban area greatly enriched 3D building information, indicating a generalizable framework that can be applied to any other city.

1. Background

Three-dimensional (3D) building models are one of the most important geospatial datasets that have been widely used in city science (Ying et al., 2019, 2020; Yan et al., 2023). Particularly, building surface information (e.g., materials, textures, colors, and albedos) is crucial for a variety of studies. For example, Cao et al. (2021) evaluated 33 studies that investigated different planning strategies to reduce the urban heat island (UHI) effect using 3D models, and it was found that buildings, green areas, and pavements contributed the most to the UHI effect. Albedo information on city-scale building façades is also an important attribute for 3D building models (Calcabrini et al., 2019; Jakubiec & Reinhart, 2013; Li et al., 2016). Due to high albedos of large-area glass equipped on building façades, daytime light pollution is increasingly getting severe in high-density urban areas, which negatively affects human health, the ecological environment, and energy use among an array of problems (Zhang et al., 2024). Zhu et al. (2019, 2020) estimated solar irradiation distribution on 3D urban surfaces, incorporating direct, diffuse, and reflective irradiation into a 3D building model that has been used for estimating solar photovoltaic potential in cities. However, these studies lacked spatially heterogeneous albedo information on façades that can generate multiple reflections on 3D urban surfaces. To address this problem, they either ignored reflective irradiation (Walch et al., 2020; Park et al., 2021; Assouline et al., 2017) or assumed that building surface materials and albedos were homogenous, which can cause large uncertainty in the estimated results. Therefore, 3D city models associated with accurate and heterogeneous material and albedo information are urgently needed.

* Corresponding author.
E-mail address: zhur@ihpc.a-star.edu.sg (R. Zhu).

https://doi.org/10.1016/j.scs.2025.106414
Received 29 October 2024; Received in revised form 21 March 2025; Accepted 27 April 2025
Available online 28 April 2025

A variety of methods have been developed to collect material and albedo data of façades to advance 3D models, which, however, face considerable drawbacks. For instance, using traditional land surveying to measure building façades and identify various materials is extraordinarily accurate, but is costly and time-consuming (Mishra et al., 2024). Large-scale semantic segmentation of rooftop areas and materials from high-resolution satellite imagery is another effective way to obtain building surface information. However, it cannot obtain scenic information about the rest partitions of building surfaces from a 3D standpoint, which is necessary when accounting for studies closely related to façades, such as modelling the distribution of multi-surface solar radiation on façades (Sánchez & Izard, 2015; Boccalatte et al., 2020).

Street View Images (SVIs), with the obvious advantages of wide coverage in entire cities, easy accessibility, and abundant building information, have been used to address the challenges (Xing et al., 2023; Fan & Biljecki, 2024). There are two key groups of studies related to the use of SVIs. The first are those that develop deep learning (DL) models, such as GLNet (Lin et al., 2020) and adaptive multi-scale dual attention network (Wang et al., 2021), to segment objects in SVIs. The second is studies that compare the performance of various state-of-the-art (SOTA) models at segmenting objects in SVIs (Minaee et al., 2022). These DL models can effectively distinguish façade materials even though (i) they are of similar spectral characteristics, which is more challenging than identifying façade components with distinct shapes, and (ii) they have visual impediments like blurring varied distances from the viewpoint and unwanted obstacles.

However, these studies did not use multi-scale inputs to enhance feature representation in existing SOTA models, which neglected the possible benefits multi-scale inputs could introduce, including improved generalization to object sizes and preservation of global and local information. Alternatively, some studies impose assumptions to decrease the details of façade material representation, such as assuming each façade has two types of materials at most (Xu et al., 2023). We observe that the challenge of accurately detecting and segmenting a variety of building façade materials remains. Therefore, the major objective of this study is to efficiently obtain city-scale heterogeneous façade albedos, by developing a research framework that can comprise data collection, data preparation, DL-based building façade segmentation from SVIs, and projection of façade albedos onto 3D building models.

Subsequent sections of the paper are organized as follows. Section 2 reviews related work, enlightening us to propose an ensemble strategy using a modified UNet3+ network architecture to achieve accurate segmentation of building façade materials in SVIs. Section 3 introduces the methodology and assumptions of the experiment. Section 4 assesses the performance of the model across datasets and in comparison, with the bench line UNet variants whilst conclusions and discussions are in Section 5.

2. Literature review

For complex tasks such as façade material identification, the above challenges become important considerations for the creation of an advanced façade segmentation model. It has been emphasized that the varying distances between the buildings and the camera, necessitate an ideal composition of image dimensions, i.e., width, depth, and resolution. For example, Xu et al. (2023) have proposed an effective method of combining multi-scale inference with object contextual representation to balance the trade-offs between the demand for details and contextual comprehension capability on large objects like buildings, achieving novel results. Meanwhile, several methods have excelled in different fields but have not been introduced to building façade segmentation. A notable example is the MDAN-UNet proposed by Liu et al. (2020), which has combined both multi-scale inference, multi-scale feature aggregation, and dual attention on Optical Coherence Tomography images to achieve noteworthy segmentation performance.

The quintessential factor, however, is the need to balance segmentation performance amongst the façades themselves while ensuring the remaining classes (non-façades) do not get compromised. Training a model on only relevant classes may not guarantee an accurate representation of the background, which is still necessary when projecting predictions onto a 3D model. Conversely, training the model in all classes introduces the possibility of the model placing too much focus on the learning of non-façade classes. Hence, this "compensates" for a dip in performance for façade material classes.

This study will test the feasibility and performance of a multi-scale variant of the proposed model against other variants of UNet on the Cityscapes dataset (Cordts et al., 2016), which contains a variety of street-view instances and classes. Following this, we will implement transfer learning using the best model to carry its prerequisite knowledge of street view classes in Cityscapes over to the new task, whilst multiplicatively combining the outputs of the model trained on binary segmentation of buildings and multi-class segmentation of façades. Lastly, we will use 5-fold cross-validation to obtain the best model.

2.1. Street view and material recognition datasets

Street view datasets like WHU-Urban3D (Han et al., 2024) are popular for the semantic segmentation of urban scenes. Another example is Cityscapes, a large-scale dataset that is often used to benchmark models, containing 3475 annotated images of urban SVIs from 50 cities and 19 different classes of dynamic objects (Cordts et al., 2016). However, their labels comprise an array of street-view objects and do not have specific façade material information associated with buildings. Conversely, popular material recognition datasets, like CUReT (Dana et al. 1999) and OpenSurfaces (Bell et al., 2013), lack very important characteristics unique to a street-view scene like varying distances from the viewpoints and blurring. This makes them ineffective at representing façades in SVIs. Therefore, this study will create a new SVI dataset containing detailed labels for façade materials to suit the task of segmenting façade materials in SVIs.

To achieve that, we plan to use a dataset containing 24,219 panorama images in two central regions in Singapore (i.e., Bishan and Toa Payoh districts) and a subsample of 400 images in each district consisting of six refined façade classes to train our proposed model, after transfer learning from Cityscapes. The chosen areas have a diversity of building structures and thus an equally diverse range of façade materials, comprising Housing Development Board (HDB) flats, high-rise condominiums, complexes, landed properties (e.g., terrace houses), etc.

2.2. *Applications of deep learning models in segmenting building façades in SVIs*

Recently, Xu et al. (2023) proposed a novel architecture that uses attention modules and multi-scale object contextual representation to weight contextual information to produce noteworthy segmentation results and performance metrics. The key addition of multi-scale inputs allowed us to handle the trade-off between larger receptive fields and complex details at higher resolutions, which was the specific cause behind the model's success. Conjunctively, SOTA models constantly have their effectiveness at segmenting urban street view scenes evaluated via surveys (Minaee et al., 2022). One example of a renowned SOTA model is the UNet model, which has inspired various UNet-based architectures that have demonstrated promising performance when evaluated on the Cityscapes dataset (Kazerouni et al., 2021; Liu et al., 2024). This is because UNet uses plain skip connections that integrate feature maps of stark differences from the encoder and decoder layer (Ronneberger et al., 2015). UNet++ further improves on the base UNet by including nested dense skip connections to facilitate smoother gradient flow during the backpropagation of the loss (Zhou et al., 2018). Moreover, a recent variant of UNet called UNet3+ replaces the dense skip connections with full-scale skip connections, which concatenate smaller and identical size feature maps from encoders and larger feature

maps from decoders, proving far more effective at capturing both fine detail and coarse semantics than both UNet and UNet++ (Huang et al., 2020).

2.3. Limitations of existing deep learning models

However, existing models face limitations in segmenting heterogeneous building façade materials. Traditional models such as UNet and UNet++ struggle due to their reliance on plain skip connections and local receptive fields, which limit their ability to distinguish materials with subtle texture variations. The absence of attention mechanisms prevents them from effectively emphasizing critical regions, leading to errors in differentiating materials with similar spectral properties, such as glass reflections and transparent surfaces. Transformer-based models capture long-range dependencies but introduce challenges such as high computational costs and scale inconsistency, often resulting in over-segmentation of repetitive patterns or under-segmentation of fine façade details. Lastly, SOTA approaches, such as the multi-scale attention-based model by Xu et al. (2023) and MDAN-UNet by Liu et al. (2020), improve contextual understanding. Still, they face trade-offs in feature refinement and struggle with materials exhibiting high intra-class variance. Additionally, their reliance on computationally expensive attention mechanisms limits their scalability for large-scale urban applications.

2.4. Potential enhancements in existing deep learning models

To improve façade material segmentation, we propose AdvUNet3+, an enhanced variant of UNet3+ that integrates multi-scale inputs, attention mechanisms, and optimized feature fusion to balance fine-detail preservation with broader contextual awareness. Unlike MDAN-UNet (Liu et al., 2020), which applies multi-scale aggregation within UNet++ but lacks a fully integrated multi-scale encoder, AdvUNet3+ ensures seamless multi-scale feature fusion throughout the network. It also addresses the limitations of the model proposed by Xu et al. (2023), which prioritizes contextual weighting but lacks direct spatial-channel optimization, by incorporating the Dual Attention Module (DAM) (Fu et al., 2019; Ji et al., 2023) to enhance spatial and channel relationships and Atrous Spatial Pyramid Pooling (ASPP) (Chen et al., 2018, 2022, Sun et al., 2023) to extract multi-scale contextual features. Additionally, full-scale skip connections help retain high-resolution details while lightweight attention mechanisms maintain computational efficiency. By combining these elements, AdvUNet3+ improves segmentation accuracy for complex façade compositions while remaining suitable for large-scale urban analysis and 3D modelling applications.

2.5. Contributions

This study is innovative in three aspects. First, the proposed framework can accurately segment city-scale building façade materials from panoramic SVIs, and efficiently project the façade material information onto the 3D building model. Second, we proposed a DL model that leverages the combinatory effect of multi-scale inputs and abstract bottleneck operations to segment small and confusable textures in SVIs. Third, this study creates a new SVI dataset that contains (i) accurate labels of building façade materials across various types of buildings (i.e., flats, complexes, historical sites) and (ii) panoramic metadata including geographical coordinates, azimuth, and angle of elevation that can be used together with 3D building models.

3. Methodology

3.1. Research framework and workflow

Our study proposed a cohesive research framework with three core components, including the collection of data, cleaning and labeling of data, and a three-phased training procedure. The entire workflow is presented in Fig. 1. Module 1 represents data collection, where panoramic images within our study area and their corresponding metadata are downloaded using the app "Street View Download 360″. Module 2 represents data cleaning and preparation, where only panoramic images with façades are kept making our dataset, and a small sub-sample is reserved and labeled for training and validation. Module 3 depicts the training process, consisting of three interconnected steps. Firstly, transfer learning trains the model on Cityscapes with many classes, and then uses the learned weights on our constructed dataset to accurately predict façade materials in urban SVIs. Secondly, inspired by Huang et al. (2020), two separate models for binary and multiclass segmentation are trained. The binary output is multiplied with the multiclass output, zeroing out incorrect building labels meant to be background, thereby improving segmentation accuracy. Thirdly the framework effectively combines information from the panoramic metadata with the model predictions to locate the exact geographical positioning of panoramas, identify matching buildings in the vicinity, and project façade material and albedo information onto 3D building models.

Particularly, Fig. 2 illustrates the full process of how the models are trained and their outputs are combined. In contemporary segmentation models, the choice between the inclusion of the background class has been highly debated. Conceptually, the inclusion of the background class entails that the model deliberately learns the background class during training and possibly neglects the performance of other classes of higher importance. Conversely, doing the opposite by predicting the remaining classes and specifying an ignored index in the loss function, can cause the model to mistakenly predict the background class as something else given the lack of a penalty.

3.2. Development of AdvUNet3+

3.2.1. Model structure

In this paper, we propose a multi-scale input encoder-decoder structure called AdvUNet3+ that is augmented by ASPP and DAM to enhance the focus on fine yet important spatial details from high-level feature maps (Fig. 3). Specifically, the model will be trained and assessed together with other UNet variants to determine the validity of this approach. First, the model integrates multi-scale inputs into the encoder layers of UNet3+ to mitigate the loss of spatial information via constant down-sampling operations. Next, the feature map produced from the deepest encoder layer is entered into both the DAM and ASPP to capture spatial relationships and multi-scale contextual information simultaneously. The outputs from each module are then concatenated, and the concatenated feature map then passes through a 1×1 convolution layer and weighted summation with the original input to preserve spatial information. Lastly, the summed output ascends a decoder structure that uses full-scale skip connections to learn precise details and coarse contexts from both same to higher-level encoders and lower-level decoders respectively. Consequently, the model can better capture nuanced visual patterns in façades and street view classes together with coarser details common in backgrounds or classes that span large areas of the input image. The benefits of each addition will be further elaborated on in their subsections.

The primary innovation of AdvUNet3+ lies in its ability to integrate multi-scale inputs into the encoder layers, which is responsible for extracting hierarchical features from the input image by progressively reducing spatial resolution while capturing increasingly abstract and high-level representations. These multi-scale inputs reduce the loss of spatial information that typically occurs due to repeated sampling operations. Down-sampling refers to a process in convolutional neural networks that reduces the resolution of feature maps to extract higher-level information. At the bottleneck layer, which represents the deepest part of the network where feature representations are highly abstract, the output feature map is processed by both DAM and ASPP. DAM applies spatial and channel-wise attention, helping the model focus on

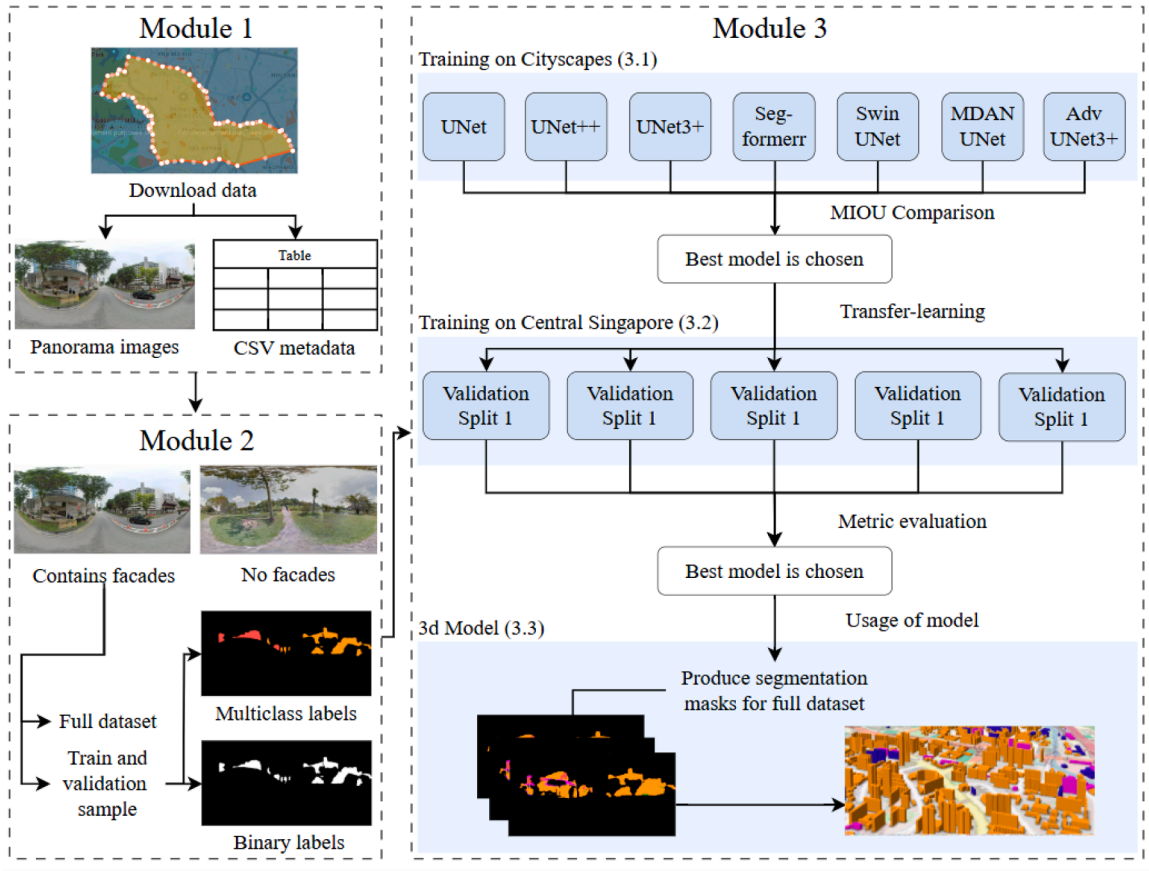


Fig. 1. Proposed components of methodology.

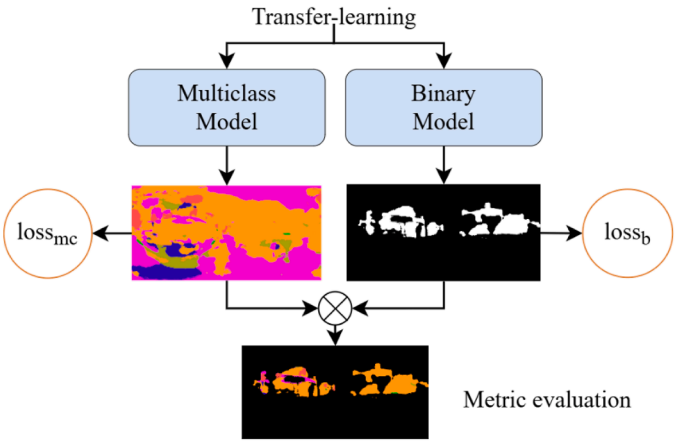


Fig. 2. The process of training models and combining their outputs.

the most relevant regions within the feature map. At the same time, ASPP captures multi-scale contextual information by applying convolutions with different dilation rates to account for varying object sizes. The outputs from these modules are concatenated, passed through a 1×1 convolution layer that refines features while maintaining spatial integrity, and combined with the original bottleneck input via weighted summation. This ensures that the model balances newly extracted high-level features and the original deep encoder output, preserving important spatial details. The final processed feature map then moves through the decoder structure, which employs full-scale skip connections—a mechanism in which features from deep and shallow layers are aggregated across multiple spatial resolutions to maintain fine details and broader scene structures. This allows the model to refine its segmentation outputs by leveraging both high-level contextual cues and low-level image details, improving its ability to distinguish façade materials from the surrounding urban environment.

For our backbone, we utilize EfficientNet-B0 as the encoder for its balance between accuracy and training time because it performs better than ResNet50 for medical segmentation tasks (Kansal et al., 2024). The image input first undergoes an initial convolution block (EC) that consists of a 3×3 convolution, batch normalization, and a down-sampling layer. The resulting output from EC then incurs a weighted sum with $a \times 0.5$ scaled version of the image, where the result is passed onto the encoder layers (EX_{0-3}). At each encoder layer, the original image is scaled and convoluted (C) to match the spatial and channel dimensions of the encoder output, where both undergo weighted summation to produce an input for the next encoder layer. Following the encoder, the ASPP and DAM receive the most abstract feature map at the Bottleneck Layer (BL), where their outputs are concatenated and subsequently, undergo weighted summation with their original input. This approach can enhance the model's ability to balance local detail awareness with global context understanding while sensibly weighting their contributions against the original input. Refined encoder and bottleneck outputs are then passed onto a decoder structure identical to that of UNet3+, which uses full-scale skip connections in each decoder layer (DX_{0-3}) to combine information across scales and generate a more accurate feature map. The final output from the segmentation head (SH) is a tensor representing the raw logits of different classes which will be sent to the loss function ($L^{(0)}$).

3.2.2. *Multi-scale facilitated encoders*

One predominant issue faced by variants of UNet models is the gradual loss of spatial information as the feature map descends the

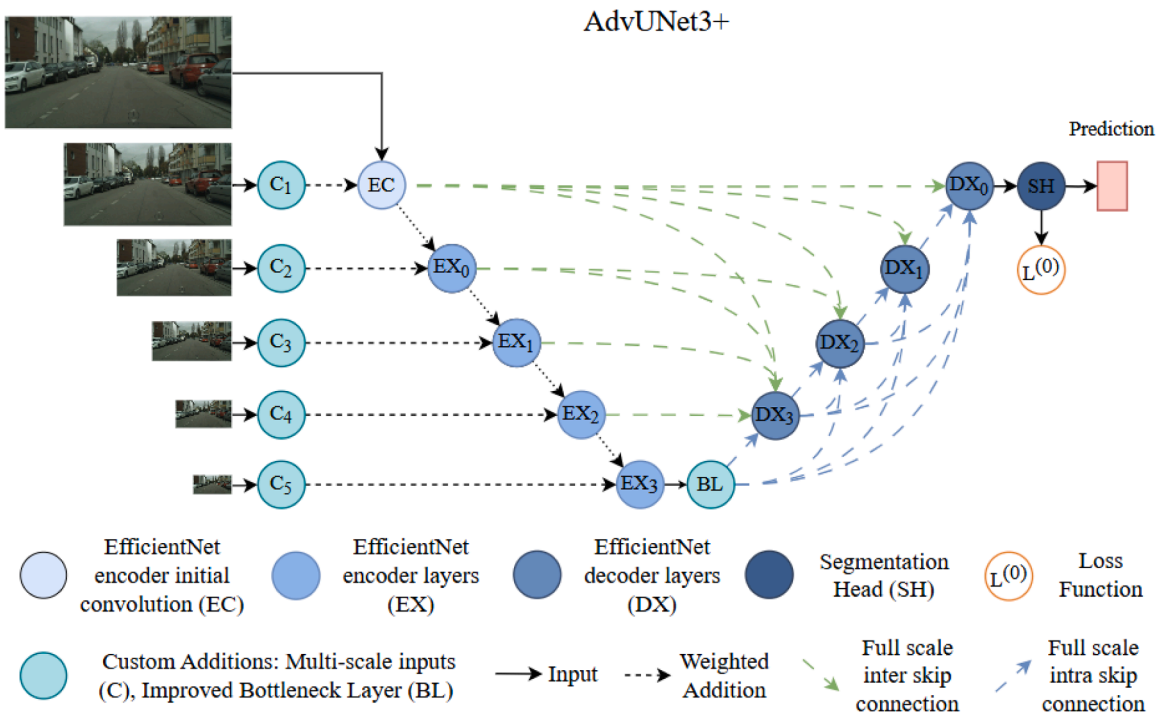


Fig. 3. AdvUNet3+ integrates multi-scale inputs, ASPP, and DAM.

encoder layers and experiences more pooling operations. Incorporating multi-scale inputs by summing them with encoder outputs thus allows us to preserve such information, thereby improving the model's segmentation result overall (Liu et al., 2020). To achieve downscaled versions of the image, previous works have constructed multi-scale inputs via average pooling layers (Fu et al., 2018). Another study using a similar network architecture takes the first value found in the upper left corner for every 2×2, 4×4, 8×8, and 16×16 non-overlapping area respectively, which is distinctly unique from max-pooling operations with a stride equal to 2 (Liu et al., 2020). Fig. 4 illustrates the difference between all three of these methods.

We use max pooling as the method of obtaining different scaled versions of the original image because of its ability to capture the most prominent features that better represent the input image overall. Specifically, following our defined scales, we take the maximum values within every 2×2, 4×4, 8×8, 16×16, and 32×32 windows that do not overlap with each other. Each of the scaled inputs is then passed through a 1×1 convolution to enforce identical channel dimensions, before being weighted and summed with the outputs of its corresponding encoder layer outputs. In the case of EfficientNet-B0, its encoder comprises an initial convolution block and several encoder layers. The processes of the initial convolution block and the encoder layers are highlighted in Fig. 5.

Fig. 5(a) illustrates the logic behind the addition of scaled inputs with EfficientNet-B0 encoder layer outputs. Additionally, Fig. 5(b) further elaborates on the operations within a single EfficientNet-B0 encoder layer. Our encoder section utilizes learnable weights defined using W_1 and W_2 (Fig. 5(c)), which are associated with the convolution output tensor and encoder output tensor, respectively. Each weight has a pre-defined value and is stored in an array, before passed through a softmax activation function and subsequently multiplied to their corresponding output tensor. Across Sum 1–5, W_1 always equals 1 to ensure the encoder output is given more priority. For W_2, they are defined in Eq. (1):

$$W_{2_{(x)}} = 2^{-(6-x)} \tag{1}$$

Where x represents the sum number from 1–5. This equation makes W_2, be 0.03125, 0.0625, 0.125, 0.25, and 0.5, which is the inverse of our defined scale sizes $0.5 \times$, $0.25 \times 0.125 \times$, $0.0625 \times$, $0.03125 \times$. Hence, the weighted contribution from the multi-scale input feature map increases as the encoder output tensor's spatial dimensions and its decrease, thereby allowing the multi-scale inputs to gradually enforce spatial information deeper in the network.

Integrating multi-scale inputs into the encoder improves material differentiation by preserving fine details while maintaining broader contextual awareness, reducing misclassification in complex urban environments. This approach helps distinguish materials with similar textures and spectral properties, particularly in diverse façades

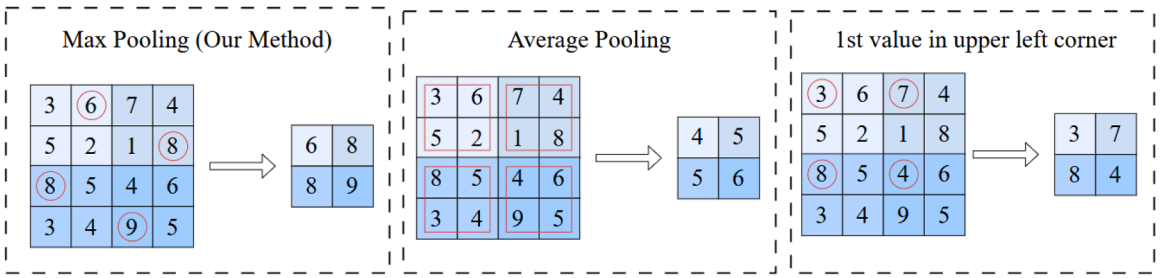


Fig. 4. Illustrations of each pooling operation were discussed.

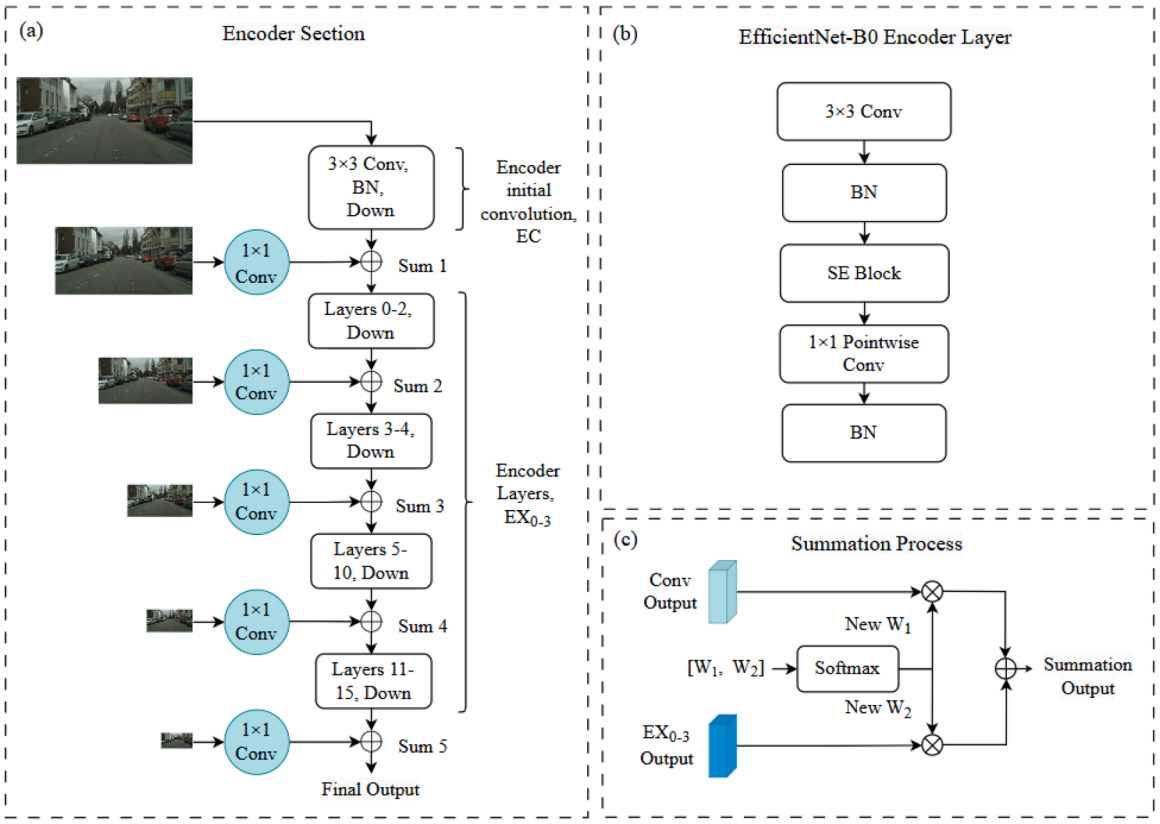


Fig. 5. The encoder processes and the weighted summation process. (a) Encoder Section with multi-scale inputs incorporated. (b) Processes within EfficientNet-B0 Encoder Layer. (c) Summation Process between encoder output and multi-scale inputs.

containing brick, glass, and painted concrete. By dynamically weighting scaled feature maps, spatial information is progressively reinforced, leading to more accurate segmentation, better boundary delineation, and improved generalization across different building typologies, regardless of their size within an SVI.

3.2.3. *Abstract bottleneck operations*

Typically, the output of the encoder at the bottleneck layer is directly passed onto the decoder structure, save for a few convolution operations. This paper introduces a DAM, which performs spatial and channel attention functions concurrently in conjunction with an ASPP module that gathers contexts across scales. The full operations of the DAM are depicted in Fig. 6. The original bottleneck's encoder output gets passed to the channel and spatial attention module, where their original outputs go through a 3 × 3 convolution layer, batch normalization layer, and ReLU activation function that halves their channel size. These transformed outputs are now summed to produce a unique tensor before they undergo the same operations as before to halve their channel size once again. Finally, the operations of the DAM conclude by concatenating all output tensors together.

At the same time, as presented in Fig. 7, the bottleneck's encoder output is fed into the ASPP, where multiple atrous convolutions of different dilation rates (1, 6, 12, 18) are applied to it in parallel to produce different variations of the feature map. In addition, a global

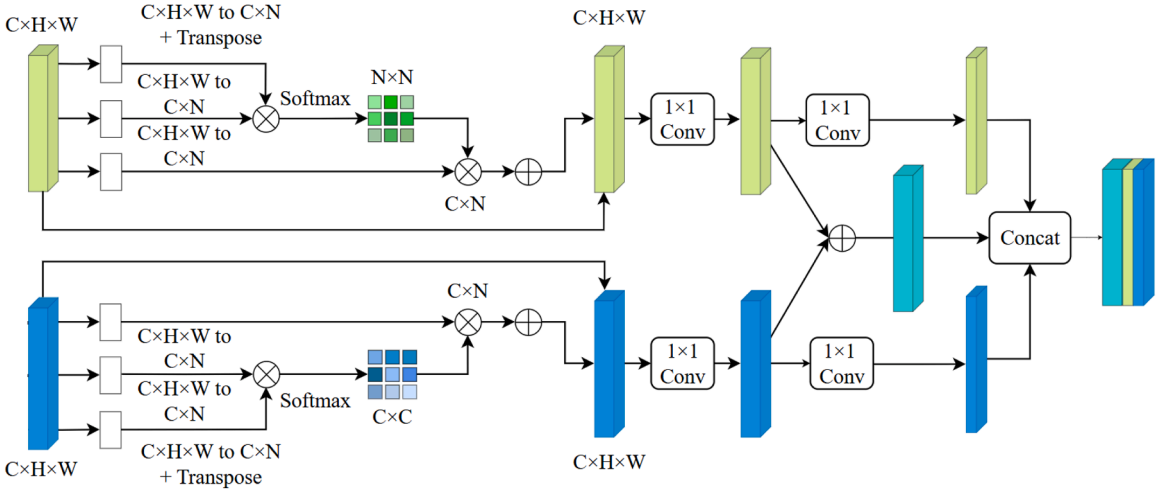


Fig. 6. Design of the Dual Attention Module (DAM).

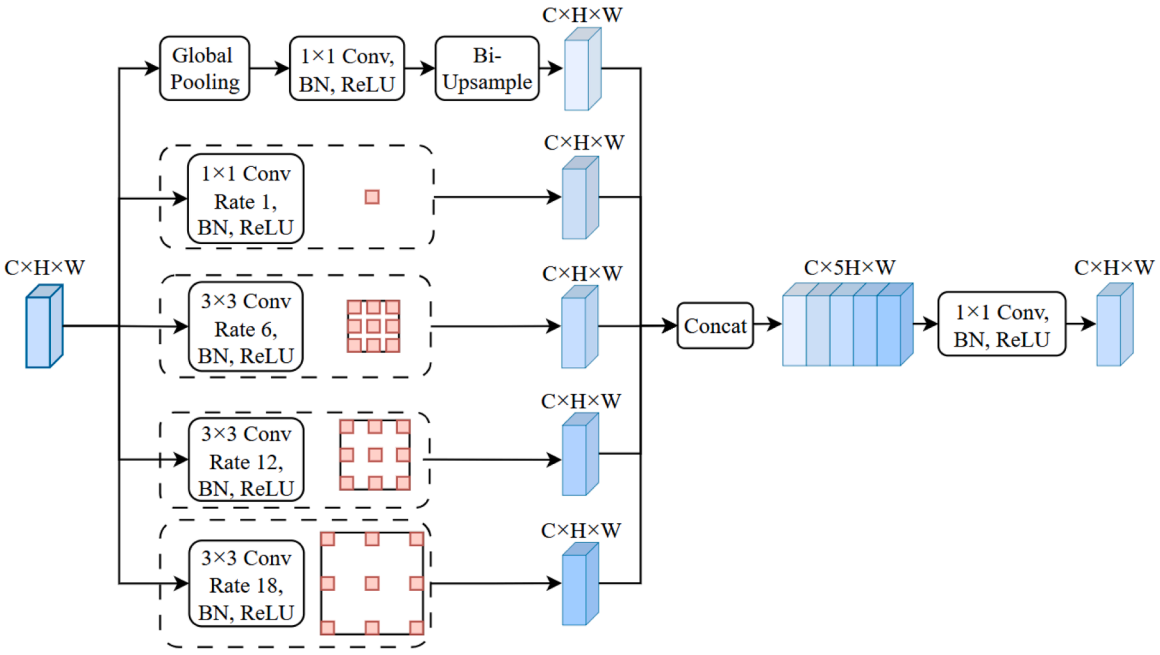


Fig. 7. Design of the Atrous Spatial Pyramid Pooling (ASPP).

average pooling operation is applied to derive a global context vector that summarizes the entire feature map. Finally, all feature maps are concatenated along the channel dimension and undergo a 1 × 1 convolution layer, ensuring the final output is compact but rich.

Once both modules have produced their respective output tensors, their outputs will be concatenated along the channel dimension to combine the information learned. Afterwards, the concatenated feature map will enter through a 1 × 1 convolution layer which halves its channel dimensions, producing an abstracted feature map with the same number of channels. The original input and the abstracted feature map will then be weighted and summed together to achieve an enhanced representation of the original input. In Fig. 8, the learnable weights are defined as 1 for both the original input ($W_{original}$) and the abstracted feature map ($W_{abstract}$).

The integration of DAM and ASPP enhances segmentation by capturing both fine-grained local features and broader contextual

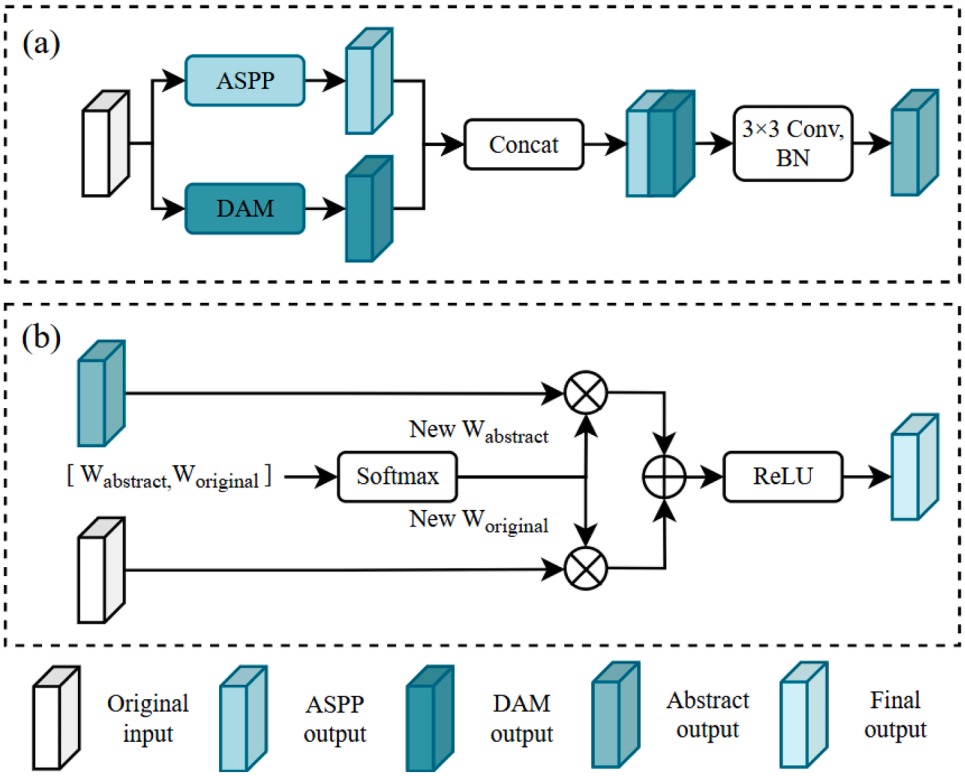


Fig. 8. ASPP and DAM Fusion with Weighted Summation. (a) ASPP and DAM concatenation and convolution. (b) Weighted summation of abstract and original features.

relationships, providing a more comprehensive representation of diverse façade materials. Unlike traditional methods that rely on basic convolutions at the bottleneck, this approach refines features before decoding, leading to more accurate segmentation, especially for complex textures. The use of weighted summation helps balance original and abstracted features, reducing information loss and improving material differentiation. These enhancements address key limitations in existing segmentation models, resulting in fewer misclassifications, stronger generalization across urban environments, and improved suitability for large-scale urban analysis and 3D façade reconstruction.

3.2.4. Loss functions

For our experiment, we used a combination of cross-entropy loss and dice loss with defined weights as the primary loss function. Dice loss can enforce shape and regional accuracy by maximizing the overlap between the predictions and the ground truth whereas cross-entropy loss ensures the correct classification of each pixel by measuring the difference between the predicted probabilities and actual labels. For the training in Cityscapes, only cross entropy and dice loss are used while for training on our constructed dataset, we use cross entropy loss and dice loss as the primary loss for multiclass outputs, and binary cross entropy loss with dice loss as the main loss of our binary output. The reason is that for our dataset, two distinct models are being trained separately from each other, meaning that each model requires its loss function designed to fit its specific task. The individual loss functions are presented in Eqs. (2) and (3).

$$\text{loss(BCE or CE)} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{2}$$

$$\text{loss(DE)} = 1 - 2\frac{\sum_{l=1}^{2} w_l \sum_n r_{ln} p_{ln}}{\sum_{l=1}^{2} w_l \sum_n r_{ln} + p_{ln}} \tag{3}$$

In this context, N denotes the total number of classes, y_i serves as the binary indicator for class i, and p_i indicates the predicted outcome for class i. For our model, we opt for binary cross-entropy loss as the variant used during backpropagation of the binary segmentation model and cross-entropy loss for multi-class segmentation. For dice loss, W_l is employed to ensure invariance to various properties of label sets. The primary loss functions for Cityscapes and our constructed dataset can thus be expressed using Eqs. (4)-(6), respectively.

$$\text{loss}_{\text{total}} = \alpha \bullet \text{loss(CE)} + \beta \bullet \text{loss(DE)} \tag{4}$$

$$\text{loss}_{\text{b}} = \alpha \bullet \text{loss}_{\text{b}}\text{(BCE)} + \beta \bullet \text{loss}_{\text{b}}\text{(DE)} \tag{5}$$

$$\text{loss}_{\text{mc}} = \alpha \bullet \text{loss}_{\text{mc}}\text{(CE)} + \beta \bullet \text{loss}_{\text{mc}}\text{(DE)} \tag{6}$$

Where α and β represent the weights of each loss and are both defined with a value of 1. loss_{b} and loss_{mc} are calculated using the binary predictions and multiclass predictions S^{b} and S^{mc} respectively. The loss $\text{loss}_{\text{total}}$ is used during backpropagation of the network when trained on Cityscapes while loss_{b} and loss_{mc} are used during backpropagation of the binary and multiclass versions of the network when trained on our dataset.

3.3. Study area

The study area comprises two districts named Bishan and Toa Payoh, which are in the northernmost part of the central region of Singapore (Fig. 9(a)). Both districts fall within the Bishan-Toa Payoh Town Council and Bishan-Toa Payoh Group Representation Constituency. As of 2024, both districts span a combined area of 15.79 km^2 with an approximated total population of 209,000 and are matured residential towns. A mature residential town is defined as any residential town older than 20 years. The collated SVIs (Fig. 9(b)-9 (c)) reveal that Bishan and Toa Payoh are ideal districts for obtaining information on various building architectures for two reasons. The first reason is that these districts have the standard architectural designs that can be found in other residential districts in Singapore. A large majority of buildings in any district can be categorized into either a residential estate or multi-purpose building, the latter being more commonly associated with amenities or office buildings with unique designs. The second reason is that these districts also have a rare group of buildings which add to the robustness of the dataset.

Fig. 9. Location of the study area. (a) Locations of districts Bishan and Toa Payoh. (b) Satellite view of Bishan and Toa Payoh. (c) Panoramic SVIs within Bishan and Toa Payoh.

3.4. Façade segmentation and 3D projection using SVIs and building model dataset

We pre-train the proposed model on Cityscapes (Cordts et al., 2016) then fine-tune our proposed dataset with similar network configurations to transition into a street view façade identification task. Meanwhile, we construct a dataset comprising SVIs in central regions of Singapore with a diversity of façade materials and architectural structures. We purchased Google SVI API and downloaded 32,549 panoramic images in the districts of Bishan and Toa Payoh, located in the central area of Singapore. The geographical boundaries were defined using precise coordinates, and images were captured at intervals of 50 m along the roads. After a thorough visual inspection, 24,219 images were retained. Labelling was conducted using an online platform called *LabelBox* with a customized ontology for façades and non-façades, and the SVIs were used to verify the accuracy of labels. Additionally, we obtained the building dataset from Singapore Buildings (https://data.humdata.org/dataset/hotosm_sgp_buildings) to project our segmentation outputs. This dataset includes metadata features like building level, which are essential for calculating the assumed height of the building, thus allowing the conversion into a 3D building model.

3.4.1. Assumptions

In this study, we propose three crucial hypotheses to address the challenges in labelling a variety of geo-object classes appearing in SVIs in Singapore, as well as promoting effective and precise segmentation. The first assumption is that each building has at most three materials per visible façade. This is based on the classification of Singapore's architecture into three main categories, standalone structures, mixed-use complexes, and unique structures as seen in Fig. 10. Compared to one novel study that only associates a maximum of two materials *per building* (Xu et al., 2023), our labelling procedure entails the inspection and differentiation of materials *per visible façade*. This assumption makes the segmented material closer to reality while making the whole process more complicated and challenging. The second assumption is that each façade consists of only one primary material, as most façades are constructed using multiple materials. To address this issue, the study disregards less prominent materials and posits that each component is made up of a single primary material. Therefore, the façade of any building compromises no more than three materials, including one primary material and possibly two other subsidiary materials. Lastly, we estimate residential building heights in our 3D building model by multiplying the number of levels by 3 meters (m), a standard approximation for typical floor-to-ceiling height in urban planning. Buildings with missing level data are assumed to be landed properties with a height of 9 m.

3.4.2. Labelling methodology

The labelling process follows an iterative approach to ensure accurate material classification in LabelBox. Firstly, AI-assisted labelling tools were used to generate bounding boxes over façade materials, streamlining the segmentation process by providing an initial structured representation of object locations. Secondly, manual refinement was applied to correct these inaccuracies, particularly for thin elements like metal beams and window dividers. A structured preprocessing pipeline was then used to refine and organize the annotations efficiently. Once finalized, annotation metadata, including class labels, bounding box coordinates, and segmentation masks, were extracted from LabelBox in JSON format. A custom script parsed this metadata, retrieved annotated images, and converted bounding boxes into pixel-wise segmentation masks. This post-processing step merged multiple annotations into a single mask per image, ensuring consistency across the dataset. Any remaining inconsistencies were addressed through additional verification, resulting in high-quality ground truth data for segmentation tasks.

3.4.3. Façade classification and labeling criteria

Our study area comprises three main types of architecture as shown in Fig. 10(a)-10 (c), including stand-alone flats, multi-purpose buildings,

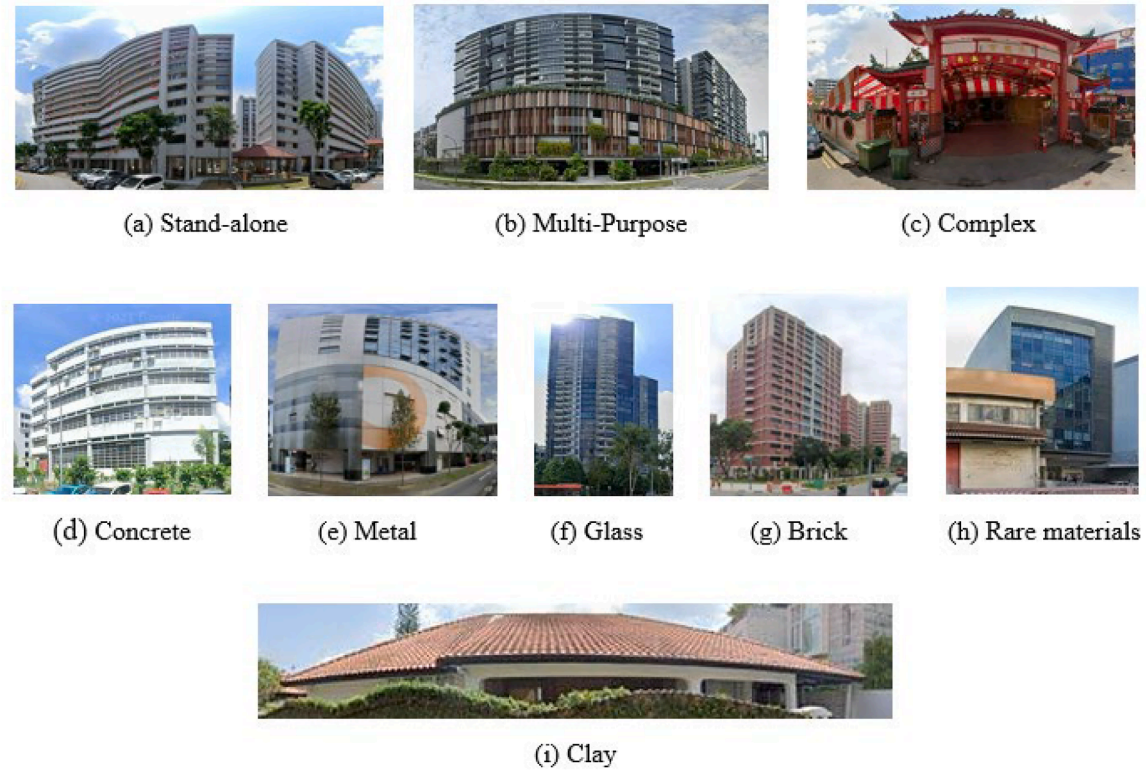


Fig. 10. Images of common building architectures and façade materials. (a-c) Most common categories of building structures. (d-i) Façade materials in our constructed dataset.

and complexes, as well as six different façade materials shown in Fig. 10 (d)-10 (i). Based on the pixel counts within labeled ground truth masks, we confirmed that paint in Fig. 10(d) is the most common façade material across Singapore. They are commonly used to heighten the aesthetic appeal of domestic buildings, especially those falling under the residential and industrial categories. Furthermore, they serve as an extra façade layer on top of the exposed brick of cement, thereby making it resistant to environmental wear and tear. A contemporary façade that is largely used in the same category of buildings is the brick (Fig. 10(g)), which are common façades for some of the older buildings. Contrastingly, buildings of a commercial nature, often taking the form of malls or offices, are typically made of glass, metal, or a combination of either, given their sleek appearance.

Each material class was defined based on its visual distinguishability and spectral characteristics. Additionally, a key consideration is whether these materials are also commonly used in other countries, thereby justifying their relevance. For the remaining façades that have not been elaborated in greater detail, clay primarily consists of terracotta roofs which exist in low-rise landed properties and high-rise HDB flats (Fig. 10 (i)). Their ample appearance within the dataset which extends into the appearance of some façades coupled with their discernible appearance made them a justifiable addition to the dataset and albedo calculation process. Metal comprises metallic materials like aluminum, iron, or alloy façades found in commercial buildings (Fig. 10(e)), while glass refers to most office buildings or commercial structures in general (Fig. 10(f)). Here we omit the usage of background during evaluation. Rare materials are a unique minority class that comprises uncommon façade materials that, if labeled as their own class, would make an even smaller minority that could affect the model's performance (Fig. 10(h)). They include materials like ceramic and mosaic, which are rare in Singapore's architectural landscape. As a result, the Central-Singapore dataset consists of seven annotations including background, paint, metal, glass, brick, rare material, and clay, as summarized in Fig. 10.

3.4.4. Inter-annotator agreements

Inter-annotator agreements were established through collaborative visual inspections, predefined classification criteria, and structured quality control measures to ensure consistent labeling standards. Material classes were defined based on their visual characteristics, spectral properties, and relevance to urban environments. Visual inspection served as the primary classification method, with Google Street View used for verification in cases where reflections, shadows, or occlusions made identification difficult. A hierarchical classification strategy prioritized the dominant material on a façade, while secondary materials were annotated only if they occupied a significant portion of the surface. Decorative elements such as signage, murals, or trims were generally excluded unless they were integral to the structure. Classification rules ensured consistency across different architectural styles, such as labeling glass façades with metal frames as glass if they covered over 70 percent of the surface.

Reflective surfaces were classified based on their structural material rather than their appearance under changing lighting conditions. Additional agreements were introduced to address complex cases, including material transitions, overlapping features, and structural variations that could lead to misclassification. Special attention was given to distinguishing thin metal overlays on glass and ensuring weathered surfaces were labeled based on their original construction rather than temporary wear. To maintain uniformity, multiple rounds of visual analysis were conducted, comparing buildings with similar compositions and resolving discrepancies through structured review sessions. These quality control measures reinforced dataset reliability, ensuring a consistent and accurate representation of urban façades for high-quality segmentation models.

3.4.5. Characteristics of the customized SVI dataset

In comparison to other datasets (Teboul et al., 2011; Riemenscheider et al., 2012; Kong & Fan, 2021; Wang et al., 2024), ours present three distinctive advantages. First, our dataset is considerably larger compared to other datasets depicted in Table 1 apart from the Hong Kong Street View dataset, containing 800 carefully selected and annotated images with over 5 thousand buildings. The resolution of our images is 512 × 1024 which is standard when compared to other street view datasets like Cityscapes with images containing matching resolutions. Second, our dataset captures the appearance of façades from multiple different angles and distances, which is far superior to those that contain images taken from a single angle of regular façade shapes up close. Additionally, our dataset contains nuanced foreground occlusions including vegetation, trees, signage, and traffic under various lighting conditions. We believe that the diverse quality of images would enhance the model's ability to generalize. Third, our method also collates supplementary metadata, primarily due to the choice of our software for data collection. Our metadata is specific to each picture taken. These metadata include things like geographical coordinates, angle of elevation, angle of rotation from the north bearing, and so on. Intuitively, the use of this information is extremely beneficial within the context of replicating the segmentation results on a 3D model. Given these attributes, our dataset provides a well-rounded representation of real-world street-view façades, making it suitable for training segmentation models that must handle diverse urban conditions.

4. Empirical experiments

4.1. Training details

We used Pytorch (Paszke et al., 2019) to construct and evaluate the proposed network architecture. All experiments are conducted using a NVIDIA GeForce MX450 GPU. In the network, input images are defined in scales 1.0 ×, 0.5 ×, 0.25 × 0.125 ×, 0.0625 × and 0.03125 × to produce feature inputs of identical spatial dimensions to the corresponding feature maps in their designated encoder layer. As the scale progressively gets smaller, fine details get traded for larger receptive fields and vice versa. Furthermore, images are cropped to 256 × 512 and batch size set to 2 per GPU to reduce computational costs, particularly to fit within the memory constraints of the apparatus. The rest of the configurations are listed in Table 2.

To validate the effectiveness of the developed model, we compared segmentation results produced by UNet, UNet++, and UNet3+, using the Cityscapes dataset. For all the models, their encoder structures follow the EfficientNet-B0 architecture and have weights that have been pre-trained on the "Imagenet" dataset (Deng et al., 2009). The proposed model then applies transfer learning by selecting building and vegetation weights from the segmentation head and fitting them into a redefined version of the model with identical configurations. This allows for the model to leverage information gained previously from Cityscapes while allowing untouched weights in the model to assimilate to the new task.

Table 1
Comparison of existing façade related datasets.

Dataset Name	Size	Occlusion	Single View	Diversity	Citation
ECP2010	104	×	✓	Low	Teboul et al., 2011
Graz2012	50	×	✓	Low	Riemenschneider et al., 2012
MCUbes	500	✓	×	High	Liang et al., 2022
Hong Kong SVIs	2003	✓	×	High	Xu et al. 2023
Our Central-Singapore Dataset	800	✓	×	High	N.A.

Table 2
Specific experiment configurations.

Item	Configuration
Image Scales	{1.0 ×, 0.5 ×, 0.25 × 0.125 ×, 0.0625 ×, 0.03125 ×}
Crop Size	256 × 512
Batch Size	2 per GPU
Learning Rate	0.001 - 0.00001
Optimizer	Adam
Learning Rate Scheduler	Cosine Annealing with Warm Restarts
Loss Function	Cityscapes - Cross Entropy Loss, Dice Loss Constructed Dataset - Cross Entropy Loss, Binary Cross Entropy Loss, Dice Loss

4.2. Metrics

To quantitatively assess performance, our selection of metrics includes the mIoU, IoU, Precision, Recall, F1-Score, and accuracy for the analysis of experimental outcomes. We also employed macro-averaging (Sokolova & Lapalme., 2009) to calculate the average values of these metrics. In this study, we have excluded background classes during the calculation and assessment of these metrics. Equations (7) – (12) are used to represent the metrics stated:

$$\mathrm{IoU} = \frac{\mathrm{TP_i}}{\mathrm{FP} + \mathrm{FN} + \mathrm{TP}} \tag{7}$$

$$\mathrm{IoU} = \frac{1}{\mathrm{N_{class}}} \sum_{i=1}^{\mathrm{N_{class}}} \frac{\mathrm{TP(i)}}{\mathrm{TP(i)} + \mathrm{FP(i)} + \mathrm{FN(i)}} \tag{8}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$\mathrm{F1-Score} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})} \tag{11}$$

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{12}$$

where TP represents the true positives or number of samples predicted as positive which were correct, FP represents the false positives or number of samples predicted as positive which were incorrect, TN represents the true negatives or number of samples predicted as negative which were correct, FP represents the false negatives or number of samples predicted as negative which were incorrect and N representing the number of classes.

4.3. Results

This study trains and evaluates the segmentation performance of UNet variants, transformer-based models, and hybrid architectures on the Cityscapes dataset. According to Table 3, AdvUNet3+ outperforms all UNet-based encoder models (UNet, UNet++, and UNet3+), transformer-based models (SwinUNet, and SegFormer), and the hybrid MDAN-UNet, achieving the highest mIoU of 61.4 %. Notably, AdvUNet3+ demonstrates superior segmentation in complex and low-proportion classes, particularly for train (43.8 %), motorcycle (34.5 %), truck (49.4 %), and bus (58.7 %), highlighting its ability to handle rare yet structurally distinct objects. Compared to encoder-based UNet variants, which tend to over-smooth small objects, AdvUNet3+ retains fine-scale details and better preserves object boundaries. Unlike SwinUNet and SegFormer, which struggle with object separability in crowded urban environments, AdvUNet3+ provides more precise delineation of façade materials and traffic elements. Additionally, while MDAN-UNet improves over standard UNet architectures through attention-based mechanisms, its performance on rare classes is inconsistent, whereas AdvUNet3+ maintains strong segmentation across both high- and low-frequency categories.

Despite these strengths, AdvUNet3+ faces challenges in segmenting thin objects or elements that would typically be grouped into large expanses of non-façades in street view images, such as sidewalk and road (74.3 % and 96.6 %) or sky and terrain (92.6 % and 55.6 %). However, given that these classes already occupy large portions of the dataset and do not require high segmentation precision, this trade-off does not significantly impact the model's practical utility. Therefore, it demonstrates that adding multi-scale inputs and abstract bottleneck operations could improve the understanding of highly specific details, and possibly segment rarer classes in the proposed dataset with higher accuracy.

The visual comparison of segmentation results highlights the strengths of AdvUNet3+ in delineating objects more clearly and reducing segmentation noise, particularly in complex urban environments (Fig. 11). Specifically, Fig. 11(a), Fig. 11(c), and Fig. 11(e) illustrate how AdvUNet3+ excels in preserving fine-grained details and improving class distinction. In Fig. 11(a), the red-circled area demonstrates the model's ability to maintain sharp façade boundaries,

Table 3
The IoU of segmented classes, using UNet architecture and Cityscapes.

Classes	UNet (%)	UNet ++ (%)	UNet3+ (%)	Seg former (%)	SwinUNet (%)	MDAN-UNet (%)	Adv UNet3+ (%)	Portion (%)
Road	96.7	**96.9**	96.5	96.2	96.0	96.8	96.6	37.6
Sidewalk	74.8	**75.9**	74.7	72.0	70.9	75.2	74.3	5.4
Building	87.4	87.4	**87.5**	85.4	85.4	87.4	87.3	21.9
Wall	36.6	31.8	36.5	31.7	33.3	33.5	**38.3**	**0.7**
Fence	**38.9**	35.2	37.2	32.3	30.5	36.6	38.1	0.8
Pole	48.1	**49.7**	49.4	35.9	41.5	48.5	47.2	1.5
Traffic Light	50.3	**51.2**	48.9	40.9	40.2	50.3	47.6	0.2
Traffic Sign	57.9	**62.3**	58.5	51.4	51.3	59.8	58.9	0.7
Vegetation	89.1	**89.2**	89.0	87.1	87.7	88.9	88.8	17.3
Terrain	55.8	52.3	**56.5**	54.0	52.4	55.3	55.6	0.8
Sky	92.1	**93.0**	92.7	89.9	91.4	92.7	92.6	3.4
Person	66.5	**68.2**	66.6	61.6	57.9	66.2	65.3	1.3
Rider	43.3	44.4	41.9	34.9	31.9	39.8	38.9	0.2
Car	89.6	89.6	89.5	87.0	85.4	89.6	**89.7**	**6.5**
Truck	41.6	45.8	48.4	42.8	34.9	47.7	**49.4**	**0.3**
Bus	57.6	54.5	56.0	55.2	46.8	55.9	**58.7**	**0.4**
Train	32.6	24.8	37.0	24.9	22.8	31.2	**43.8**	**0.1**
Motorcycle	25.1	31.9	32.6	23.8	14.4	**35.3**	34.5	0.1
Bicycle	62.4	**63.2**	63.0	56.6	54.0	61.8	61.6	0.7
mIoU	60.3	60.4	61.2	56.0	54.5	60.6	**61.4**	

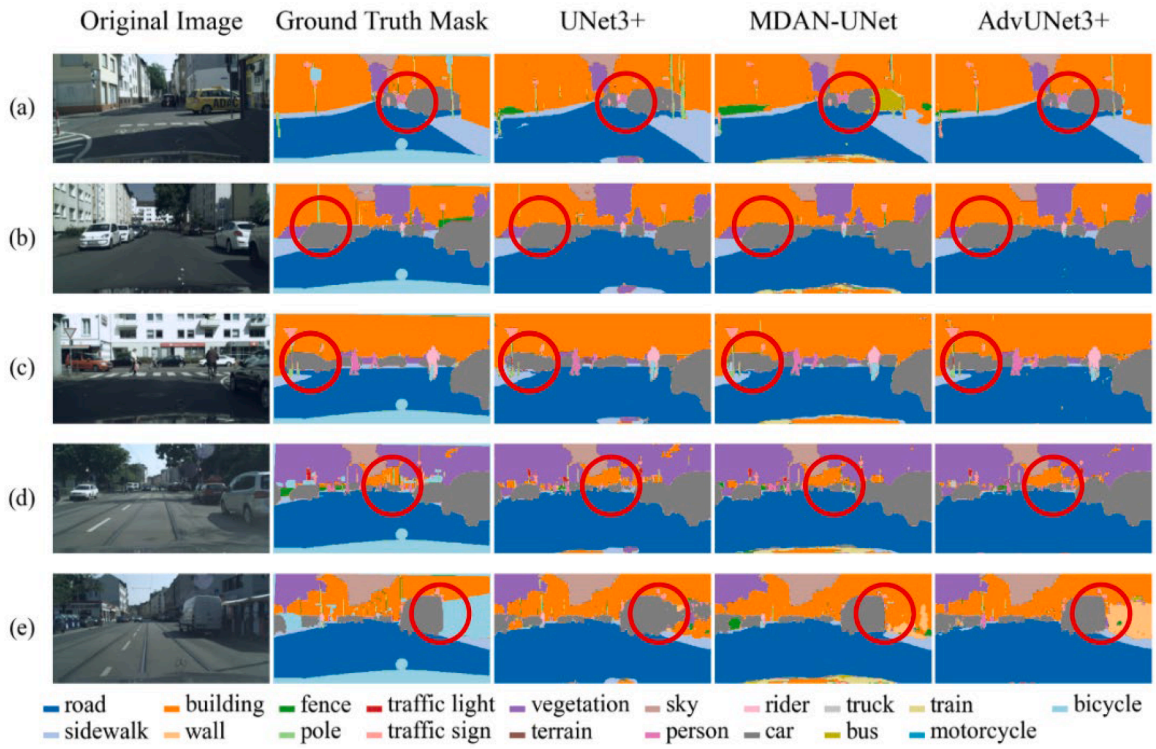


Fig. 11. Visual comparison of AdvUNet3+ model and bench line outputs.

correctly separating road, sidewalk, and building surfaces with minimal blending. Similarly, Fig. 11(c) showcases how AdvUNet3+ distinguishes traffic signs and poles from adjacent building structures, which UNet and UNet++ often fail to differentiate, leading to class merging. In Fig. 11(e), despite a more challenging scene, the model provides clearer segmentation of road edges and vehicles, reducing noisy artifacts frequently appearing in UNet3+ and UNet++. Unlike these models, which over-smooth and blend multiple façade materials into a single class, AdvUNet3+ ensures sharper transitions between materials, providing a more accurate and visually coherent segmentation. Additionally, the model preserves structural integrity, ensuring that large objects like roads and buildings remain distinct while minimizing unnecessary noise.

However, Fig. 11(b) and Fig. 11(d) highlight some of AdvUNet3+'s limitations, particularly when segmenting thin objects and closely positioned structures. In Fig. 11(b), the red-circled areas indicate instances where poles and thin traffic signs are merged with the background or nearby objects, leading to occasional misclassification. Similarly, Fig. 11d illustrates the model's difficulty in distinguishing thin architectural elements, where narrow structures are sometimes absorbed into broader categories such as vegetation or road surfaces. Additionally, Fig. 11(e) reveals that while the model's segmentations are more defined and less noisy, they can occasionally lead to incorrect classifications. The red-circled area highlights instances where misclassifications occur despite the clear object delineation, suggesting that while AdvUNet3+ enhances boundary sharpness, its class predictions may still require refinement. This indicates that the model may prioritize structural clarity over classification accuracy in certain cases, necessitating further adjustments to balance both aspects effectively. Nonetheless, AdvUNet3+ remains the most consistent and structurally aware segmentation model, outperforming its predecessors by reducing noise, improving object boundaries, and maintaining visual coherence while preserving the integrity of large-scale urban structures.

Furthermore, we train and evaluate AdvUNet3+ on the created dataset using 5-fold training and validation (80 % and 20 % of the 800 labeled SVIs, respectively). Overall, the model performed the best on fold 3 with a mIoU of 52.9 % across all classes, demonstrating acceptable performance on common and confusable, rare textures. Across all validation splits, the model was able to distinguish the most common façade materials based on pixel count percentages, paint, and brick, with the highest degree of accuracy, netting an average IoU of approximately 84 % and 80 % respectively. It is also found that the best performance of the model obtained a mIoU of 52.9 % in split 3, while others got a mIoU lower than 50 % (Table 4). This is mainly achieved by the precise segmentation of *metal,* with an IoU equaling 50 % in split 3 compared to other splits with an IoU below 30 %. Additionally, the model also exhibits satisfactory potential for segmenting glass, and clay façades, obtaining maximum IoUs of 53.5 % for glass and 54.9 % for clay across splits.

To supplement the investigation of accuracy metrics, this study also compares the Precision, Recall, F1-Score, and Accuracy of each façade material class. Table 5 denotes that for fold 3, the model produces acceptable results for precision and recall in all classes, except for rare materials and paint respectively. Additionally, the F1-Score effectively harmonizes the calculations of Precision and Recall, affirming the model's adequacy at segmenting all façade materials as it obtained an average score of 60.4 %. Close inspection of the F1-Score also reveals that the model achieves a respectable balance between precision and recall for metal, glass, brick, and clay façades based on their

Table 4
The IoU of segmented classes, using 800 labeled SVIs in central Singapore.

Classes	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold4 (%)	Fold 5 (%)
Paint	83.9	82.2	87.4	83.0	81.4
Metal	20.9	17.9	50.0	18.7	17.0
Glass	42.9	47.9	45.9	53.5	49.3
Brick	88.5	72.0	85.6	79.2	81.9
Rare Materials	12.7	2.0	3.2	10.1	9.4
Clay	40.8	54.9	45.1	48.3	42.1
mIoU	48.3	46.2	**52.9**	48.8	46.9

Table 5
The indicators of segmented classes in fold 3, using 800 labeled SVIs in central Singapore.

Classes	IoU (%)	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Paint	87.4	94.5	34.9	51.0	69.7
Metal	50.0	66.3	71.6	68.8	90.0
Glass	45.9	59.2	83.7	69.3	91.3
Brick	85.6	94.9	93.5	94.2	98.0
Rare Materials	3.2	3.3	82.5	6.3	83.8
Clay	45.1	58.5	97.0	73.0	92.8
Mean	52.9	62.8	77.2	60.4	87.6

corresponding F1-Score of 68.8 %, 69.3 %, 94.2 %, and 73.0 %. Moreover, the overall accuracy of classes is demonstrably high, with an average of 87.6 % and a maximum of 98.0 % from brick, which is close to perfection.

Table 6 shows that rare materials make up the smallest proportions in the training and validation datasets, accounting for just 0.7 % and 0.9 % of pixels, respectively. As a result, 96.7 % of rare material pixels are misclassified, with 70.6 % incorrectly labeled as paint and 21.6 % as metal (Fig. 11). The scarcity of rare materials and their frequent misclassification into other categories largely explains their low IoU, Precision, and F1-Score. In contrast, paint dominates the dataset in terms of pixel count (Table 6) and is the material most predicted in place of others (Fig. 11). The frequent occurrence of paint and its tendency to be overpredicted helps explain its higher recall, F1-Score, and Accuracy. The model's inability to distinguish rare materials as a unique category is likely due to their low dataset representation and varied visual characteristics, making them difficult to classify reliably. Similarly, glass presents challenges, with 38.48 % of its pixels mislabelled as paint. Despite achieving a high recall of 83.7 %, glass segmentation remains inconsistent, especially when it appears alongside metal overlays or reflective surfaces, leading to frequent misclassification.

Misclassification patterns are also observed in metal and clay surfaces, though their performance is slightly better. Fig. 12 shows that metal is correctly identified 66.33 % of the time but is misclassified as paint in 24.56 % of cases, suggesting that metallic elements often blend into painted structures, particularly in cladding or roofing applications. Similarly, clay roofs are frequently misclassified as paint (39.83 %), reinforcing the model's tendency to absorb smaller façade components into the dominant building material. In contrast, brick and paint achieve the highest segmentation accuracy, with 94.87 % of brick correctly classified, likely due to its textured pattern and strong association with older HDB flats. Paint benefits from its overrepresentation in the dataset (73.6 %), leading to higher recall and frequent overprediction in ambiguous cases. While these trends highlight areas for improvement, the model remains effective in identifying large-scale façade structures, with errors primarily affecting minority and composite material classes rather than the overall segmentation framework.

Figs. 13(a)–13(b) demonstrate that the model successfully captures dominant painted surfaces but struggles with clay roofs, often merging them into the main building structure and labelling them as paint. While this results in a loss of fine-grained detail, the overall segmentation remains largely accurate and visually coherent, as paint is the most prevalent material in the dataset. Similarly, Figs. 13(f)–13(g) show that painted surfaces in high-rise buildings are consistently identified, but the model occasionally blends glass and metal elements into painted regions, likely due to reflectance similarities in panoramic imagery. Despite this, the segmentation of large-scale façades remains structurally sound, with most building components correctly identified. However, the model continues to struggle with metal roofs, as seen in several examples where they are frequently mislabelled as either painted surfaces or other building materials. This suggests that the model has difficulty distinguishing between metal used in structural overlays and metal used as a core façade material.

Table 6
Proportions of the pixels in training and validation, using 800 labeled SVIs in central Singapore.

Classes	Train (%)	Validation (%)	Total (%)
Paint	73.6	73.7	73.6
Metal	8.4	5.3	7.8
Brick	8.4	10.8	8.9
Glass	7.3	7.7	7.4
Clay	1.6	0.9	1.5
Rare Materials	0.7	1.7	0.9

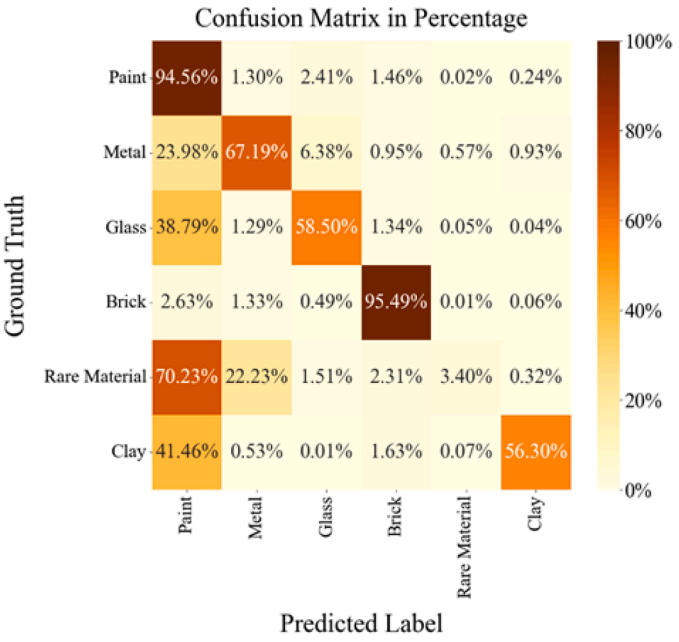


Fig. 12. Confusion matrix of the percentages of pixels in six classes. The rows represent the ground truth, and the columns represent predicted labels.

For rare materials, Figs. 13(c)–13 (e) highlight their frequent misclassification as paint or metal, underscoring the model's difficulty in recognizing materials that lack distinctive visual patterns or architectural placement. Unlike brick, which is often tied to older HDB flats, or clay, which typically appears on roofs, rare materials lack clear spatial context, making them difficult to distinguish. Fig. 13(h) further illustrates this issue in construction-heavy environments, where the presence of scaffolding and partially completed structures introduces additional segmentation ambiguity. In such cases, the model struggles to define material boundaries, often absorbing rare materials into larger, more dominant classes. Another persistent challenge is seen in glass façades with metal overlays or components, where the model inconsistently assigns portions of the façade to metal rather than recognizing the full surface as glass. Overall, this suggests that while the model captures transparent surfaces effectively, it struggles with composite façades that integrate multiple materials within the same structure. However, these errors have a limited impact on overall segmentation accuracy, as rare materials constitute a small proportion of the dataset and do not significantly alter the identification of major façade materials or their corresponding albedos.

Conclusively, the evaluation of AdvUNet3+ on our central Singapore dataset reveals that the accuracy metrics on average and per class are acceptable based on the following reasons. The first is that the proposed model has shown from individual class F1-Score and IoU to be able to accurately segment façade materials like paint, metal, glass, clay, and brick. Secondly, the infrequency of rare materials in the set of panoramas used for training is resemblant to that of our entire central

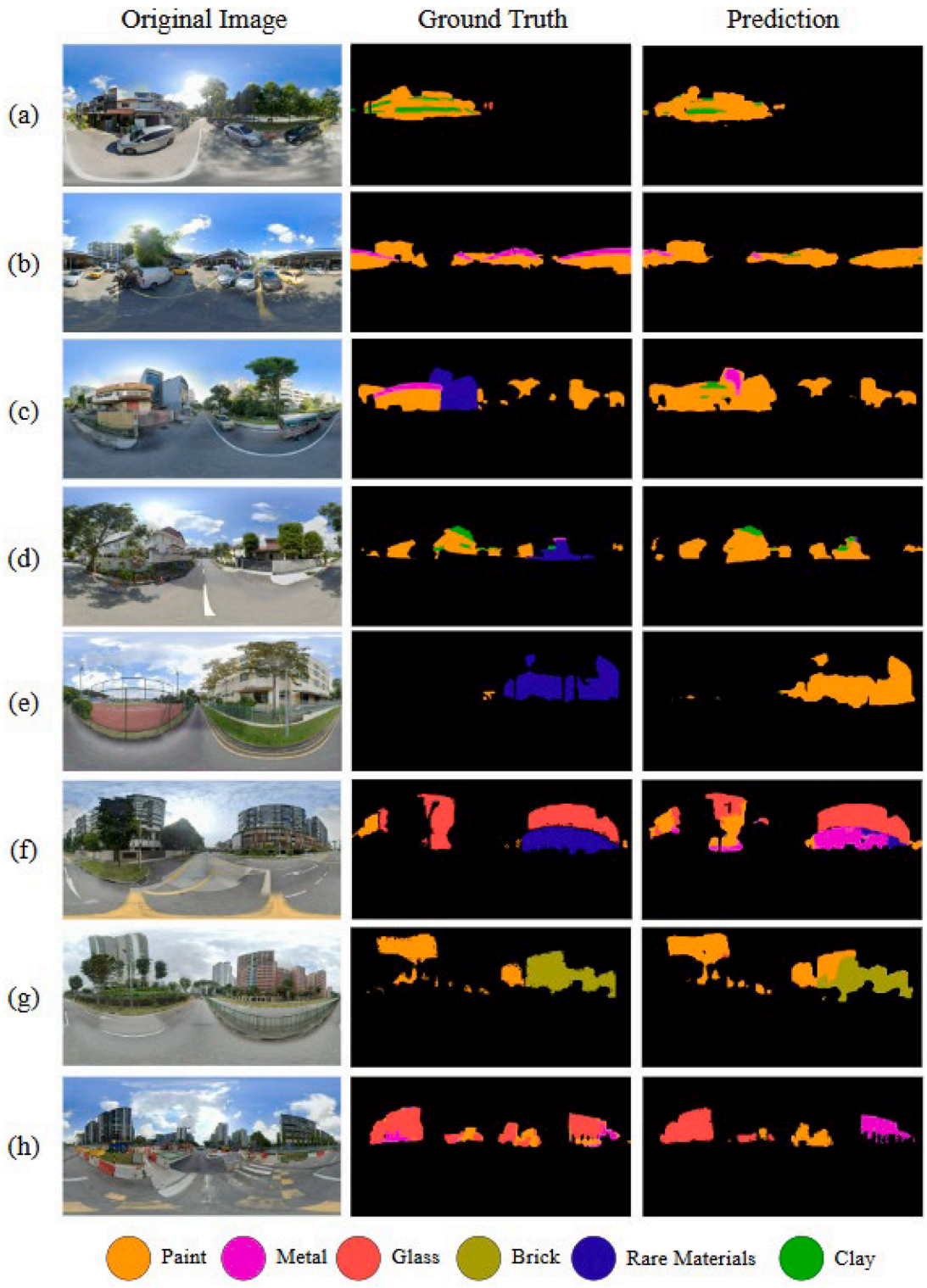


Fig. 13. Comparison between ground truths and predictions derived from AdvUNet3+.

Singapore dataset that comprises 24,219 images. This means that the segmentation results for rare materials can be overlooked since inaccurate predictions of such a rare class are unlikely to detriment the overall accuracy of predictions across the entire dataset. Thirdly, the difference in segmentation results between our study and others such as from Xu et al. (2023) can be attributed to highly specific labeling principles. For example, our labels do not merge instances of non-façades with building materials no matter how small the area they obscure is. Additionally, we make sure to only distinguish amongst building façades in our labels to provide an accurate and fair representation of segmentation results for purely façade materials. Therefore, our model is deemed to have demonstrated proficiency in segmenting building façade materials in general with acceptable accuracy metrics, especially accounting for this enhanced difficulty introduced by our labeling strategy.

4.4. Projecting façade material information onto 3D building models

We follow a coordinated 3-step process to incorporate the predictions into a 3D model (Fig. 14). Firstly, we identify the panorama location by matching the prediction name with the filename of the original image found in our collated meta-data file. Then, we obtain the geographical coordinates of that panorama like latitude and longitude to locate it on the 3D model and use the variable rotation or azimuth to simulate the direction the camera was facing at the point the panorama was taken. Next, we vertically split the predictions into 4 equal parts. To do this, we first assume a line down the center of the prediction indicating 0 from the panorama's point of view. Subsequently, we form two new lines adjacent to the left and right of the first, indicating the angular point of view of the panorama at 270 and 90, respectively. As a result, we can identify the direction of a building from the panorama concerning the range of angles it falls in, then specifically locate the exact building in the 3D model.

In step 3, we developed a Python script that performs two key tasks to enhance the integration of façade material predictions with our 3D model. First, the script determines the dominant material for each building by analyzing the segmentation results of nearby façades within a specified proximity. It counts the frequency of each material class in the pixels corresponding to a building's footprint, and the most frequent material is assigned as the dominant material. The resulting pool of dominant materials thus consists of paint, metal, glass, brick, and rare materials. Clay is never considered the dominant material since it usually consists of a small minority of a building's façade, typically its roof. This dominant material is then integrated into the metadata of the 2D building model in tabular form, allowing for easy reference. Second, the Python code projects the segmentation outputs onto the 2D building model, ensuring the façade materials are accurately aligned with the building's footprint. These results are then converted into a 3D polygonal representation using QGIS and the QGIS2threejs tool, effectively mapping the façade materials onto a 3D model, where the dominant façade materials, except for clay, color of each building. In addition, the heights of each building are represented by multiplying their building levels by 3 meters (m) or assuming a constant height of 9 m for typical landed property type buildings without building level values.

Figs. 15(a)–(e) show close-up views of 3D buildings from selected sections of the Bishan and Toa Payoh areas, with each building color-coded based on its dominant façade material. Fig. 15(a) represents painted façades, while Fig. 15(b) highlights buildings primarily composed of metal. Fig. 15(c) illustrates structures with glass façades, whereas Fig. 15(d) showcases buildings with brick materials. Lastly, Fig. 15(e) focuses on structures classified under rare materials. This approach ensures that the segmented façade materials are accurately mapped onto the corresponding buildings in the 3D model using available spatial metadata, providing a structured method that enables (i) further analysis of urban façades in a geospatial context, (ii) support of applications in urban planning, (iii) energy modelling, and (iv) architectural studies.

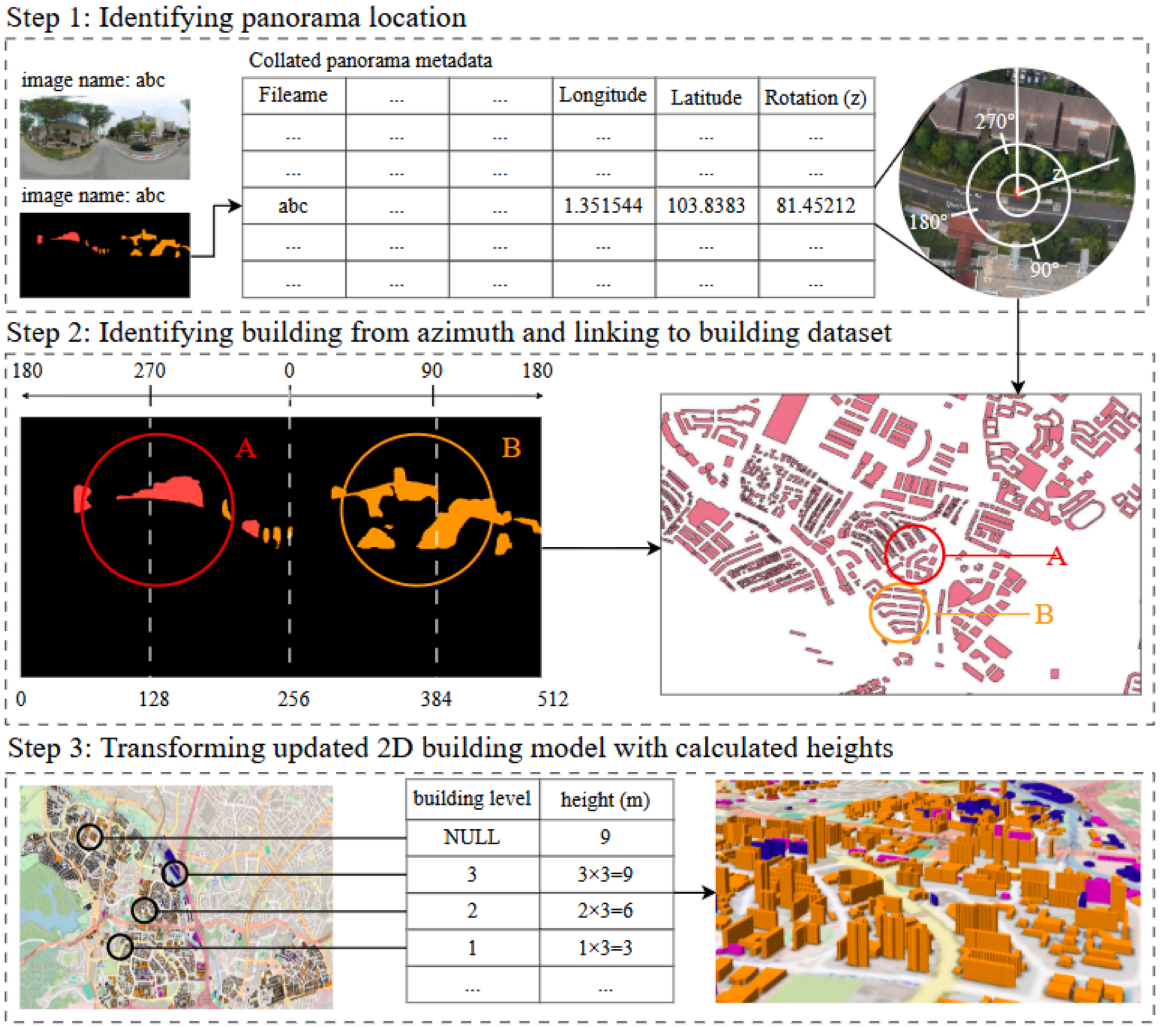


Fig. 14. Projection of segmented building façade material onto the 3D building models.

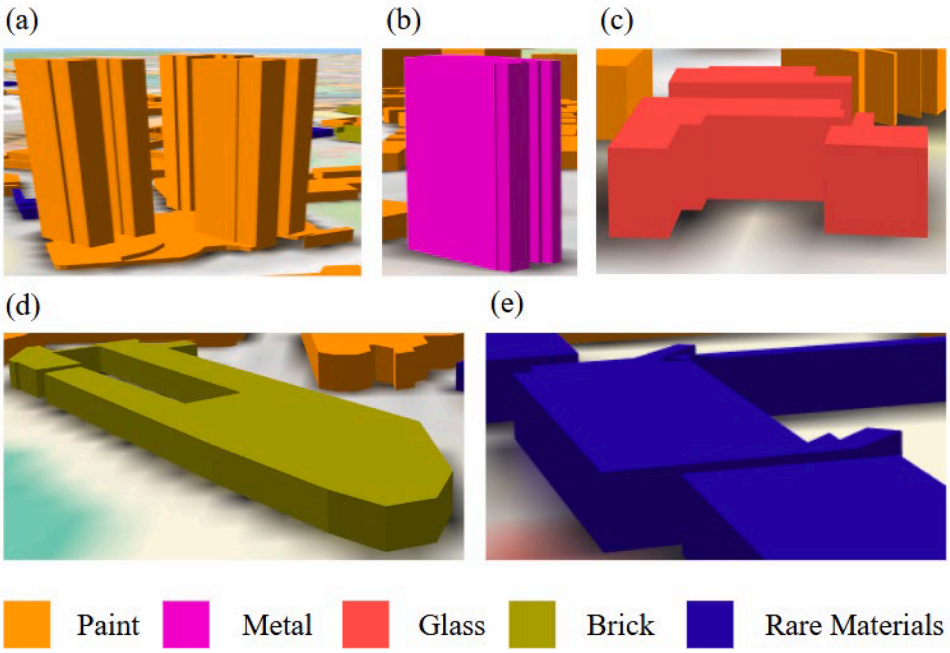


Fig. 15. Geo-visualization of the 3D building model in Bishan and Toa Payoh area.

5. Discussion

The proposed research framework effectively combines an advanced multi-scale model that proficiently handles the tradeoff between high detail and large receptive fields, together with a cohesive and interconnected training procedure that harnesses good strategies like transfer learning and cross-validation. As a result, the research framework can precisely segment façade materials from SVIs of varying and complicated environments and identify spatially heterogeneous albedos that can be projected into 3D building models. Additionally, the framework can either be used directly to produce predictions of façade materials in other cities or adapted to identify façade materials specific to a city by changing the labeling principles, making it extremely robust.

The implication of our research stretches far beyond the realms of just street view image analysis or deep learning. Compared to other studies, our framework harmonizes all aspects of the data lifecycle including data collection, data preparation, analysis, modelling, and publication to form a uniquely adaptable solution that can generate albedo information on a city-wide scale. Furthermore, the comprehensive design of our framework places greater emphasis on the aftermath of training than other studies by (i) adapting the data collection and preparation sections to produce metadata that can be used with predictions and (ii) discussing how the predictions and metadata can be used in correspondence with a 3D model. Summarily, the conception of our framework facilitates the advancement of 3D models via the provision of albedo information through a well-defined and efficient pipeline. Most importantly, more studies can utilize such advanced 3D models to develop more well-informed solutions for urban planning and energy efficiency analysis.

Our framework is significant in three aspects. The first is that the segmentation framework can contribute to urban sustainability initiatives, particularly in estimating solar photovoltaic potential, assessing light pollution, and analyzing the urban heat island effect. By accurately identifying façade materials, the model enables more precise solar reflectance and absorption calculations, which are critical for optimizing photovoltaic panel placement and predicting energy yields in dense urban environments. Secondly, the classification of glass and reflective surfaces can support studies on artificial light dispersion, helping to assess the impact of illuminated buildings on night-time brightness and ecological disruption. Thirdly, the differentiation of materials such as concrete, metal, and brick allows for better thermal property estimation and provides insights into heat retention and dissipation patterns across cityscapes in the context of urban heat islands. These applications demonstrate the broader significance of façade segmentation beyond visual classification, reinforcing its value for environmental assessment and sustainable urban planning.

Our study has three uncertainties: limitations in dataset diversity and labelling assumptions, segmentation accuracy, and evaluation robustness. First, the dataset primarily consists of a specific geographic region, limiting the model's adaptability to different architectural styles. In addition, labelling assumptions, such as restricting buildings to three façades, may reduce annotation accuracy in complex urban settings. Secondly, the model's accuracy can be improved since current panoramic images introduce distortions that affect segmentation performance, while class imbalances hinder the model's ability to identify rare façade materials. Thirdly, the evaluation process lacks multiple training runs and statistical significance testing, making assessing the model's true generalizability difficult. Future work will address these uncertainties by expanding the dataset to incorporate diverse architectural styles, refining labelling practices to improve annotation quality, and mitigating distortions by splitting panoramic images into smaller sections based on calculated 90° segments. Class imbalances will be tackled using augmentation techniques like CutMix and balanced dataset strategies while the robustness of the experiment can be improved through multiple training runs, statistical significance testing, and cross-dataset evaluations. Lastly, the 3D building model will be extended to include environmental factors like building temperatures and solar irradiation to enhance its relevance for urban planning. Nonetheless, the study demonstrates its effectiveness in façade material segmentation, even managing to project the segmentation outputs into a 3D building model, thus laying the groundwork for future advancements in large-scale urban analysis.

6. Conclusion

This study developed a novel framework that achieved high accuracy in the semantic segmentation of urban façades and their materials. The DL-based network presents encouraging and inspiring results,

demonstrating the effectiveness of combining multi-scale inputs with an encoder-decoder structure for urban scene analysis. The created SVIs and façade labels in central Singapore also provide an authentic representation of common materials in an urban landscape, which can be used in transfer learning to apply the model to other cities. The successful projection of façade material predictions onto 3D building models resolves the lack of spatially heterogeneous albedo information and presents more opportunities to obtain valuable insights for urban planning and energy efficiency evaluation. Our model has a satisfactory generalization capability, and our study is significant in 3D solar potential modelling, urban energy assessment, and smart city applications. Future work could (i) refine context transfer, (ii) expand datasets and review labelling ontologies for robustness, and (iii) improve rare material segmentation through appropriate evaluation tests and complex augmentation techniques.

CRediT authorship contribution statement

Jing Kai Daniel Tan: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis. **Rui Zhu:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Data curation, Conceptualization. **Jie Song:** Writing – review & editing. **Zheng Qin:** Writing – review & editing. **Yanqing Xu:** Writing – review & editing. **Yumin Chen:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

R. Zhu thanks the funding support from A*STAR Career Development Fund (C243512020).

Data availability

Data will be made available on request.

References

Assouline, D., Mohajeri, N., & Scartezzini, J.-L. (2017). Quantifying rooftop photovoltaic solar energy potential: A machine learning approach. *Solar Energy, 141*, 278–296. https://doi.org/10.1016/j.solener.2016.11.045

Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2013). OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics, 32*(4), 111. https://doi.org/10.1145/2461912.2462002. Article.

Boccalatte, A., Fossa, M., & Ménézo, C. (2020). The best arrangement of BIPV surfaces for future NZEB districts while considering urban heat island effects and the reduction of reflected radiation from solar façades. *Renewable Energy, 160*, 686–697. https://doi.org/10.1016/j.renene.2020.07.057

Calcabrini, A., Ziar, H., Isabella, O., & Zeman, M. (2019). A simplified skyline-based method for estimating the annual solar energy potential in urban environments. *Nature Energy, 4*(3), 206–215. https://doi.org/10.1038/s41560-018-0318-6

Cao, Y., Zhang, M., Liu, W., Song, J., Wang, Z., Li, X., & Li, C. (2021). Progress of 3D printing techniques for nasal cartilage regeneration. *Aesthetic Plastic Surgery, 46*(2), 947–964. https://doi.org/10.1007/s00266-021-02472-4

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, Q., Li, Y., Tang, J., Wang, Q., Lin, H., & Chen, Z. (2022). Automatic and visualized grading of dental caries using deep learning on panoramic radiographs. *Multimedia Tools and Applications, 82*(15), 23709–23734. https://doi.org/10.1007/s11042-022-14089-z

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3213–3223). https://doi.org/10.1109/CVPR.2016.350

Dana, K. J., Van Ginneken, B., Nayar, S. K., & Koenderink, J. J. (1999). Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics, 18*(1), 1–34. https://doi.org/10.1145/300776.300778

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Fan, Z., & Biljecki, F. (2024). Nighttime street view imagery: A new perspective for sensing urban lighting landscape. *Sustainable Cities and Society, 116*, Article 105862. https://doi.org/10.1016/j.scs.2024.105862

Fu, H., Xu, Y., Lin, S., Zhang, D., & Lao, S. (2018). Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Transactions on Medical Imaging, 37*(7), 1597–1605. https://doi.org/10.1109/TMI.2018.2791488

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3141–3149). IEEE. https://doi.org/10.1109/CVPR.2019.00326.

Han, X., Yu, L., Fan, X., Jiang, J., Yang, Y., & Zhang, Y. (2024). WHU-Urban3D: An urban scene LiDAR point cloud dataset for semantic instance segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing, 209*, 500–513. https://doi.org/10.1016/j.isprsjprs.2024.02.007

Huang, H., Lin, M., Ruan, D., Huang, C., & Shao, X. (2020). UNet 3+: A full-scale connected UNet for medical image segmentation. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1055–1059). https://doi.org/10.1109/ICASSP40776.2020.9053405

Jakubiec, J. A., & Reinhart, C. F. (2013). A method for predicting city-wide electricity gains from photovoltaic panels based on LiDAR and GIS data combined with hourly daytime simulations. *Solar Energy, 93*, 127–143. https://doi.org/10.1016/j.solener.2013.03.022

Ji, A., Zhang, L., Fan, H., Xue, X., & Dou, Y. (2023). Dual attention-based deep learning network for multi-class object semantic segmentation of tunnel point clouds. *Automation in Construction, 156*, Article 105131. https://doi.org/10.1016/j.autcon.2023.105131

Kansal, K., Chandra, T. B., & Singh, A. (2024). ResNet-50 vs. EfficientNet-B0: Multi-centric classification of various lung abnormalities using deep learning. *Procedia Computer Science, 235*, 70–80. https://doi.org/10.1016/j.procs.2024.04.007

Kazerouni, I. A., Dooly, G., & Toal, D. (2021). Ghost-UNet: An asymmetric encoder-decoder architecture for semantic segmentation from scratch. *IEEE Access, 9*, 97457–97465. https://doi.org/10.1109/ACCESS.2021.3094925

Kong, G., & Fan, H. (2021). Enhanced façade parsing for street-level images using convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing, 59*(12), 10519–10531. https://doi.org/10.1109/TGRS.2020.3035878

Li, L., Luo, F., Zhu, H., Ying, S., & Zhao, Z. (2016). A two-level topological model for 3D features in citygml. *Computers, Environment and Urban Systems,, 59*, 11–24. https://doi.org/10.1016/j.compenvurbsys.2016.04.007

Liang, Y., Wakaki, R., Nobuhara, S., & Nishino, K. (2022). Multimodal material segmentation. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 19768–19776). https://doi.org/10.1109/CVPR52688.2022.01918

Lin, C.-Y., Chiu, Y.-C., Ng, H.-F., Shih, T. K., & Lin, K.-H. (2020). Global-and-local context network for semantic segmentation of Street View images. *Sensors, 20*(10), 2907. https://doi.org/10.3390/s20102907. Article.

Liu, W., Sun, Y., & Ji, Q. (2020). MDAN-UNet: Multi-scale and dual attention enhanced nested U-net architecture for segmentation of optical coherence tomography images. *Algorithms, 13*(3), 60. https://doi.org/10.3390/a13030060. Article.

Liu, D., Du, J., Li, C., Yu, C., & Zhang, M. (2024). Multi-unit stacked architecture: An urban scene segmentation network based on UNet and ShuffleNetv2. *Applied Soft Computing, 165*, Article 112065. https://doi.org/10.1016/j.asoc.2024.112065

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968

Mishra, A., Agnihotri, A. K., Pipil, S., Gaur, S., & Ohri, A. (2024). Surveying techniques for urban areas. In A. Kumar, P. K. Srivastava, P. Saikia, & R. K. Mall (Eds.), *Earth observation in urban monitoring* (pp. 69–91). Elsevier. https://doi.org/10.1016/B978-0-323-99164-3.00013-6.

Park, S., Lee, J., Kim, Y., Kim, J., Lee, H., & Lee, D. (2021). Prediction of solar irradiance and photovoltaic solar energy production based on cloud coverage estimation using machine learning methods. *Atmosphere, 12*(3), 395. https://doi.org/10.3390/atmos12030395. Article.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS), 32*, 8024–8035. https://doi.org/10.48550/arXiv.1912.01703

Riemenschneider, H., Krispel, U., Thaller, W., Donoser, M., Havemann, S., Fellner, D., & Bischof, H. (2012). Irregular lattices for complex shape grammar façade parsing. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1640–1647). IEEE. https://doi.org/10.1109/CVPR.2012.6247857.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28.

Sánchez, E., & Izard, J. (2015). Performance of photovoltaics in non-optimal orientations: An experimental study. *Energy and Buildings, 87*, 211–219. https://doi.org/10.1016/j.enbuild.2014.11.035

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sun, X., Zhang, Y., Chen, C., Xie, S., & Dong, J. (2023). High-order paired-ASPP for deep semantic segmentation networks. *Information Sciences, 646*, Article 119364. https://doi.org/10.1016/j.ins.2023.119364

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P., & Paragios, N. (2011). Shape grammar parsing via reinforcement learning. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2011.5995319. June 2011.

Walch, A., Castello, R., Mohajeri, N., & Scartezzini, J.-L. (2020). Big data mining for the estimation of hourly rooftop photovoltaic potential and its uncertainty. *Applied Energy, 262*, Article 114404. https://doi.org/10.1016/j.apenergy.2019.114404

Wang, W., Wang, S., Li, Y., & Jin, Y. (2021). Adaptive multi-scale dual attention network for semantic segmentation. *Neurocomputing, 460*, 39–49. https://doi.org/10.1016/j.neucom.2021.06.068

Wang, S., Park, S., Park, S., & Kim, J. (2024). Building façade datasets for analyzing building characteristics using deep learning. *Data in Brief, 57*, Article 110885. https://doi.org/10.1016/j.dib.2024.110885

Xing, Z., Yang, S., Zan, X., Dong, X., Yao, Y., Liu, Z., & Zhang, X. (2023). Flood vulnerability assessment of urban buildings based on integrating high-resolution remote sensing and street view images. *Sustainable Cities and Society, 92*, Article 104467. https://doi.org/10.1016/j.scs.2023.104467

Xu, F., Wong, M. S., Zhu, R., Heo, J., & Shi, G. (2023). Semantic segmentation of urban building surface materials using multi-scale contextual attention network. *ISPRS Journal of Photogrammetry and Remote Sensing, 202*, 158–168. https://doi.org/10.1016/j.isprsjprs.2023.06.001

Yan, L., Zhu, R., Kwan, M.-P., Luo, W., Wang, D., Zhang, S., Wong, M. S., You, L., Yang, B., Chen, B., & Feng, L. (2023). Estimation of urban-scale photovoltaic potential: A deep learning-based approach for constructing three-dimensional building models from optical remote sensing imagery. *Sustainable Cities and Society, 93*, Article 104515. https://doi.org/10.1016/j.scs.2023.104515

Ying, S., Chen, N., Li, W., Li, C., & Guo, R. (2019). Distortion visualization techniques for 3D coherent sets: A case study of 3D building property units. *Computers, Environment and Urban Systems,, 78*, Article 101382. https://doi.org/10.1016/j.compenvurbsys.2019.101382

Ying, Y., Koeva, M. N., Kuffer, M., & Zevenbergen, J. A. (2020). Urban 3D modelling methods: A state-of-the-art review. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 699–706. https://doi.org/10.5194/isprs-archives-xliii-b4-2020-699-2020. XLIII-B4-2020.

Zhang, H., Kang, M.-Y., Guan, Z.-R., Zhou, R., Zhao, A.-L., Wu, W.-J., & Yang, H.-R. (2024). Assessing the role of urban green infrastructure in mitigating summertime Urban Heat Island (UHI) effect in metropolitan Shanghai, China. *Sustainable Cities and Society, 112*, Article 105605. https://doi.org/10.1016/j.scs.2024.105605

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-net architecture for medical image segmentation. In *, 11045*. *Proceedings of the Deep Learning in Medical Image Analysis (DLMIA) and Multimodal Learning for Clinical Decision Support (ML-CDS) Workshops at MICCAI 2018* (pp. 3–11). Springer. https://doi.org/10.1007/978-3-030-00889-5_1.

Zhu, R., You, L., Santi, P., Wong, M. S., & Ratti, C. (2019). Solar accessibility in developing cities: A case study in Kowloon East, Hong Kong. *Sustainable Cities and Society, 51*, Article 101738. https://doi.org/10.1016/j.scs.2019.101738

Zhu, R., Wong, M. S., You, L., Santi, P., Nichol, J., Ho, H. C., Lu, L., & Ratti, C. (2020). The effect of urban morphology on the solar capacity of three-dimensional cities. *Renewable Energy, 153*, 1111–1126. https://doi.org/10.1016/j.renene.2020.02.050