THE INTERSECTION OF TECHNOLOGY AND ARCHITECTURE: SMARTPHONE PHOTOGRAPHY IN CARBON ANALYSIS

IUAN KAI FANG[1] and SHEN GUAN-SHIH[2]
[1,2]*Department of Architecture, National Taiwan University of Science and Technology.*
[1]*iuankai.fang@gmail.com, 0009-0002-7693-2579*
[2]*sgshih@mail.ntust.edu.tw, 0000-0001-7108-9960*

Abstract. Our research introduces an innovative methodology that employs smartphone imaging for measuring dimensions and utilizes deep learning to estimate carbon emissions associated with facade materials. The dimensions of various components of building exteriors are obtained through smartphone imaging, and a network model on a cloud server automatically segments these components in the images, calculating their respective areas. By combining user-input material specifications such as thickness and density， with standard values of material carbon coefficients, estimations for each component's material carbon footprint are derived. This approach offers the advantage of individual estimations for diverse materials, aiding in the design of low-carbon facades. Additionally, it features a user-friendly interface enabling swift carbon estimation through portable devices. The method provides a convenient and efficient means for assessing carbon emissions in building facades, contributing to sustainable efforts and informed material selections for a greener future.

Keywords. Carbon Emission, Façade, Part Segmentation, Smartphone.

1. Introduction

Life Cycle Assessment (LCA) considers the carbon emissions associated with materials' production, transportation, installation, and usage phases. It calculates the overall carbon footprint of entire building materials, including those used for the exterior appearance of buildings. While the façade constitutes a part of the architectural framework, distinct from the main structure and mechanical systems, its design plays a pivotal role in enhancing environmental conservation, energy efficiency, and sustainable development. For instance, the Double-Skin Façades (DSF) design aims to enhance a building's energy efficiency, reduce environmental impact, and promote energy conservation and sustainability. (Andrea Zani et al., 2021). By isolating external climate influences, improving thermal performance, and minimizing energy wastage, it balances indoor comfort and environmental friendliness (Sabrina Barbosa & Kenneth Ip,2014) (Baldinelli, 2009).

However, improvement projects for façade in compliance with green building

ACCELERATED DESIGN, Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2024, Volume 1, 189-198.

standards often achieve energy-saving objectives without accounting for the carbon footprint generated by components enveloping the façade. This oversight might inadvertently harm environmental sustainability despite achieving energy-saving effects(Zahra S. Zomorodian & Mohammad Tahsildoost,2018).

According to the Ministry of Environment in Taiwan's Voluntary Disclosure Review (VDR) report for sustainability from 2020 to 2022 released in 2023, there's a severe shortage of certified personnel in carbon footprint assessment, leading to numerous companies facing challenges in obtaining carbon footprint certifications. When considering ESG regulations, the carbon footprint of existing corporate office spaces, factories, and other buildings serves as a significant evaluation criterion for a company's compliance with ESG standards. There's an urgent need for a rapid and convenient assessment of building carbon emissions.

In light of those issues, we are endeavoring to explore a path from the domain of computer vision. We aim to research a method capable of using smartphone imaging to capture real-world architecture, subsequently enabling the identification and quantitative assessment of visible exterior materials of buildings with different textures. The carbon footprint of a product can be calculated using methods such as direct monitoring instruments, energy and mass balance, or emission factors. However, within the construction industry, the emission factor method is commonly adopted. This method employs the fundamental formula: the carbon footprint of activity equals the activity data (mass/volume/kilowatt-hours/kilometers) multiplied by the emission factor (per unit of carbon dioxide equivalent). Mass equals volume multiplied by density; hence, once the volume and density of a material are known, its carbon footprint can be calculated.

2. Methodology

2.1. DEVELOPMENT PROCEDURE

The development process of the entire system begins with smartphone imaging and measurement of building dimensions. The captured images and dimensional information are then transmitted to a cloud server. Within this cloud server, there are two integrated systems: one is a proprietary model that improves upon the Generative Adversarial Network (GAN)(Ian Goodfellow et al.,2014) using Pix2Pix(Phillip Isola et al.,2017) as a baseline. It incorporates self-attention and self-proliferation mechanisms to enhance feature weights, accomplishing the semantic segmentation task within the images, and enabling precise identification of building components in photos. The other system is an expert system composed of prior knowledge. It encodes different component material thicknesses, densities, and carbon emission coefficients into the program. Segmented image blocks are analyzed within this system to compute individual material areas, estimating their areas based on RGB pixel points.

The user interface on the smartphone will display recognized building components, presenting a menu for users to select potential materials and specifications based on their professional construction experience, as illustrated in (Figure 1.). For instance, if the component is a window, the user can choose the possible thickness values for the glass. Then, the total area of the segmented and identified glass windows multiplied by the selected thickness provides the total volume. The cloud server matches the density and carbon emission coefficients of that material from the expert system, multiplying them accordingly, and transmits the results to the smartphone's user interface. The interface lists the estimated carbon emissions of various materials captured in the

smartphone's images of the facade, culminating in the final comprehensive data result.

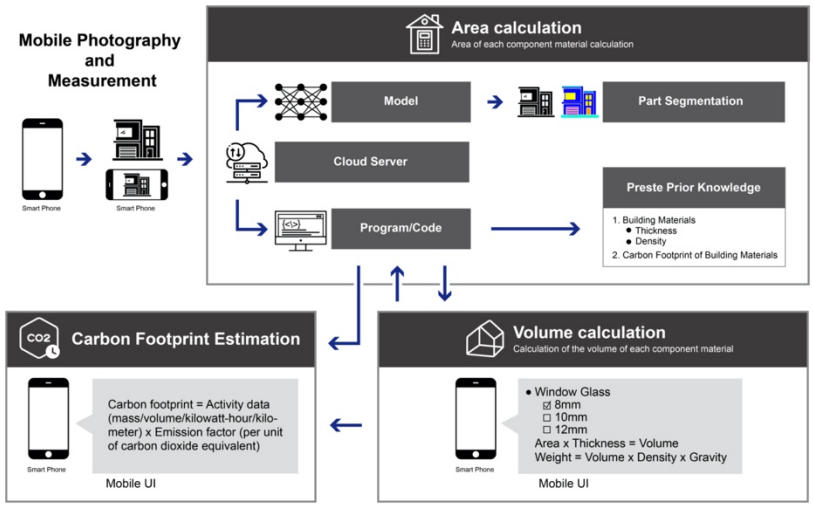

Figure 1. The technical development process diagram of our study.

Our research focuses primarily on the part segmentation of the facade throughout the development process. The functionalities of smartphone imaging and measurement are already well-established, and various countries have established standard reference carbon coefficients for different building materials. Hence, our emphasis lies in the automatic identification of building exterior components, followed by the computation of component quality. Assuming precise measurements of component quality, the accuracy of carbon emission estimation would also improve. Therefore, within the entire development process, the precision of part segmentation stands as our paramount objective in machine learning.

2.2. DATASET

The process of semantic segmentation involves assigning each pixel in an image to its corresponding class label. The value of semantic understanding lies in providing explanatory categorization for meaningful objects in the real world. In contrast to object detection and recognition, semantic segmentation achieves pixel-level classification of images. However, current semantic understanding has not shown substantial development and progress in the tasks of component identification and segmentation in real-world object images. This implies that apart from leveraging large-scale image datasets with diverse labels such as ImageNet (Deng, Jia, et al.,2009), ShapeNet(Angel X.Chang et al.,2015), COCO (Tsung-Yi Lin et al.,2014), ADE20K(Bolei Zhou et al.,2017), etc., for the recognition and understanding of real-world objects, there's still limited inference regarding the finer constituents and functionalities of these objects.

ImageNet, ShapeNet, COCO, ADE20K, and others are renowned large-scale datasets comprising over 200,000 images. However, in comparison to all the objects in the world, their image data for building exteriors is relatively limited. Moreover, part annotations for objects are absent in ImageNet, ShapeNet, COCO, and ADE20K. Our self-built dataset enhances the recognition of information regarding the exterior

components of buildings.

Firstly, we collected 500 photographs of building exteriors, forming what we call Raw Data. Next, we removed backgrounds and unnecessary obstructions from the building images, creating what we term Object Data(Figure 2). Subsequently, the components within the building exterior were categorized into ten main types: doorway, window, exposed beam, exposed column, exposed slab, projecting balcony slab, railing, eaves, decorative wall panel, and curtain wall. These ten types were manually annotated based on functional attributes, establishing what we refer to as Notation Data (Figure 2).

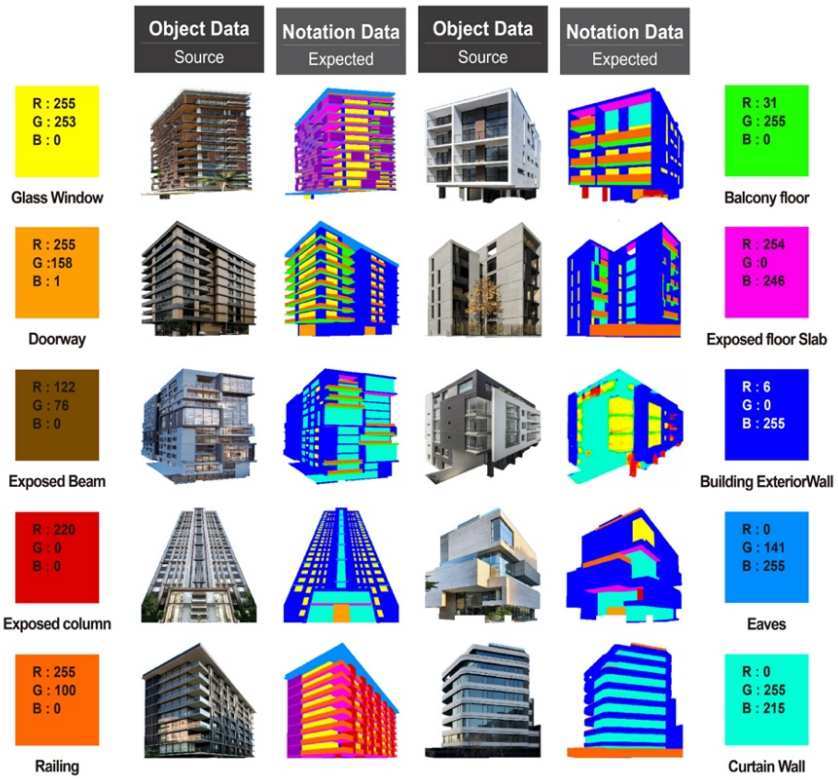

Figure 2. RGB Setting, Object data samples and Notation data samples.

The Façade dataset from UC Berkeley's official directory of Pix2Pix Datasets contains 506 photos of building façades alongside corresponding annotated images. However, aligning with our research objectives, capturing entire building façades conveniently with a mobile phone camera might not always be feasible due to perspective angles. Hence, the Façade dataset might not be entirely suitable for our machine-learning purposes. Moreover, the dataset lacks finer categorization for external building components. Additionally, the BuildingNet (Pratheba Selvaraju et al.,2021) dataset comprises 2,000 annotated architectural models, but it's a 3D model dataset, differing from our 2D real image data. The annotation methods also differ logically. Nonetheless, in the future, we could render BuildingNet's 3D architectural models into 2D images to expand our dataset. However, currently, we prefer using

actual images as our experimental foundation to approach real-world scenarios. Annotating real-world photos for component classification aims to enable our trained network models to be practically applied and achieve high precision in real-world scenarios. The genuine composition of real objects offers reliable volumetric forms for carbon footprint calculations.

2.3. MODEL

Once the components of a facade are accurately identified, estimating carbon emissions through the calculation of material coverage areas becomes feasible. This approach offers a rapid initial estimation for broad-scale carbon calculations. To achieve this goal, we propose a novel part segmentation model. Given Pix2Pix's room for improvement in accuracy regarding target details and boundaries, we introduced two techniques(Figure 3): self-proliferation(Yuan-Fu Yang & Min Sun,2021) and self-attention(Ashish Vaswani et al.,2017). The self-proliferation aids in generating meaningful feature maps, while self-attention provides a more refined way to enhance features for improved precision in semantic understanding and part segmentation.

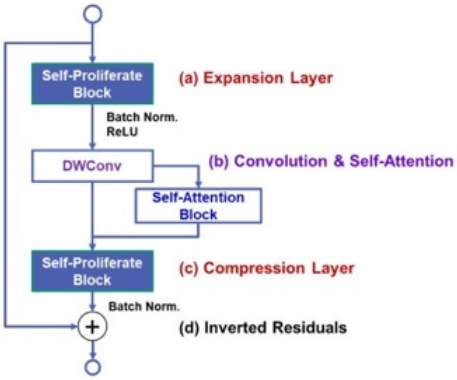

Figure3. incorporated self-attention and self-augmentation mechanisms into our baseline model.

Image segmentation requires positional information for each pixel. It demands full-resolution semantic estimation, making it impossible to reduce computational complexity using pooling or dilated convolution networks as in classification tasks. Therefore, we adopted an encoder-decoder model structure. In the Pix2Pix framework, the encoder part utilizes downsampling to reduce spatial resolution, generating low-resolution feature maps. We employed an embedded attention mechanism to alleviate the limitations of pooling. Furthermore, the self-expansion mechanism is an extension derived from MobileNet(Andrew G. Howard et al.,2017), a lightweight network structure based on depth-wise separable convolutions. We embedded it into the encoder to achieve low power consumption and enhanced speed.

Pix2Pix, developed in 2014, emerged from the Conditional Generative Adversarial Network (CGAN)(Mehdi Mirza & Simon Osindero. 2014.), specifically designed for image translation tasks. CGAN extends the basic GAN, enabling the generation of images that satisfy specific conditions or features, facilitating the transformation from patterns to images. Despite being an older development, Pix2Pix remains powerful in its image-to-image tasks and even exhibits faster output generation compared to the current prevalent Diffusion models. For large-scale estimations of carbon emissions based on building exteriors, the focus is on approximate calculations, eliminating the

need for excessive granularity in image quality. The emphasis lies in the immediate availability of information through mobile devices for real-time estimation of building carbon footprints, prioritizing rapid results. Hence, we chose Pix2Pix as the baseline model for development.

In addition to integrating the generator and decoder based on these foundational principles, this model architecture also incorporates the concept of U-Net (Ronneberger, O. et al.,2015).This involves establishing skip connections between multiple levels in the encoder and decoder to address information loss and resolution reduction issues in semantic segmentation tasks. Conventional encoder-decoder structures, due to repetitive downsampling and upsampling operations, tend to reduce resolution gradually, potentially leading to the loss of fine features. Skip connections allow rich semantic information transfer from low-level to high-level feature maps into the decoder, preserving more details while maintaining high resolution. By integrating CGAN's ability to conditionally generate meaningful feature maps, combining PatchGAN's (Phillip Isola et al.,2017) detailed discriminator, and incorporating U-Net's method for retaining fine features at high resolutions, Pix2Pix demonstrates flexibility and robust performance in various image-to-image translation tasks. These include both image-to-image and image-to-label transformation tasks.

Our network model is based on the Pix2Pix generative adversarial architecture, enhancing feature quality through improvements in the generator. Firstly, the model utilizes a series of linear transformations to generate additional feature maps with lower computational costs. This segment incorporates a self-augmentation mechanism. Subsequently, it captures the long-term dependencies of feature maps through self-attention mechanisms in both channel and spatial domains. Our model consists of a seven-layer Convolutional Neural Network (CNN)(Figure 4). As data passes through each layer of the CNN, the model duplicates the same number of feature maps and incorporates them into the self-attention mechanism, where the generated feature maps, processed through softmax, are added back to the original data.

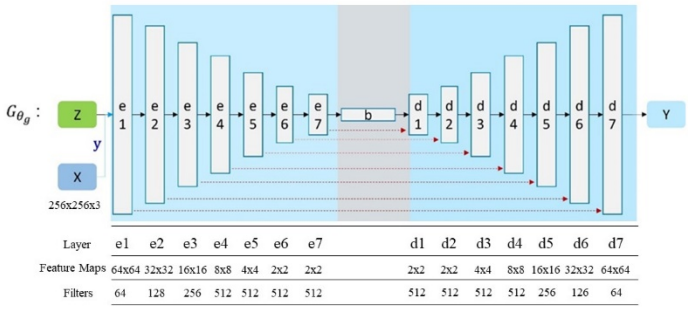

Figure 4. This diagram shows the size of feature maps and the number of filters in each layer of encoder and decoder.

3. Experiment

In the training process of machine learning, we continually adjusted several hyperparameters, including epoch, batch size, and learning rate. To achieve better computational efficiency, convergence, and prevent overfitting, we settled on setting 50 epochs for the model to train on different parts of the dataset each time. Additionally,

we fixed the Batch Size at 75, meaning that during each training step, the model randomly selects 75 images from the training data as a batch for gradient computation and weight updates. The gradients of the model's weights are calculated based on the loss function of these 75 images, and the model's weights are updated accordingly. Throughout our training process, both the generator and discriminator had a learning rate of 0.0002. We utilized the Adam optimizer and set the loss weight for the generator and discriminator to 0.5. For hardware specifications, we employed a TESLA T4 16G GPU and conducted the training using Keras.

About the evaluation metric, we utilized PSNR (Peak Signal-to-Noise Ratio) in our part segmentation experiments. A higher PSNR value indicates greater similarity between images, suggesting lower distortion. PSNR measures the ratio between the maximum possible power of a signal and the power of noise that affects its accuracy. In the context of images, PSNR offers a relatively objective and quantifiable means of assessing image distortion. This allows for a more comprehensive evaluation of the quality of our segmentation results.

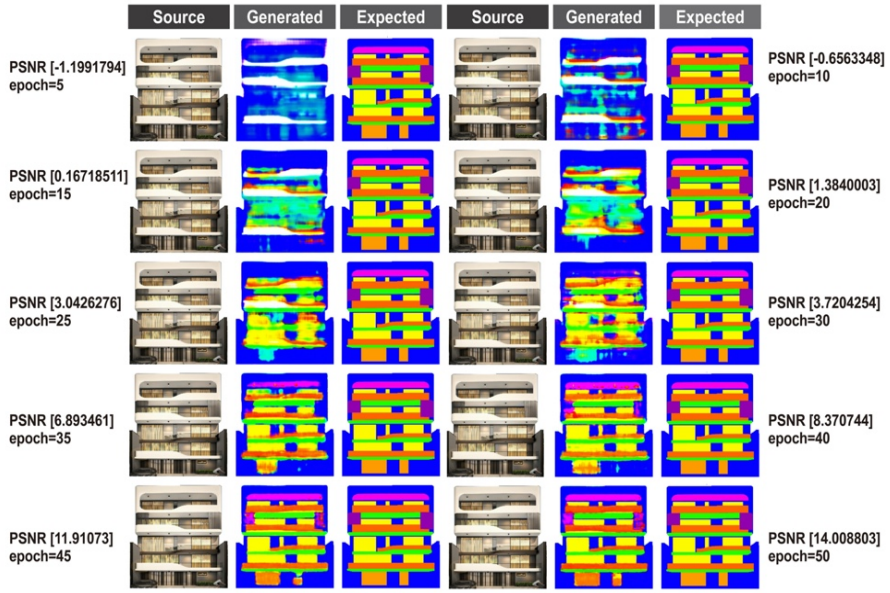

Figure 5. Segmented images selected at different training epochs in the recognition network.

We partitioned the data into training and testing sets with a ratio of 9:1. We evaluated the impact of different epoch counts on model improvement. As shown in (Figure 5), the results of part segmentation become increasingly similar to the expected values with more training epochs. Our model exhibits better PSNR values at 50 epochs, indicating its significant capability in the task of part segmentation.

In addition to training the machine using datasets, we conducted real-time tests on ongoing construction projects. After capturing the exterior of the buildings, we input the images into our self-developed convolutional neural network model. The automatically segmented results are illustrated in (Figure 6). We verified the results using a pixel-based comparison method, where the total sum of similar-colored pixels

represents the area of the component. In this case, the pixel comparison error rates for the main building's exterior walls and metal curtain walls were below 10% (Table 1). For windows and railings, the pixel comparison error rates were below 30% (Table 1). This suggests that when calculating areas based on pixel counts, the segmentation results closely approximate the actual component areas.

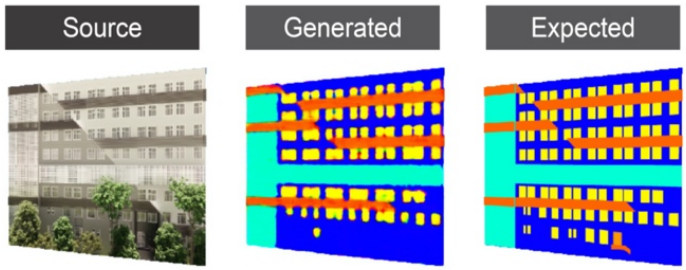

Figure 6. Mobile photography of an ongoing construction site and its generation result of part segmentation.

Category / Component	**RGB**	**Generated** / unit : Pixels	**Expected** / unit : Pixels	**Error Rate**
Glass Window	(255,253,0)	7946	11414	**32.9%**
Railing	(255,100,0)	6164	8499	**27.47%**
Curtain Wall	(0,255,215)	11194	12284	**8.87%**
Building ExteriorWall	(6,0,256)	22688	23780	**4.59%**

Table 1. We conducted a pixel-based analysis by comparing the segmented results from *Figure 6*. with the expected values to assess the error rates in pixel estimation. Pixel-based calculations provide estimations of areas, allowing us to compare them with actual area measurements.

4. Result & Discussion

4.1. PIXEL COUNTING VS. PSNR.

The task of our segmentation model is to accurately identify and delineate architectural components. The best performance of our component segmentation model in terms of PSNR values ranges between an average of 18-20. While not exceptionally high, the precision in comparison between detected RGB pixel values and actual color block areas is high. Consequently, from such experimental results, we observe that the precise shape of the color block segmentation isn't as crucial. What matters more to us is the area covered by these color blocks. If the overall area approximation tends towards reality, the area calculation becomes more realistic. Hence, in pursuit of automatically deriving areas through segmentation, the error rate in RGB pixel estimation holds higher importance and reference value compared to the distortion rate evaluated through PSNR between images.

4.2. DATASET REVIEW

During our research, we identified lower recognition rates for certain architectural components, such as the eaves. This is primarily due to the limited representation of

eaves data in our training dataset, resulting in what is known as an 'Imbalanced Dataset.' This refers to a significant disparity in the number of samples across different categories in the dataset, impacting the model's performance in predicting and classifying minority categories. Additionally, in semantic understanding, the definition of eaves might be multifaceted. Therefore, in labeling, we might need further delineation regarding forms and functions, such as styled eaves, shading eaves, and traditional sloping roof eaves, among others.

4.3. COMPARISON OF DATASETS

The Façade dataset from UC Berkeley consists entirely of Western traditional architectural facade photos, all limited to buildings with seven floors or less. These images exhibit a higher degree of order in façade design, making it relatively easier for the model to discern various components. In contrast, our dataset emphasizes the collection of photographs from modern architectural settings, where there's more design variation and a combination of diverse materials in components. This diversity presents considerable ambiguity in defining components, posing a significant challenge for model training. However, it is precisely for this reason that we need to gather a vast amount of data from modern architectural contexts to meet the current demand.

5. Conclusion & Future Work

Overall, our research provides a rapid and convenient way to perform preliminary assessments of carbon emissions in physical constructions: (1). It offers an initial estimation of the total carbon emissions from façade components made of composite materials, allowing a deeper understanding of an object's environmental impact. (2). It enables architects designing architectural exteriors to improve and optimize designs with a lower carbon footprint, encouraging sustainable practices. (3). It provides a highly convenient method where, through the camera and measurement functions on a mobile phone, one can upload data to a cloud server for computation and then transfer results back to the phone, enabling assessments of component material carbon emissions, addressing the shortage of experts in carbon footprint assessment.

The focus of our paper lies in our precise calculation of the area of architectural exterior components to obtain the material quantity necessary for carbon footprint calculations. Future research aims not only to enhance the segmentation model's accuracy but also to develop a complete system focusing on expert systems and user interfaces, facilitating practical industrial applications.

REFERENCES

Andrea Zani1, Oluwateniola Ladipo, Antonio D'Aquilio, Carmelo Guido Galante, Matthew Tee and Tania Cortés Vargas.(2021). Facade Design Process to Establish and Achieve Net Zero Carbon Building Targets. *8th International Building Physics Conference (IBPC 2021) 25-27 August 2021, Copenhagen, Denmark.*

Zahra S. Zomorodian, Mohammad Tahsildoost. (2018). Energy and carbon analysis of double skin façades in the hot and dry climate. *Journal of Cleaner Production. Volume 197, Pages 85-96.*

Sabrina Barbosa, Kenneth Ip (2014). Perspectives of double skin façades for naturally ventilated buildings:A review. *Renewable and Sustainable Energy ReviewsVolume 40, December 2014, Pages 1019-1029.*

G. Baldinelli,(2009). Double skin façades for warm climate regions: Analysis of a solution with an integrated movable shading system. *Building and Environment. Volume 44, Issue 6, Pages 1107-1118*

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. (2014). Generative Adversarial Nets.*Neural Information Processing Systems (NeurIPS).*

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. (2017).Image-To-Image Translation With Conditional Adversarial Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, Jia, et al. (2009). Imagenet: A large-scale hierarchical image database. *IEEE conference on computer vision and pattern recognition*.

Angel X.Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu.(2015). ShapeNet: AnInformation-Rich3DModelRepository.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár & C. Lawrence Zitnick(2014). Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision(ECCV)*, *Pages740-755.*

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, Antonio Torralba. (2017).Scene Parsing Through ADE20K Dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Pratheba Selvaraju, Mohamed Nabail, Marios Loizou, Maria Maslioukova, Melinos Averkiou, Andreas Andreou, Siddhartha Chaudhuri, Evangelos Kalogerakis.(2021) BuildingNet: Learning to Label 3D Buildings. P*roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

Yuan-Fu Yang, Min Sun. (2021). A Novel Deep Learning Architecture for Global Defect Classification: Self-Proliferating Neural Network (SPNet).*32nd Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC).*

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. (2017). Attention Is All You Need. *arXiv:1706.03762v7*

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam. (2017).MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *Conference on Computer Vision and Pattern Recognition (CVPR). arXiv.1704.04861.*

Mehdi Mirza, Simon Osindero. (2014).Conditional Generative Adversarial Nets. *Conference on Neural Information Processing Systems(NIPS).arXiv:1411.1784.*

Ronneberger, O., Fischer, P., & Brox, T. "U-net. (2015).Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention(MICCAI)pp. 234-241, Springer.*