ISPRS Journal of Photogrammetry and Remote Sensing 202 (2023) 158–168

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Semantic segmentation of urban building surface materials using multi-scale contextual attention network

Fan Xu [a], Man Sing Wong [a,*], Rui Zhu [b], Joon Heo [c], Guoqiang Shi [a]

[a] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, 999077, Hong Kong, China
[b] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, Singapore, 138632, Republic of Singapore
[c] School of Civil and Environmental Engineering, College of Engineering, Yonsei University, 134 Shinchondong, Seodaemungu, Seoul, 03722, Republic of Korea

ARTICLE INFO

Keywords:
Façade
Material
Segmentation
Deep learning
Multi-scale

ABSTRACT

Distributed solar photovoltaic (PV) harvesting is an effective way to collect solar energy in existing metropolitan cities. To facilitate the installation of PV modules at solar abundant locations, an accurate estimation of solar PV spatial potential is indispensable. Solar energy could be reflected on high-albedo building surfaces inside the urban canyon. However, using conventional ways to construct albedo datasets for different building surfaces is extremely labor-intense. In this work, we proposed to use semantic segmentation to identify façade materials from street view images. Due to the distinguishable features between materials in terms of the subtle texture and patterns rather than just their shapes and colors, identification requires more details from images, which makes multi-scale inference structure a promising solution. Compared with existing methods combining scales features at pixel-level, we proposed a novel Multi-Scale Contextual Attention Network (MSCA), using a Multi-Scale Object-Contextual Representation (OCR) block to exploit and combine contextual information from different scales in high dimensional layers. The experimental results show that the proposed model significantly outperforms the existing models, achieving a mean Intersection over Union (mIOU) of 70.23%. The results indicate that the MSCA can effectively obtain the materials information from street views and can be a reliable solution to providing urban albedo information for solar estimation.

1. Introduction

Solar photovoltaic (PV) systems are becoming increasingly popular in metropolitan cities. From 2018 to 2019, the solar energy produced in Hong Kong raised from 47 TJ to 74 TJ, with a significant increase of 57% in two years (Electrical and Department, 2021). To enhance the efficiency of PV equipment deployments, some studies estimate the urban PV potentials (Gassar and Cha, 2021; Choi et al., 2019), which provide an essential basis for energy policy decision-making and panel deployment in metropolitan cities (Zhu et al., 2019; Dehwah et al., 2018; Zhu et al., 2022c,b). However, most studies only consider urban areas as a two-dimensional plane, using satellite images and cloud coverage data to estimate the solar radiation received by rooftops (Walch et al., 2020; Park et al., 2021; Assouline et al., 2015, 2017). Multi-reflective solar radiation made by façades has been omitted in the estimation in city-wide studies, which should be deemed as a significant component of received solar radiation (Sánchez and Izard, 2015; Boccalatte et al., 2020). Although some studies have proposed 3D models for the estimation (Redweik et al., 2013; Calcabrini et al., 2019; Jakubiec and Reinhart, 2013; Li et al., 2016), radiation reflections are not incorporated in the calculation. In the research of Zhu et al. (2020), they incorporate reflection into the 3D model by applying a constant value to represent the albedo of all urban façades. This is challenged by the missing of the façade albedo information whereas using conventional ways to collect such dataset is exceptionally labor-intensive at a city scale.

In this study, we proposed to use a multi-scale contextual attention network to extract the material information from street views, which provides the opportunity to link the results with 3D models and hence improves the accuracy of solar estimation. In the past few years, previous works (Liu et al., 2017; Ma et al., 2020; Dai et al., 2019) have applied convolutional neural networks (CNN) to façade parsing, which have achieved better performances than traditional models (Gadde et al., 2016). However, several challenges still exist in the current research. Firstly, although there are some datasets for material identification (Bell et al., 2015; Sharan et al., 2009), no specific dataset

* Corresponding author.
E-mail addresses: fanfan.xu@connect.polyu.hk (F. Xu), ls.charles@polyu.edu.hk (M.S. Wong), zhur@ihpc.a-star.edu.sg (R. Zhu), jheo@yonsei.ac.kr (J. Heo), guoqiang.shi@polyu.edu.hk (G. Shi).

https://doi.org/10.1016/j.isprsjprs.2023.06.001
Received 12 July 2022; Received in revised form 30 May 2023; Accepted 1 June 2023
Available online 15 June 2023

is dedicated to façade material identification. Compared with common objects like carpets, hair, and sofa, façade materials have higher similarity in appearance and farther distance from the viewpoint. Secondly, most façade related research was conducted by utilizing non-street-level images (Liu et al., 2017; Ma et al., 2020), which simplifies complicated environments and obstacles and thus reduces the generalization of methods. Furthermore, the colors and shapes of different materials may have similar visible spectral characteristics in street view images. This makes identifying materials much more challenging than identifying façade components, like windows or balconies. This study aims to estimate surface albedo via building materials on street-level images. For better converting the results into 3D models in other research, this study chooses to use semantic segmentation instead of object detection to identify building materials.

In semantic segmentation models, it is critical to balance the network dimensions (i.e., width, depth, and resolution) (Tan and Le, 2019). Some studies use low-resolution images as input to cover a relatively larger receptive field and lower FLOPS (Richter et al., 2021). Façade materials identification has a high demand on pixel-level details to differentiate specific materials, which means remaining the image original resolution is crucial. In addition, higher-resolution inputs usually mean better performance in detecting small objects while lower is good for large ones. Using multi-scale inference is a popular way to handle the trade-off. Lin et al. (2017) used average pooling to combine the features between scales. Tao et al. (2020) proposed the attention mechanism to determine the weighted mask and then use the mask to trade off the information of different scales. However, it is still a pixel-level operation, which cannot fully exploit the contextual information of the output tensor.

In this paper, we presented a new dataset containing 2,003 street view images of Hong Kong. Different from typical façade related datasets (Korc and Förstner, 2009; Teboul et al., 2010; Riemenschneider et al., 2012), the proposed dataset is specifically developed for façade material identification with sufficient façade styles and complicated urban environments. Based on the reflectivity and visual distinguishability, the dataset divides common façade materials into six categories. Furthermore, a multi-scale contextual attention network (MSCA) is proposed for the façade segmentation. Compared with previous works (Chen et al., 2016; Tao et al., 2020), MSCA combines contextual information in high-dimension feature space to achieve feature-level fusion between scales. The main idea of our work is to use the attention layers to fuse hierarchical information in a revised Object-Contextual Representation (OCR) module (Yuan et al., 2019) rather than after the segmentation head of the network. This can significantly improve the contextual comprehending ability of the network and thus could better handle the trade-offs between high demand on details and contextual comprehension ability on large objects. The contributions of this study are listed below:

(1) This study presents a street-level dataset for façade material identification. This developed method is superior to using visual interpretation to acquire information on façade materials in a complex urban environment.
(2) To exploit contextual information at different scales efficiently, this study proposed a multi-scale contextual attention network. The experiments suggested that the proposed structure could increase the efficiency and robustness of obtaining urban albedo information.

The remainder of the paper is organized as follows. Section 2 reviews related work and current technology. Section 3 presents the assumption, data preparation, and details of our work. In Section 4, we evaluate the performance of the proposed method and compare it to the latest models. Conclusions and future work are discussed in Section 5.

2. Related work

2.1. Material recognition

Unlike object detection or scene understanding, it is challenging to find image features that can reliably distinguish materials since a material could span a diverse range of appearances. Early works on material recognition tend to identify textures from close-up pictures without any background. For instance, Dana et al. (1999) proposed a dataset, CUReT, collecting 61 material surfaces under more than 200 illumination and viewing conditions. Nevertheless, a class in CUReT only contains a single instance, which leads to poor generalization. The KTH-TIPS (Fritz et al., 2004) and KTH-TIPS2 (Mallikarjuna et al., 2006) provide more samples for about ten materials like corduroy, linen, and cotton. However, recognizing materials of real-world objects is more challenging than distinguishing close-up textures. Sharan et al. (2014) proposed a 1,000 images dataset, FMD, which contains complete objects in ten categories with a little irrelevant background interference. Subsequently, OpenSurfaces (Bell et al., 2013) released over 25,000 indoor images in regular view instead of using close-ups. Then, Bell et al. (2015) presented MINC with more uniform samples of materials, including 7,061 labeled material segmentations in 23 material categories.

However, most existing datasets only contain indoor objects. By contrast, façades in street views are often captured from further distances with blurrier details, a more complex environment, and less discriminating shapes.

2.2. Façade imagery analysis

Due to the advantages of easy access and crowdsourcing, street view images can provide datasets for façade imagery analysis. Gadde et al. (2016) proposed using an unsupervised clustering method, gathering simple rules as complex patterns instead of giving handcrafted grammar for parsing a specific façade class of façade. Liu et al. (2017) proposed a network with a symmetric regularizer that can make use of the structure of man-made architectures. The study reckoned that building façades have highly regularized shape priors and fine-grained details, this is the reason why directly applying standard deep learning approaches (Schmitz and Mayer, 2016) could not always yield the optimal results. Due to the high dependence of grammar-based approaches on prior knowledge and the poor performance of directly using existing segmentation models, Kong and Fan (2020) proposed to apply a novel CNN pipeline that combines pixel-wise segmentation and global object detection on a complicated street-level dataset to achieve a better generalization. However, the difference in shape and spectral characteristics between different materials (e.g., glass or tiles) is much smaller than that between façade components (e.g., windows or balconies). That makes extracting façade material categories more challenging than the existing identifying façade components task.

2.3. Multi-scale segmentation

In the latest pixel semantic segmentation networks, low output stride backbones are typically applied to resolve fine-grained details. However, the small receptive fields lead to poor performance when identifying large objects (Wei et al., 2017). To balance the trade-off, PSPNet (Zhao et al., 2017) uses the pyramid pooling module to aggregate the multi-scale context. Some studies use encoder–decoder structure and skip connection, like UNet (Ronneberger et al., 2015), to pass the contextual information between different depth layers. The DeepLabv3 (Chen et al., 2017) proposed atrous convolutions with multiple atrous rates in cascade or parallel ways, which can perceive the context more flexibly.

However, Chen et al. (2016) found that employing average or max pooling after feature extraction leads to features at each scale either

equally important or sparsely selected. They introduce the attention mechanism to fuse the features from scales to address this problem, which allows the model to focus on the most relevant scales adaptively. Based on this model, Tao et al. (2020) reduced the training cost by predicting a relative weighting between adjacent scales without introducing extra scales. Nevertheless, no matter whether the operation is max-pooling or attention-based, in the task of materials segmentation, the crucial problem is distinguishing which material a pattern comes from is challenging at the pixel level (Schwartz and Nishino, 2016). This study proposes to handle the contextual information at the feature level by applying the attention module before the segmentation head instead of employing it at the end of the network.

3. Methodology

This section introduces the dataset and the refined model. The model aims to identify the façade materials, especially in metropolitan cities like Hong Kong. Therefore, we cooperated with the Hong Kong Highways Department to developed a façades segmentation dataset with 2,003 street-level images. However, to adapt the model for complex styles of modern façades and variable solar conditions, some compromises on assumptions are introduced.

3.1. Assumption

Regarding the difficulties in identifying materials from RGB images, two assumptions were introduced to reduce the labeling difficulties and the corresponding labor cost.

- **Each building has at most two components.** As shown in Fig. 1(b), considering that the dominant materials of different parts of an individual building may vary greatly, this study divides buildings into several components. Given the prevalent building types in Hong Kong, this research assumes the specific number of components is two. As shown in Fig. 1, most buildings in Hong Kong can be roughly divided into three categories. (1) Free-standing buildings, which are primarily for residential functions. The entire façade of this type of building is usually made of consistent material. (2) Complexes, which typically consist of two parts, lower for commercial function while upper is a residential zone. The façade material of each component can be independent. (3) Special buildings like museums. The façade of special buildings can be irregular and novel.
- **Each component consists of only one primary material.** It can be observed from the images that most façades consist of more than one material, like glass windows, metal pipes, and decorated coating. In the current stage, it is technically challenging to obtain material information about all accessories on the façade. Besides, if more than one material is required to label, like the windows in 1(a) and the decoration in 1(c), alternating and irregular patterns would cost a considerable investment of time. To tackle this problem, this study ignores the non-dominant material and assumes each component consists of only one primary material. Considering the first assumption, it is assumed that the façade of an individual building is composed of two primary materials at most.

3.2. Dataset

Metropolitan cities typically have complex street environments and crowded buildings. However, most façade related datasets (Korc and Förstner, 2009; Teboul et al., 2011; Riemenschneider et al., 2012) only have single-view images with sparse buildings and few obstacles, which are not practical enough to support the analysis of metropolitan city façades. In this work, we select Hong Kong as the research site and present a dataset with street-level images that can be more representative of the actual environment of the metropolis.

Table 1
Comparison between the existing façade-related datasets.

	Size	Occlusion	Single view	Diversity of building style
eTRIMS	60	✗	✓	Low
ECP2011	104	✗	✓	Low
Graz2012	50	✗	✓	Low
FaçadeWHU	900	✓	✗	Medium
Proposed dataset	2003	✓	✗	High

3.2.1. Data specifications

The Highways Department used a vehicle-based mobile mapping system to collect data from 2017 to 2019, covering the areas of Shek-Mun, LaiChiKok, WestKowloon, NorthPoint, and Central in Hong Kong. These images include old residential buildings and flourishing business districts under different solar conditions. Cameras in 8 directions can ensure that the acquired data comes from multiple views. Then, we excluded those overexposed, repetitive, blurred, and illegible images. Finally, 2,003 images were selected to build the dataset. Among them, 1,463 (73.0%) are used for training, 360 (18.0%) for validation, and 180 (9.0%) for testing. In this work, we used labelme (Russell et al., 2008) to perform the labeling. After that, cross-checking was conducted to ensure the consistency and correctness of the annotation. Compared to the existing datasets (Korc and Förstner, 2009; Teboul et al., 2011; Riemenschneider et al., 2012; Kong and Fan, 2020), our dataset has the following merits.

(1) **Sufficient façade styles**: As shown in Table 1, existing façade-related datasets only have a small number of images for training and testing. For example, eTRIMS (Korc and Förstner, 2009) only has 60 annotated images, while ECP2011 (Teboul et al., 2011) has 104, and Graz2012 (Riemenschneider et al., 2012) consists of 50. The size of FaçadeWHU (Kong and Fan, 2020) is quite large, which is 900. Besides, current datasets only contain monotonous building types, making them less generalizable when applying the results to other cities. In this study, we selected 2,003 images and manually annotated more than 10 K buildings to provide sufficient façade styles in the dataset. Besides, the resolution of our dataset is 2046×2046, which is also higher than most of the existing datasets.

(2) **Complicated environment**: Typically, current façades related datasets only contain buildings with regular façade shapes that are captured from the front view of buildings. These images usually include only a few or no obstacles to avoid the occlusions in front of target buildings. In contrast, our street-level images contain complex foreground occlusions like trees, commercial advertisements, and dense traffic. Besides, various light conditions also affect the hue, brightness, and saturation of façades differently in the images. We believed this heterogeneous quality of images will improve the model's generalization.

3.2.2. Classes and annotations

As stated in the survey of HO et al. (2004), mosaic and ceramic tiles are the most common façade materials of domestic buildings in Hong Kong, which attributes to their self-cleaning property and economical maintenance cost. By contrast, commercial buildings in central business districts are typically high-rise, making glass and glass mixed facades more preferred, given their safety and superior appearance.

In this dataset, nine annotation classes are defined, including background, ceramic tile, glass, hybrid, metal, mosaic tile, paint, tree, and unidentified materials. Classes were selected based on their reflectivity, visual distinguishability, and annotation effort. This study did not include the classes that are too rare in the dataset. Given the difficulty of identifying materials from images, visual distinguishability is the most critical consideration when doing the labeling. The difference in reflectivity does not mean high visual distinguishability, especially in

(a) Free-standing (b) Complexes (c) Special

Fig. 1. Common building structures in our dataset.

(a) Ceramic tile (b) Glass (c) Hybrid (d) Metal (e) Mosaic tile (f) Paint

Fig. 2. Examples of façade material pictures for each class.

distancing views. That is why we compromised on the differential of materials reflectivity to some extent. Specifically, as shown in Fig. 2(e), façades consisting of tiny tiles are common in the residential zone. *Mosaic tile* is the most representative type. Although the tiles may have similar appearances, the consisted materials could be various, like ceramic, wooden, or brick-like. However, it is exceptionally challenging and labor-consuming to distinguish and label them. Thus, based on the visual distinguishability, we combined all façades with tiny tiles into the category *Mosaic tile*.

Similarly, it is difficult to distinguish whether a façade is made from marble or marble-like ceramic. As shown in Fig. 2(a), this study defines the *Ceramic tile* as the façades consisting of large tiles. *Hybrid* in the dataset is defined as the façades that are usually fifty-fifty made of glass and another material, as shown in Fig. 2(c). *Metal* includes metallic materials (e.g., aluminum plate or other alloy façades), typical in commercial areas. *Paint* is the façade with ordinary paint coating that can usually be seen in old communities. *Glass* refers to the most common office building like Fig. 2(b). Besides, this study labels those buildings that cannot be identified from images, typically caused by a long distance or severe occlusions, as *Unidentified materials*. Similar to the *Background* (the unannotated data), this study will not use the accuracy of *Background* and *Unidentified materials* to evaluate the performance of this model.

In the annotation work, we used field investigation and Google Street View to provide multiple perspectives helping verify the classification of façades.

3.3. Multi-scale contextual segmentation

3.3.1. Architecture

In this paper, we propose a multi-scale attention structure to understand the contextual information in high-level features. Compared with the latest pipeline (Tao et al., 2020), HRNet+OCRNet, we have made some innovative modifications to adapt our task: the multi-head attention (MHA) after HRNet, attention within OCRNet, and the residual block at the end. The former structure embeds the extracted features into diverse representation subspaces to obtain comprehensive perspectives on different classes. The attention within OCRNet helps the network adaptively focus on the most relevant semantic information at the feature level.

For more details, as shown in Fig. 3, we use HRNet-W48 (Sun et al., 2019) as the backbone of our network. Compared with ResNet-101 (He et al., 2016), it can significantly preserve the low-level information from the high-resolution images. The MHA projects the features obtained from the backbone to different representation subspaces. Each head of MHA represents a separate way to interpret the semantic information. After preliminary experiments, this study sets the heads as the number of classes to obtain optimal results. The output of MHA is regarded as the preliminary result and thus used for estimating the auxiliary loss. In the next step, the fine-tuned features from two scales are then transformed into the query and value matrices, and the feature from the high-resolution one is simultaneously used as the key matrix to calculate the attention score. Then the proposed multi-scale OCR could combine the significant contextual information by the scores and send it to the segmentation head. To further improve the performance of the model, we have attempted to make the segmentation head deeper. However, there are some drawbacks, i.e. amplified the degradation problem and lower the accuracy level. This study uses a segmentation head with a residual block to replace a deeper structure. At the end of the network, the output tensor is a probability map for different classes.

3.3.2. Multi-head attention

Instead of directly using the results from the backbone, this paper performs multiple attention functions to conduct linear projections, allowing the model interprets the features from different representations in parallel. The calculations can be performed as follows:

$$head_i = Attention_i(F_{backbone}) \tag{1}$$

$$MultiHead(F_{backbone}) = \rho(Concat(head_1, \ldots, head_N)) \tag{2}$$

Where $F_{backbone}$ is the feature obtained from the backbone in the size of [batch size, 720,256,256]. As shown in the lower left part of Fig. 3, multiple attention blocks are used to process $F_{backbone}$. The number of attention blocks, N, is 8, which has been tested to achieve optimal results in the preliminary stage of experiments. Then the information from all attention heads is packed together by convolutional layers $\rho(\cdot)$ and trained jointly. In the experiments, two different attention blocks are used in distinct modules. The attention in Multi-Heads Attention Module is conceptually similar to that of Tao et al. (2020).

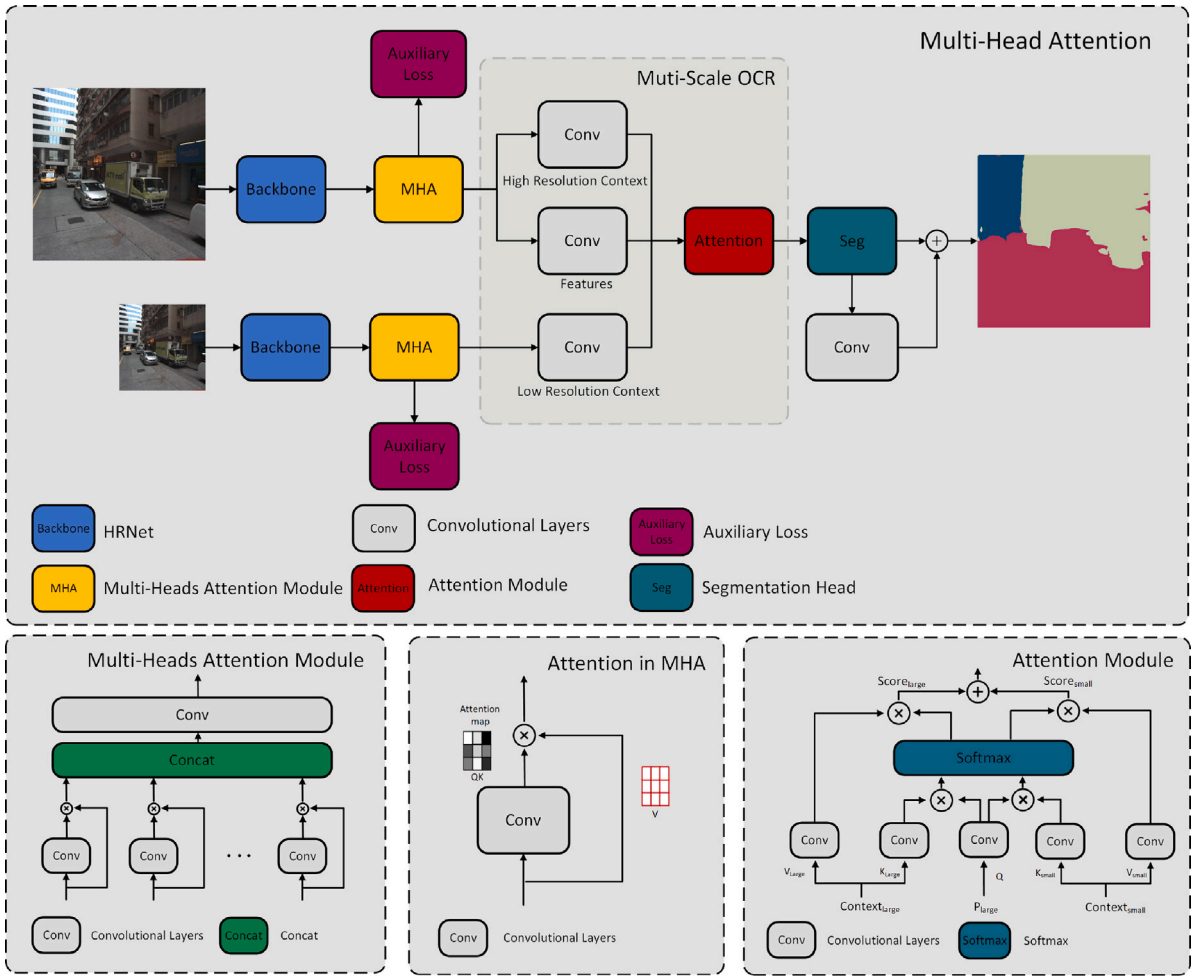

Fig. 3. Network architecture. The upper figure shows the overall architecture of the network. Images in two scales are fused in Multi-Scale OCR after HRNet and MHA, and the final result is obtained through a residual block. The details of the Multi-Heads Attention Module and Attention Module are shown in the bottom part. Specifically, two different attention blocks are used in distinct modules.

As shown in the middle plate of Fig. 3, it calculates a dense mask from features instead of query and key matrices; and then performs pixel-wise multiplication to obtain the final results. By contrast, the attention in Multi-scale OCR is a self-attention module. In this module, attention calculates scores of different scale inputs, then combines them after the softmax.

3.3.3. Multi-scale OCR

Multi-scale OCR aggregates the contextual information from scales at the feature level. Specifically, given a pair of input images features F_{large} and F_{small} that are generated by the MHA, we first calculated the pixel representation by Eq. (3):

$$P_{large} = g(F_{large}) \tag{3}$$

where F_{large} is the features obtained from the large size image. P_{large} is the corresponding pixel representation. $g(\cdot)$ is the convolution operations consisting of a 3×3 convolution layer, a batch normalization layer, and a ReLU layer. Then, the contextual information of large scale, $Context_{large}$, can be aggregated by Eq. (4):

$$Context_{large} = f(F_{large}, P_{large}) \tag{4}$$

where $Context_{large}$ is the object region representation in OCRNet. $f(\cdot)$ uses $Softmax(\cdot)$ to calculate the degrees of P belonging to each object region and then multiplies by F. In the next step, we use $Attention\ Module$, as shown in Fig. 3, to calculate the attention and combine the features from different scales:

$$Score_{large} = Attention(P_{large}, Context_{large}) \tag{5}$$

$$Score_{small} = Attention(P_{large}, Context_{small}) \tag{6}$$

$$MultiScaleFeature = r \cdot Score_{small} + Score_{large} \tag{7}$$

In Eqs. (5) and (6), the network uses self-attention to calculate the score between P and $Context$. Since the P from the large scale has more detailed information than the small one, this study uses P_{large} to calculate both scores with $Context_{large}$ and $Context_{small}$. As shown in the lower right of Fig. 3, P_{large} is used to calculate the query matrix and then estimate the attention by multiplying with key and value matrices from scales. To exploit the fine-grained details, the proposed network uses $Score_{large}$ as the major source of the $MultiScaleFeature$. $Score_{small}$, which has larger receptive fields, is regarded as an enhancement mask in this structure. This study introduces the drop-out function on the branch of $Score_{small}$. In Eq. (7), r is a vector of independent Bernoulli random variables, which is set as 0.5 in this study. This design could prevent over-fitting and reduce the sensitivity of the network on vaguer features from the small scale. The multi-scale OCR can be described as generating contextual information and using it to compute weighted output. The weights assigned to the values are determined by the relations of the pixel representation and the region representation from the multi-scale contexts.

Table 2
Details of experiment configuration.

Item	Configuration
Image scale	{1, 0.5}
Crop size	896 × 896
Batch size	2 per GPU
Learning rate	0.02
Optimizer	SGD
Momentum	0.9
Learning rate scheduler	Linear
Minimum learning rate	0.0001
Loss function	Cross-Entropy Loss

3.3.4. Loss function

In this work, we use cross-entropy loss as the loss function.

$$loss = -\sum_{i=1}^{N} y_i log(p_i) \tag{8}$$

where N is the number of classes, y_i is the binary indicator for class i, and p_i represents the predicted result of class i. The total loss is computed by Eq. (9).

$$loss_{total} = \alpha \cdot loss^{s}_{aux} + \beta \cdot loss^{l}_{aux} + loss_{main} \tag{9}$$

Where, α and β are the weights of auxiliary loss. In this study, α and β are defined as 0.5, the same value as Hierarchical MSA (Tao et al., 2020) The auxiliary loss $loss^{l}_{aux}$ and $loss^{s}_{aux}$ are calculated by the preliminary results S_l and S_s, respectively. $loss_{main}$ means the loss of the final output.

4. Experiment

In this section, we introduce the training details and present the experimental results. The experimental results from the proposed model were compared with that from the latest algorithms on our Hong Kong street view dataset and FaçadeWHU. The results show that the proposed model performs better on both datasets. In addition, the ablation studies are conducted to analyze and discuss the contributions of each sub-module. The results show that the proposed modules in our network are effective.

4.1. Training details

This paper uses PyTorch (Paszke et al., 2019) to develop the proposed model and conduct the experiments on a server containing 2 TITAN RTX GPU. The input images of the pipeline are defined in two scales, 1.0x for better details and 0.5x for the larger receptive field. Due to the high computational cost, the experiments crop the images to 896 × 896 and set the batch size as 2 per GPU. We used the Stochastic Gradient Descent (SGD) for the optimizer with a momentum 0.9 and weight decay 0.0001 in training. After comparing the polynomial decay with power 1.0 and 2.0, we selected power 1.0, linear decay, as the scheduler the same way as Tao et al. (2020). The other configurations are listed in Table 2.

To verify the effectiveness of the model, we compared the proposed MSCA with the following algorithms: DeepLabv3 (Chen et al., 2017), DeepLabv3+ (Chen et al., 2018), OCR (Yuan et al., 2019), and Hierarchical MSA (Tao et al., 2020). Hierarchical MSA achieved the optimal results on the Cityscapes validation set. In this experiment, the baselines and the backbone of MSCA are first pretrained on Cityscapes and then fine-tuned on the proposed dataset with similar configurations.

4.2. Metrics

To evaluate the performance quantitatively, we selected mIOU, precision, recall, and F1-score to analyze the experimental results. Furthermore, we use macro-averaging (Sokolova and Lapalme, 2009) to estimate the mean values of the metrics. For the most evaluations in this paper, we do not take *background* and *unidentified materials* into consideration.

In addition, to discuss the relationship between receptive field and the performance on this task, this study calculates the theoretical receptive field (TRF) of pixels at the network output layer. TRF describes the maximal area in the input image that can impact a pixel in a specific layer. It can be computed by Eq. (10):

$$r = \sum_{l=1}^{L}((k_l - 1)\prod_{i=1}^{l-1} s_i) + 1 \tag{10}$$

where r is the receptive field size of the network. k_l is the kernel size of layer l. s is stride. The actual impacted area, known as the effective receptive field (ERF), is typically smaller than TRF (Gu and Dong, 2021). The specific ERF depends on the information utilization ability by different networks.

4.3. Experimental results

First, this study conducts experiments on the proposed dataset. As shown in Table 3, the overall performance of the proposed model is significantly higher than others, with a mIOU of 72.58%. Besides, from the TRF of models, it can be seen that there is no significant linear correlation between the size of the receptive field and the performance. In the dataset, the size of most buildings varies from hundreds of pixels to about two thousand pixels. Considering the gap between TRF and ERF, Hierarchical MSA and MSCA, which have the TRF closer to the building size, obtained better results. It demonstrates multi-scale structure could bring a better understanding of different level details.

Except for the *ceramic tile* and *metal*, MSCA outperforms the baselines in the rest of material classes. Specifically, for the class *metal*, since we only have a very small amount of *metal* façades data for training, the performance of the proposed models on *metal* is unsatisfactory, only 64.48%, which is the poorest among all models. However, Fig. 4 also shows that the most mismatched metal pixels are identified as background instead of other materials, which is deemed as reasonable for building-level segmentation results. Furthermore, all models achieve their worst results in the classes *hybrid* and *ceramic tile*. DeepLabV3+ performs an IOU of 39.90% on *hybrid* and 46.71% on *ceramic tile*, while the proposed model results are 58.44% and 50.40%, respectively. However, the samples of *hybrid* and *ceramic tile* are sufficient compared with *metal*. The vague label principle of *hybrid* is the primary reason for the bad performance. The models seem challenging to classify a ceramic-like and glass-like object as the class *hybrid*. Due to this reason, as shown in Fig. 4, *ceramic tile* is the category where *hybrid* is most likely to be incorrectly marked. In contrast, for the painted façades, the proposed model reaches the best result of 86.88%, while others also achieve their best performances. The reason is that the painted façades have the most considerable number of samples in the dataset and more colorful appearances than others, making them more easily to be recognized.

Fig. 4 presents the percentage matrix that the ground-truth (GT) pixels are predicted (Pred) as different classes. As shown in the figures, the proposed model is prone to label pixels as background. Notably, except for *metals*, about 3% to 5% of pixels in other categories are misclassified as background. This is mainly due to many occlusion in the street view images. Because of the building-level annotation principle, we included the obstacles, e.g., advertisements, as a part of façades. That may confuse the network when identifying the exclude and unknown objects on façades.

Table 3
Performance of MSCA versus Baselines based on the constructed dataset. Best results in each class are represented in bold.

Method	Backbone	TRF	Ceramic tile	Glass	Hybrid	Metal	Mosaic tile	Paint	Tree	mIOU
DeepLabV3	ResNet-101	3459 × 3459	54.50	71.13	54.59	68.96	69.41	85.52	**80.58**	64.25
DeepLabV3+	ResNet-101	3583 × 3583	46.71	63.08	39.90	67.22	65.65	84.20	79.57	60.91
OCR	HRNet-W48	1087 × 1087	**59.48**	73.53	53.43	**74.12**	68.67	84.17	77.95	65.28
Hierarchical MSA	HRNet-W48	2302 × 2302	53.47	66.59	46.43	67.91	68.51	84.52	75.76	69.31
MSCA(ours)	HRNet-W48	2558 × 2558	55.40	**76.46**	**58.44**	64.48	**70.09**	**86.88**	75.95	**72.58**

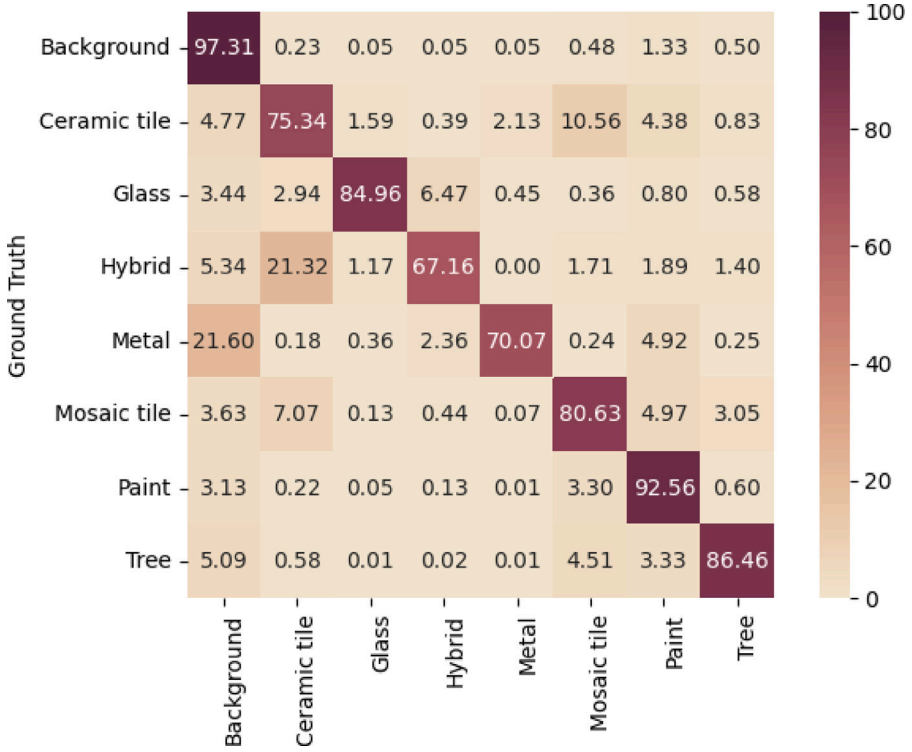

Fig. 4. The percentage of pixels that are classified into different classes. Rows represent the total pixels of this material (Ground truth). Columns represent all pixels classified into this material (Predicted class).

Table 4
Metrics of the proposed method on the Hong Kong street views dataset.

Metrics	Ceramic tile	Glass	Hybrid	Metal	Mosaic tile	Paint	Tree	Mean
Precision	0.75	0.85	0.67	0.70	0.81	0.934	0.86	0.80
Recall	0.68	0.88	0.82	0.89	0.84	0.93	0.86	0.84
F1-score	0.71	0.87	0.74	0.78	0.82	0.93	0.86	0.82

Besides, the results of hybrid façades are also related with the annotation principle. As mentioned, hybrid façades are usually fifty-fifty made of glass and other materials, typically ceramic. In that case, the proportion of materials becomes essential for judgment. Based on the ratio of materials shown in the street views, a ceramic-glass hybrid façade could be classified as ceramic tile, glass, or hybrid. However, it is subjective when the proportion is ambiguous, like 30% or 40% of glass, not 50%. This type of error causes 6.47% of glass to be mislabeled as hybrid façades, accounting for 43.02% of the misclassification of glass. Similarly, 21.32% hybrid are inferred as ceramic tiles, accounting for 64.92% of the mis-classification of hybrid. This makes hybrid has the lowest precision of 0.67 (Table 4).

About mosaic tiles, as shown in Fig. 5, the most challenging problem is that due to the erosion, fading, and distance, the grids of tiles could not be recognized confidently from images, which makes them hardly distinguished from *metal*, *paint* and some *ceramic tile*. In addition, the typical color and unique pattern of falling off tiles can also provide hints. However, there are still a considerable amount of mosaic tiles without any obvious features. That leads to the network misjudging them as painted façades. As shown in Fig. 4, 12.04% of mosaic tiles are labeled as paint or ceramic and this type of error accounts for 62.16% of all misclassification in mosaic tiles. For the same reason, 3.30% of painted façades are confused with mosaic tiles, accounting for 44.35% of the misclassification of paint. Nevertheless, because of the sufficient training data on these two categories, the proposed model still performs well with 0.82 and 0.93 F1-score (Table 4).

According to Table 3, Hierarchical MSA outperforms other state-or-the-art methods. The performance of the proposed method and the

Table 5
Performance of MSCA versus Baselines based on FaçadeWHU. Best results in each class are represented in bold.

Method	Window	Door	Wall	Balcony	Roof	Shop	mIOU
DeepLabV3	42.78	19.08	**61.82**	29.15	43.93	19.53	44.27
DeepLabV3+	**45.40**	17.39	59.04	29.39	41.42	16.98	43.24
OCR	43.66	8.23	61.32	25.24	36.94	11.46	40.07
Hierarchical MSA	43.22	20.17	60.68	33.84	42.50	19.67	44.82
MSCA(ours)	44.68	**21.70**	61.26	**36.00**	**45.41**	**24.34**	**46.69**

Table 6
Metrics of the proposed method on FaçadeWHU.

Metrics	Window	Door	Wall	Balcony	Roof	Shop	Mean
Precision	0.54	0.28	0.72	0.43	0.56	0.28	0.47
Recall	0.73	0.49	0.81	0.69	0.71	0.64	0.68
F1-score	0.62	0.36	0.76	0.53	0.62	0.39	0.55

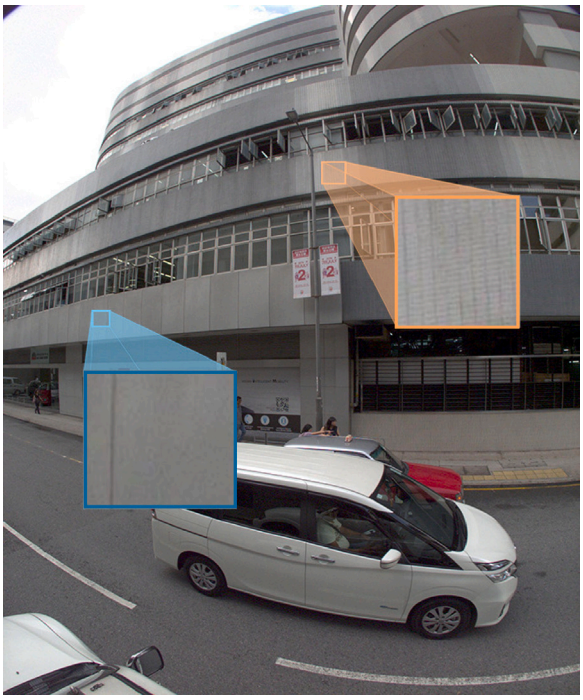

Fig. 5. Two different materials have almost the same color and luster. The lower left material is metal, and the upper right is mosaic tile. The difference between the two materials in the picture is only reflected in the pixel-level details, i.e., mosaic tiles have grids. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

comparison with Hierarchical MSA are given in Fig. 6. The significant difference between the proposed model and Hierarchical MSA is that we have adapted the attention module within the OCR module to fuse the features between scales. As shown in the first row of Fig. 6, ignoring whether the pixels are correctly inferred as its categories, the proposed model successfully detects the commercial building with the ceramic façades (left in the figure, colored in light blue) as a separated building. On the contrary, the baseline model only recognizes the upper part of the building and regards it as *hybrid* (color in light blue), same as MSCA. The lower part is reckoned as part of other residential buildings as the environment near the ground is more complex. Similarly, in the second row, the hybrid building behind the residential one (middle in the figure, colored in light blue) only shows a limited part in the picture. Hierarchical MSA thus fails to distinguish the two buildings, while MSCA correctly classifies the material and keeps its integrity well. In the third row, the architectural style of the buildings on the left side (colored in orange) is similar to that on the right, so it is difficult for the models to separate them. Likewise, in the fourth row, the baseline fails to identify the mosaic façade in the middle (colored in orange) as an independent structure for the same reason. In conclusion, the qualitative results show that MSCA has a more powerful structure for comprehending contextual information than the baseline.

To verify the effectiveness of the proposed model, this study also conducts experiments on FaçadeWHU. As shown in Table 5, the proposed model achieves the highest overall performance, with a mIOU of 46.69%, and outperforms the baselines in different classes, except for *Window* and *Wall*. Even in *Window* and *Wall*, MSCA is only 0.72% and 0.56% lower than the best model. Furthermore, compared with *Wall*, *Roof*, and *Window*, all methods have poor performances in *Balcony*, *Shop*, and *Door*. The best IOUs are only 36.00%, 24.34%, and 21.70%, respectively. As shown in Table 6, since *Wall* and *Roof* have the most expansive area, which makes them the most unlikely to be blocked by obstacles, the metrics of these categories are significantly higher than others. *Window* also has a relatively satisfactory performance due to its regular shape. The model performs worst in *Shop* and *Door*, with the lowest precision of 0.28. The potential reasons leading to the poor results could be the insufficient data volume and the indefinable object boundaries. The latter requires a strong semantic comprehension ability of models.

Nonetheless, the experimental results on two datasets show that MSCA can handle the façade segmentation in street-level images robustly and efficiently .

Table 7
Quantitative results of the ablation studies.

Ablation	Multi-Scale	MHA	Residual block	mIOU
I			✓	70.43
II	✓		✓	71.30
III		✓	✓	70.27
IV	✓	✓		71.61
MSCA	✓	✓	✓	72.58

4.4. Ablation study

This paper conducted some ablation studies to demonstrate the effectiveness of different modules in our network. Compared with a simple HRNet+OCRNet structure, four significant modifications are developed to adapt our task: multi-scale, MHA after HRNet, attention within OCRNet, and the residual block at the end. Among them, the effectiveness of the attention within OCRNet is proved by comparing it with Hierarchical MSA, which adopts the attention module after OCRNet.

Besides, as shown in Table 7, we first adopted a single-scale pipeline without the MHA, achieving a mIOU of 70.43%, which is 2.15% lower than the proposed structure. Then, the multi-scale network provides a boost of 0.87% mIOU over the first one. It demonstrates that it is helpful in comprehending contextual information by using features in different scales. However, in setting 3, the model with MHA yields 70.27% mIOU, resulting in a 0.16% decrease over the first one. Although this number is minor, it shows that simply applying MHA in this model cannot lead to significant improvements. By contrast, comparing the ablation study II with the proposed network shows that employing MHA on the model with residual block and multi-scale structure can increase the performance by 1.28%, which means the MHA is still a strong component in the specific circumstance. The experimental results of the ablation study IV and the proposed method show that after adding the residual block at the end of the network, the performance increases 0.97%. The results suggest that the residual block could be helpful in fine-tuning the preliminary output of the network. In conclusion, by incorporating these modules into the network, the model results in a total gain of 2.15%, 1.28%, 2.31%, and 0.97% mIOU compared with the above settings, respectively, which means that all the modules are effective in this study.

5. Discussion and conclusion

This study proposes a multi-scale contextual attention network to handle the trade-offs between high demand on details, like materials spectral characteristics, and contextual comprehension ability on large objects, like keeping the building integrity. We selected Hong Kong as the research site and developed a street-level dataset to evaluate the performance of the proposed model. The experiments show that our model can effectively classify the materials and achieved a better result than the other models.

The impact of this study is beyond simply conducting the imagery analysis on street views. The proposed method can significantly reduce the domain gaps in façades' information collection, providing a reliable and sufficient data source for urban albedos. The obtained information with coordinates could further project to 3D GIS systems and improve

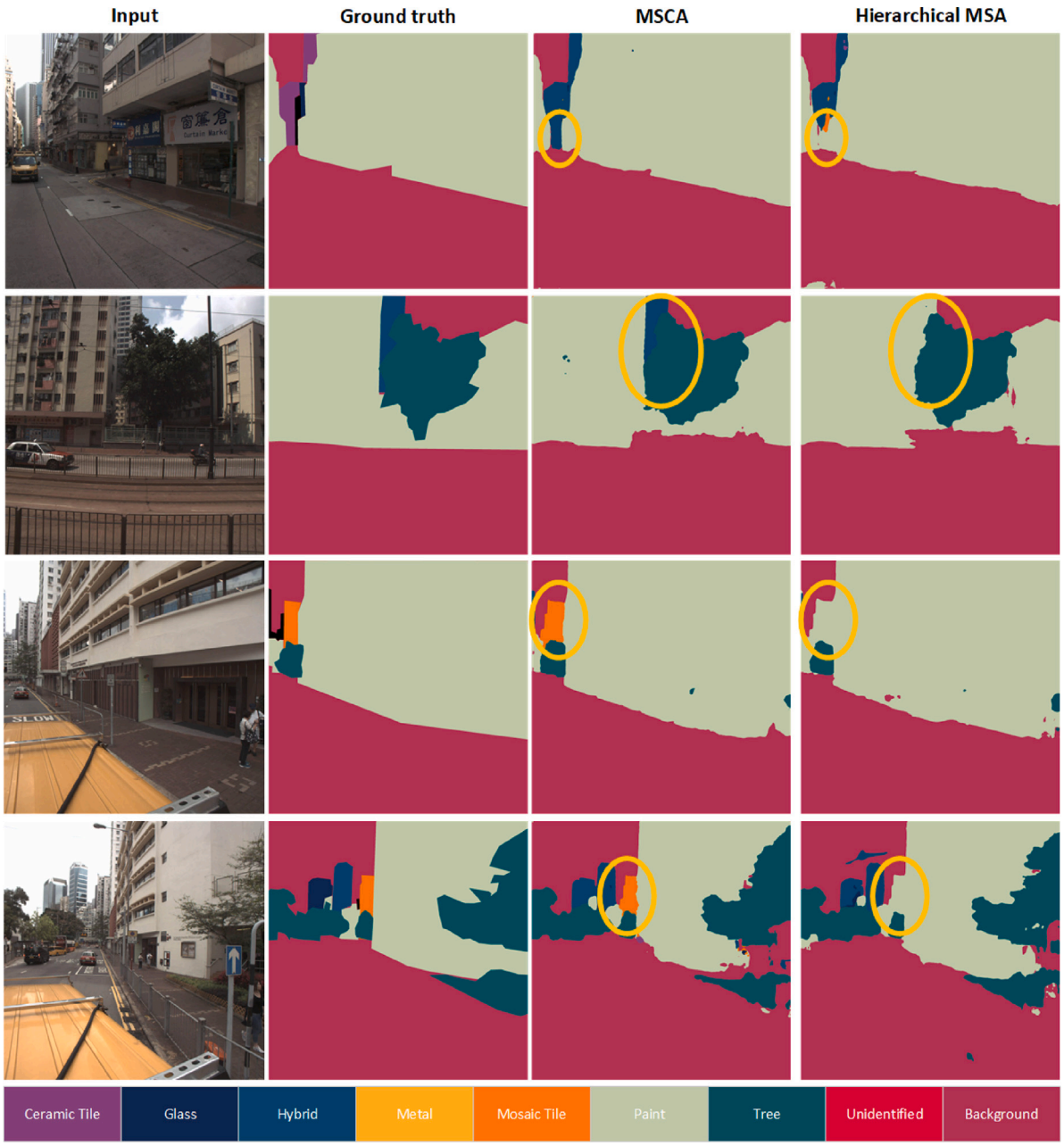

Fig. 6. Qualitative comparison between MSCA and strong baseline (Hierarchical MSA). From left to right: input, ground truth, our method, and baseline.

the accuracy of solar potential simulation. Since the simulation of reflected light is always the most challenging part of the indirect solar radiation estimation, ignoring the reflected solar or using a constant albedo to present the entire city is the typical solution that could bring considerable inaccuracy. This work explores the possibility of providing a more specific suggestion for PV deployment strategy. For instance, based on the model in Zhu et al. (2022a), we could obtain a better understanding of the relation between the urban morphology and solar capacity by incorporating the precise albedo of urban envelopes (i.e., rooftops, façades, and ground) in simulation.

However, there are still limitations to this study. First, due to the annotation cost, this study makes some compromises, assuming each building is only composed of two primary materials at most, which leads to the inconsistency on novel design buildings like theaters or museums. Besides, the building-level annotation also causes vague definitions like hybrid façades. That leads to some confusion in the classification process of the model. Secondly, due to the complex reflection characteristics of materials and the restriction of visual manner, we failed to classify the façade materials strictly by reflectivity. This limitation could diminish the potential usage of our work on solar potential estimation if we cannot supply highly albedo-differentiated results for each building. Thus, achieving a more fine-grain classification could be our future research direction.

To conclude, the proposed façade segmentation network can effectively identify the materials categories from street-level images in metropolitan cities like Hong Kong. Furthermore, this work provides a potential solution to precisely simulate the accumulative processes

of the reflective radiation and explores the possibility of conducting fine-grain urban analysis through street views.

CRediT authorship contribution statement

Fan Xu: Methodology, Software, Data curation, Writing – original draft. **Man Sing Wong:** Conceptualization, Supervision, Writing – review & editing, Resources. **Rui Zhu:** Supervision, Writing – review & editing. **Joon Heo:** Writing – review & editing. **Guoqiang Shi:** 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.

Acknowledgments

M.S. Wong thanks the funding support from the General Research Fund (Grant No. 15602619 and 15603920), and the Collaborative Research Fund (Grant No. C5062-21GF) from the Research Grants Council, Hong Kong, China. This study is also acknowledged the support from the Surveying Division in the Highways Department, HKSAR for providing the streetview images and GPS trajectories data.

References

Assouline, D., Mohajeri, N., Scartezzini, J.L., 2015. A machine learning methodology for estimating roof-top photovoltaic solar energy potential in Switzerland. In: Proceedings of International Conference CISBAT 2015 Future Buildings and Districts Sustainability from Nano To Urban Scale. (CONF), LESO-PB, EPFL, pp. 555–560.

Assouline, D., Mohajeri, N., Scartezzini, J.L., 2017. Quantifying rooftop photovoltaic solar energy potential: A machine learning approach. Sol. Energy 141, 278–296.

Bell, S., Upchurch, P., Snavely, N., Bala, K., 2013. OpenSurfaces: A richly annotated catalog of surface appearance. ACM Trans. Graph. 32 (4), 1–17.

Bell, S., Upchurch, P., Snavely, N., Bala, K., 2015. Material recognition in the wild with the materials in context database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3479–3487.

Boccalatte, A., Fossa, M., Ménézo, C., 2020. Best arrangement of BIPV surfaces for future NZEB districts while considering urban heat island effects and the reduction of reflected radiation from solar façades. Renew. Energy 160, 686–697.

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.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L., 2016. Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3640–3649.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Choi, Y., Suh, J., Kim, S.M., 2019. GIS-based solar radiation mapping, site evaluation, and potential assessment: A review. Appl. Sci. 9 (9), 1960.

Dai, M., Meyers, G., Tingley, D.D., Mayfield, M., 2019. Initial investigations into using an ensemble of deep neural networks for building façade image semantic segmentation. In: Remote Sensing Technologies and Applications in Urban Environments IV, Vol. 11157. International Society for Optics and Photonics, 1115708.

Dana, K.J., Van Ginneken, B., Nayar, S.K., Koenderink, J.J., 1999. Reflectance and texture of real-world surfaces. ACM Trans. Graph. 18 (1), 1–34.

Dehwah, A.H., Asif, M., Rahman, M.T., 2018. Prospects of PV application in unregulated building rooftops in developing countries: A perspective from Saudi Arabia. Energy Build. 171, 76–87.

Electrical, Department, M.S., 2021. Hong Kong Energy End-use Data. https://www.emsd.gov.hk/en/energy_efficiency/energy_end_use_data_and_consumption_indicators/hong_kong_energy_end_use_data/data/index.html.

Fritz, M., Hayman, E., Caputo, B., Eklundh, J.O., 2004. The kth-Tips Database. Citeseer.

Gadde, R., Marlet, R., Paragios, N., 2016. Learning grammars for architecture-specific facade parsing. Int. J. Comput. Vis. 117 (3), 290–316.

Gassar, A.A.A., Cha, S.H., 2021. Review of geographic information systems-based rooftop solar photovoltaic potential estimation approaches at urban scales. Appl. Energy 291, 116817.

Gu, J., Dong, C., 2021. Interpreting super-resolution networks with local attribution maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9199–9208.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

HO, D.C., Lo, S., Yiu, C., Yau, L., 2004. A survey of materials used in external wall finishes in Hong Kong. Ol. 15 Issue 2 December 2004.

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 daysim simulations. Sol. Energy 93, 127–143.

Kong, G., Fan, H., 2020. Enhanced facade parsing for street-level images using convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 59 (12), 10519–10531.

Korc, F., Förstner, W., 2009. eTRIMS Image Database for Interpreting Images of Man-Made Scenes. Tech. Rep. TR-IGG-P-2009-01, Dept. of Photogrammetry, University of Bonn.

Li, Y., Ding, D., Liu, C., Wang, C., 2016. A pixel-based approach to estimation of solar energy potential on building roofs. Energy Build. 129, 563–573.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Liu, H., Zhang, J., Zhu, J., Hoi, S.C., 2017. Deepfacade: A deep learning approach to facade parsing. IJCAI.

Ma, W., Ma, W., Xu, S., Zha, H., 2020. Pyramid ALKNet for semantic parsing of building facade image. IEEE Geosci. Remote Sens. Lett. 18 (6), 1009–1013.

Mallikarjuna, P., Targhi, A.T., Fritz, M., Hayman, E., Caputo, B., Eklundh, J.O., 2006. The kth-tips2 database. Comput. Vis. Active Percept. Lab., Stockholm, Sweden 11.

Park, S., Kim, Y., Ferrier, N.J., Collis, S.M., Sankaran, R., Beckman, P.H., 2021. Prediction of solar irradiance and photovoltaic solar energy product based on cloud coverage estimation using machine learning methods. Atmosphere 12 (3), 395.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Redweik, P., Catita, C., Brito, M., 2013. Solar energy potential on roofs and facades in an urban landscape. Sol. Energy 97, 332–341.

Richter, M.L., Byttner, W., Krumnack, U., Wiedenroth, A., Schallner, L., Shenk, J., 2021. (Input) size matters for CNN classifiers. In: International Conference on Artificial Neural Networks. Springer, pp. 133–144.

Riemenschneider, H., Krispel, U., Thaller, W., Donoser, M., Havemann, S., Fellner, D., Bischof, H., 2012. Irregular lattices for complex shape grammar facade parsing. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1640–1647.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T., 2008. Labelme: A database and web-based tool for image annotation. Int. J. Comput. Vis. 77 (1), 157–173.

Sánchez, E., Izard, J., 2015. Performance of photovoltaics in non-optimal orientations: An experimental study. Energy Build. 87, 211–219.

Schmitz, M., Mayer, H., 2016. A convolutional network for semantic facade segmentation and interpretation. Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci. 41, 709.

Schwartz, G., Nishino, K., 2016. Material recognition from local appearance in global context. arXiv preprint arXiv:1611.09394.

Sharan, L., Rosenholtz, R., Adelson, E., 2009. Material perception: What can you see in a brief glance? J. Vis. 9 (8), 784.

Sharan, L., Rosenholtz, R., Adelson, E.H., 2014. Accuracy and speed of material categorization in real-world images. J. Vis. 14 (9), 12.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45 (4), 427–437.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J., 2019. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tao, A., Sapra, K., Catanzaro, B., 2020. Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821.

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P., Paragios, N., 2011. Shape grammar parsing via reinforcement learning. In: CVPR 2011. IEEE, pp. 2273–2280.

Teboul, O., Simon, L., Koutsourakis, P., Paragios, N., 2010. Segmentation of building facades using procedural shape priors. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3105–3112.

Walch, A., Castello, R., Mohajeri, N., Scartezzini, J.L., 2020. Big data mining for the estimation of hourly rooftop photovoltaic potential and its uncertainty. Appl. Energy 262, 114404.

Wei, Z., Sun, Y., Wang, J., Lai, H., Liu, S., 2017. Learning adaptive receptive fields for deep image parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2434–2442.

Yuan, Y., Chen, X., Chen, X., Wang, J., 2019. Segmentation transformer: Object-contextual representations for semantic segmentation. arXiv preprint arXiv:1909.11065.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890.

Zhu, R., Anselin, L., Batty, M., Kwan, M.P., Chen, M., Luo, W., Cheng, T., Lim, C.K., Santi, P., Cheng, C., et al., 2022a. The effects of different travel modes and travel destinations on COVID-19 transmission in global cities. Sci. Bull. 67 (6), 588.

Zhu, R., Cheng, C., Santi, P., Chen, M., Zhang, X., Mazzarello, M., Wong, M.S., Ratti, C., 2022b. Optimization of photovoltaic provision in a three-dimensional city using real-time electricity demand. Appl. Energy 316, 119042.

Zhu, R., Kondor, D., Cheng, C., Zhang, X., Santi, P., Wong, M.S., Ratti, C., 2022c. Solar photovoltaic generation for charging shared electric scooters. Appl. Energy 313, 118728.

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. Renew. Energy 153, 1111–1126.

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 Soc. 51, 101738.