Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.057118

ARTICLE

Enhancing Building Facade Image Segmentation via Object-Wise Processing and Cascade U-Net

Haemin Jung[1], Heesung Park[2], Hae Sun Jung[3] and Kwangyon Lee[4,*]

[1]Department of Industrial & Management Engineering, Korea National University of Transportation, Chungju, 27469, Republic of Korea

[2]Department of Industrial Engineering, Yonsei University, Seoul, 03722, Republic of Korea

[3]Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, 03063, Republic of Korea

[4]School of Electronic Engineering, Soongsil University, Seoul, 06978, Republic of Korea

*Corresponding Author: Kwangyon Lee. Email: kylee@ssu.ac.kr

Received: 08 August 2024 Accepted: 07 October 2024 Published: 18 November 2024

ABSTRACT

The growing demand for energy-efficient solutions has led to increased interest in analyzing building facades, as buildings contribute significantly to energy consumption in urban environments. However, conventional image segmentation methods often struggle to capture fine details such as edges and contours, limiting their effectiveness in identifying areas prone to energy loss. To address this challenge, we propose a novel segmentation methodology that combines object-wise processing with a two-stage deep learning model, Cascade U-Net. Object-wise processing isolates components of the facade, such as walls and windows, for independent analysis, while Cascade U-Net incorporates contour information to enhance segmentation accuracy. The methodology involves four steps: object isolation, which crops and adjusts the image based on bounding boxes; contour extraction, which derives contours; image segmentation, which modifies and reuses contours as guide data in Cascade U-Net to segment areas; and segmentation synthesis, which integrates the results obtained for each object to produce the final segmentation map. Applied to a dataset of Korean building images, the proposed method significantly outperformed traditional models, demonstrating improved accuracy and the ability to preserve critical structural details. Furthermore, we applied this approach to classify window thermal loss in real-world scenarios using infrared images, showing its potential to identify windows vulnerable to energy loss. Notably, our Cascade U-Net, which builds upon the relatively lightweight U-Net architecture, also exhibited strong performance, reinforcing the practical value of this method. Our approach offers a practical solution for enhancing energy efficiency in buildings by providing more precise segmentation results.

KEYWORDS

Building facade image; image segmentation; edge detection

1 Introduction

As global efforts to improve energy efficiency intensify, interest in buildings, the primary consumers of energy in urban environments, has surged. In the urban areas of advanced economies,

Copyright © 2024 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

buildings account for 38% of total energy consumption and 76% of electricity usage, underscoring the significance of effective building energy management [1]. Thermal energy loss plays a critical role in building energy consumption, significantly affecting heating and cooling costs while contributing to overall carbon emissions. In 2022, buildings were responsible for 34% of global energy demand and 37% of energy and process-related carbon dioxide emissions [2]. This has led to a focus on research aimed at identifying areas of thermal loss in building envelopes. By identifying areas vulnerable to thermal loss in buildings, heat leakage can be reduced through solutions such as replacing materials or adding insulation in targeted areas.

However, directly measuring thermal loss in building envelopes not only requires experts but is also highly challenging and inefficient. Traditional methods such as *in-situ* thermal measurements, blower door tests, and manual infrared thermography are time-consuming and labor-intensive. These methods require specialized equipment and trained personnel, making them costly and impractical for large-scale assessments. Additionally, accessing certain parts of a building, especially in high-rise structures, can pose safety risks. To address this issue, recent studies have proposed a process for detecting thermal loss through the analysis of building facade images. This process is also referred to as building facades parsing [3]. Notably, Park et al. [4] segmented building facade images, applied infrared images to the results, and analyzed the temperature distribution to detect abnormal thermal losses in building envelopes.

A fundamental element of this study is the segmentation of building facade images, which effectively differentiates between the building area and the non-building background, while also pinpointing objects within the building area. This segmentation is critical, as it precisely isolates areas of interest that are susceptible to thermal loss within the building. Using a segmentation algorithm not only streamlines the process of identifying these key objects but also significantly improves the efficiency of detecting thermal loss.

In the aforementioned research, the DeepLab V3+ [5], a prominent image segmentation method based on convolutional neural network (CNN), was utilized. While this methodology demonstrates commendable performance, achieving more accurate segmentation necessitates the development of an improved algorithm that takes into account the unique characteristics of building facade images.

This study proposes a methodology aimed at achieving higher building image segmentation performance. The input consists of building facade images taken from the front, while the output is a segmentation map containing the background, building, and windows. We focus on windows as key objects of interest, as they contribute significantly to thermal loss. This method features two distinctive characteristics. The first is object-wise processing, which separately addresses the two primary components of a building facade image: the building (or exterior wall of the building) and the windows. We use two neural network models, rather than a single model, to focus on characteristics of each object type. The second characteristic is the Cascade U-Net, which is a variation on the existing U-Net [6] model. The Cascade U-Net comprises two consecutive U-Net structures and improves the final segmentation performance by using contours of the objects as guide data for the model.

The methodology unfolds over four stages. In the first stage, we isolate the building and windows in bounding boxes from the input images by object detection. In the second stage, edge detection is employed to extract the edge map within the bounding boxes and to identify the contour, which is a closed curve that encompasses the largest area. Subsequently, the edge map serves as input data while the contour serves as guide data for the Cascade U-Net, our proposed image segmentation network. Finally, the results for the building and windows are synthesized to generate the ultimate segmentation map.

To evaluate the effectiveness of the method, we conducted experiments on a dataset of 586 Korean building images. The results demonstrated that our methodology surpasses the segmentation performance of end-to-end methods, including DeepLab V3+. The results indicated that processing target objects individually proved to be more effective than attempting to segment areas across the entire image in a single step. Furthermore, incorporating contours as additional information within the Cascade U-Net demonstrated clear advantages: by using contours as direct guidance, it was evident that the network was able to produce segmentation outcomes more aligned with our objectives. We anticipate that improving the segmentation of building facade images through our methodology will significantly aid in thermal loss detection and, consequently, in building energy management.

In summary, the contributions of this research are summarized as follows:

- We propose an object-wise processing methodology that isolates building facades and windows using bounding boxes, enhancing segmentation focus and accuracy.
- We introduce the Cascade U-Net with contour guidance, which leverages contour data to improve segmentation performance through a two-stage U-Net network.
- We demonstrate superior performance compared to end-to-end methods on a dataset of Korean building images. Additionally, we provide detailed evaluations of the contributions of each component to validate our methodology and its practical applications.

This section precedes Section 2, which provides a review of existing image segmentation models. Section 3 delves into our methodology in greater detail, followed by a discussion of the experimental results in Section 4. The paper concludes with Section 5, where we discuss the study's implications and outline directions for future research.

2 Related Work

2.1 Image Segmentation Methods

Image segmentation is the process of dividing a given image I into N segments $S_1, S_2, \ldots, S_N$, where each segment represents different objects or regions within the image. The primary objective is to determine which segment S_i each pixel $p(x, y)$ in the image I belongs. This can be mathematically expressed as follows:

$$I = \cup_{i=1}^{N} S_i \tag{1}$$

where, $S_i \cap S_j = \varnothing$ for all $i \neq j$ (i.e., no two segments overlap), and each S_i is a set of contiguous pixels.

In a pixel-wise classification approach to segmentation, the goal is to learn a function f that assigns a label $l_{(x,y)}$ to each pixel:

$$l_{(x,y)} = f(p(x, y)) \tag{2}$$

Here, $l_{(x,y)}$ is the label of the pixel $p(x, y)$, and f is a function that takes a pixel as input and outputs the label of the segment it belongs to.

Typical image segmentation models utilize an end-to-end architecture that takes an image as input and outputs a segmentation map through a deep learning model, specifically using a CNN. While these models perform well, each has its own distinct advantages and limitations. U-Net introduced a novel approach by adding intermediate feature maps from its contracting path (encoder) to its expanding path (decoder) using skip connections. This model uses a patch-based recognition approach instead of the conventional sliding window method to improve speed and accuracy. U-Net has been utilized in studies such as automated crack segmentation, demonstrating strong performance [7]. U-Net has the

advantage of a simple architecture and is highly effective when working with limited data. However, it tends to lose fine boundary details, particularly in complex or noisy images. DeepLab V3+, which is an improved version of DeepLab V3 [8], minimizes the loss of original image information and enriches pixel representation through the implementation of depth-wise separable convolution and the Atrous Spatial Pyramid Pooling (ASPP) module in its decoder. DeepLab V3+ offers superior performance by capturing multi-scale context, but its increased complexity and higher computational demands make it less suitable for resource-constrained environments. High-Resolution Network (HR-Net) [9] features a unique architecture that avoids downscaling, thereby preserving information fidelity. It maintains high-resolution feature maps while concurrently applying lower-resolution maps, facilitating continuous inter-scale information exchange within subnetworks. This model is optimally utilized for addressing heat map-based problems, such as pose estimation. However, a downside of HR-Net is its increased memory and computational demand, as high-resolution features must be preserved throughout the network. This can pose challenges for scalability and efficiency.

Such models, while powerful, may not fully address the challenges posed by building facade images, which often contain repetitive patterns and subtle features. Performance limitations arise when models rely solely on raw images as input, as they may struggle to accurately segment fine details and boundaries. This observation motivates the exploration of methods that incorporate supplementary information to enhance segmentation accuracy. Advancements like Bounding Box UNet (BB-UNet) [10] and Holistically-nested Edge Detection UNet (HED-UNet) [11] build upon U-Net's foundational strengths, introducing additional information for performance. BB-UNet, based on U-Net's structure, uses bounding boxes as masks to enhance the segmentation of specific areas. It concatenates these masks to the skip connections for decoding. While this approach improves segmentation for pre-identified regions, its performance may be limited when objects lack clear bounding boxes or when precise localization of the region of interest is challenging. HED-UNet also uses U-Net, emphasizing the importance of contour information in segmentation results. It predicts contours and segmentation results together, applying contour prediction results to the segmentation outcome in the form of attention [12] to achieve the final area segmentation. However, like BB-UNet, the reliance on additional contour data increases the complexity of model training and may not be applicable to datasets where contour annotations are scarce or difficult to obtain.

Motivated by the recent insight that leveraging supplementary information can significantly improve segmentation outcomes, our research aims to harness contour information of objects within images for enhanced segmentation accuracy. In addition, by isolating objects within the images, we focus the segmentation process on each object individually, allowing for more concentrated and accurate segmentation results.

2.2 Edge Detection Methods

The methodology proposed in this study involves extracting edge maps and obtaining contours from images. For edge detection, studies have used algorithm-based methods like Canny edge detection [13], as well as deep learning-based models such as Holistically-nested Edge Detection (HED) [14], and Dense Extreme Inception Network for Edge Detection (DexiNed) [15].

Canny edge detection remains the most prevalent algorithm for contour detection due to its efficiency and independence from training. It operates through multiple stages, initially focusing on noise reduction—a crucial step as noise significantly affects contour detection. It uses a Gaussian filter to minimize image noise, followed by the application of Sobel kernels in both horizontal and vertical directions to compute the gradient magnitudes. By establishing a threshold, it discerns the significant

gradient variations among neighboring pixels, thereby isolating the precise edges. While effective at detecting clear object contours in low-noise images, its performance diminishes with increased background noise.

Xie et al. [14] sought to address comprehensive image learning and prediction challenges, including multi-level feature learning, through their contour detection model. The HED model they proposed learns hierarchical representations to mitigate the ambiguity often encountered in contour and object boundary detection. Demonstrating commendable performance on the Berkeley Segmentation Data Set 500 (BSDS 500) dataset [16]—a benchmark for contour detection—it stands out for its speed (0.4 s per image), surpassing many contemporary CNN-based algorithms in efficiency.

DexiNed is a state-of-the-art deep learning model for contour detection, notable for producing finer contours compared to predecessors like HED. Structurally, DexiNed comprises an encoder with six blocks, each linked through convolutional blocks, and includes an auxiliary network enabling information exchange among these blocks.

Due to the need for clean edge maps and precise contours in this research, DexiNed was chosen for its superior noise reduction and contour detection capabilities, making it a crucial part of our edge detection process.

3 Building Facade Image Segmentation Methodology

This section provides a detailed explanation of our algorithm, which is distinguished by two principal innovations. The first innovation is object-wise processing; instead of using the original images directly, the approach employs object-wise bounding boxes as inputs. Unlike end-to-end models, the facade and window are processed separately and combined together in the final step for a comprehensive segmentation output. The second key innovation is the use of a neural network structure called Cascade U-Net, which improves on the conventional U-Net model by integrating two U-Net architectures. This configuration utilizes contours extracted from the images as guiding data, enhancing the segmentation performance of the model. Fig. 1 presents an overview of the proposed methodology, illustrating the process from the initial building facade image to the final segmentation map.

The methodology is composed of four stages, with each step playing a specific role:

Object Isolation: In this initial stage, an object detection model is used to isolate the building's facades and windows using bounding boxes, followed by resizing to standardize the input data for further processing.

Edge and Contour Extraction: Subsequently, an edge map is generated, facilitating the identification of precise contours within the delineated object boxes, thus laying the groundwork for detailed segmentation.

Object-Wise Segmentation: Central to our methodology, the Cascade U-Net structure incorporates two sequentially connected U-Net networks. The initial network is tasked with refining object contours for accuracy, while the subsequent network leverages these refined contours to execute the segmentation.

Output Synthesis: The final stage integrates the segmented outputs for facades and windows processed through the respective Cascade U-Net, culminating in a unified segmentation map that accurately reflects the architectural elements of the building facade.

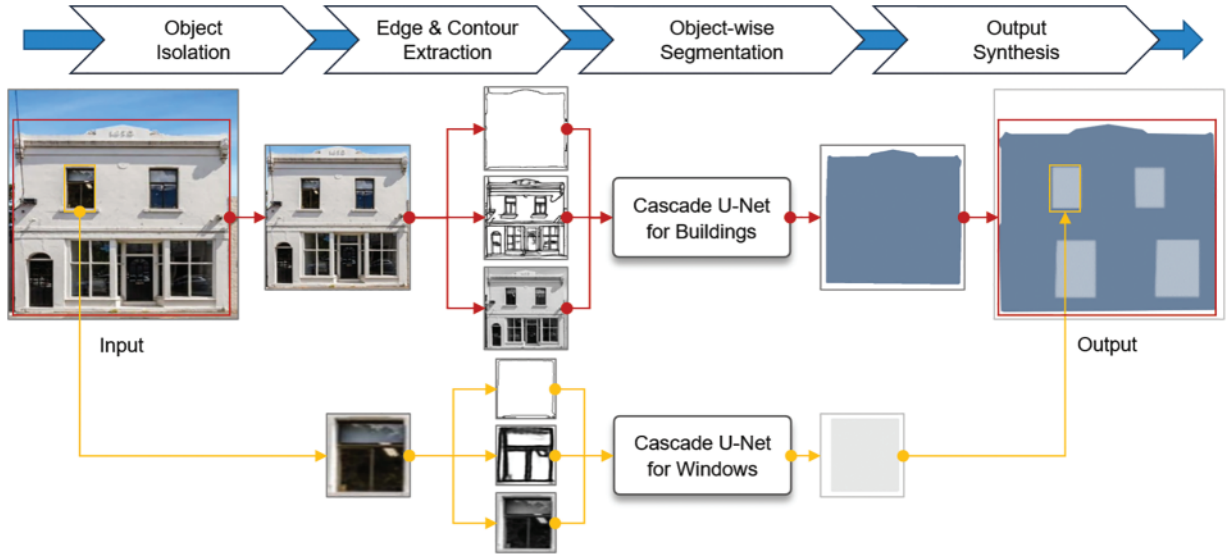

Figure 1: Overview of proposed method

3.1 Object Isolation

In the object isolation step, the facade and window objects are separated from the original image and ready to be processed in parallel for the next step. This step begins with object detection for the target objects. The bounding box obtained as a result of object detection represents the minimal, rectangular-shaped area enclosing the object, consisting of a position defined by x and y coordinates, as well as width and height. Here we assume that each input image contains only one building; therefore, because of object detection, we obtain one bounding box for the facade and multiple bounding boxes for the windows.

After identifying the bounding boxes for the objects, these areas are cropped from the original image, and their coordinates in the image are recorded. This record-keeping is crucial for the final step, where the final segmentation map is synthesized.

Given the variability in the sizes of the cropped object images, a standardization process is undertaken to resize them to a uniform input shape suitable for deep learning analysis. We chose the shape to approximate the average size of the cropped images, resizing facade images to 512×512 pixels and window images to 64×64 pixels. The rationale behind these specific resizing decisions is further elaborated upon in Section 4.

This strategy to use bounding boxes as inputs for the model, rather than the original images, was inspired by the previous works including BB-UNet [10], Kervadec et al. [17], and Lee et al. [18].

To maintain clarity and focus on our explanation, illustrations and examples provided henceforth will primarily address the processing of facade. It should be noted, however, that the same procedural framework is applied to windows as well.

3.2 Contour Extraction

In this step, the contours of objects are extracted. A contour is the out-most closed curve that represents the target object, which we consider as a key latent factor for segmentation. The contour

serves as guiding data in the subsequent Cascade U-Net stage to enhance the accuracy of building image segmentation. Fig. 2 displays the intermediate outputs of this process.

Figure 2: Contour extraction

We extract the edge map from the object image using DexiNed [15], the edge detection model that demonstrates the best performance. We utilized weights trained on the Berkeley Segmentation Dataset, specifically BSDS500 for this model.

We select DexiNed not only because it represents the state-of-the-art but also because it proved to be the most suitable for the building facade image dataset in experimental results. The Canny method [13] has the advantage of not requiring training, but it performed poorly due to its sensitivity to various noises, such as power lines or clouds around buildings. Deep learning-based contour detection models like HED [14] and HED-UNet [11] were also tested; though they were less sensitive, they often produced outputs with gaps instead of closed curves in experiments. Therefore, DexiNed was selected as the edge detection model for this study. Among the detected edges, we choose the closed curve with most extensive area as the object's contour using the OpenCV library [19]. This approach of contour extraction is based on the assumption that the closed curve with the largest area within the bounding box of the target object is most likely to represent the object.

The second and the third images in Fig. 2 illustrate the edge map and the resulting contour of the target object, respectively. As can be seen in the area marked with a red circle (Fig. 2c), the contour of the object is not accurate due to the background noises present in the building image. In the subsequent step, the front network of the Cascade U-Net will address this issue.

3.3 Object-Wise Segmentation

The Cascade U-Net builds on the traditional U-Net architecture but introduces two distinct differences. First, instead of using a single U-Net, it connects two networks in sequence to refine the process and generate intermediate results, aiming for improved performance. Another difference is the utilization of contours as guide data to steer the results in the desired direction for each network. Cascade U-Net is illustrated in Fig. 3.

The Cascade U-Net progresses through two sub-tasks. First sub-task is contour modification generates a refined contour from edge map, using the contour derived from the previous step. The next sub-task is image segmentation, which produces a segmentation map from a grayscale image, using the modified contour as guiding data.

The first network of the Cascade U-Net, called the contour refinement network, aims to refine the contours representing the object. The contour found in the previous step is used as guiding data to derive the modified contour from the edge map.

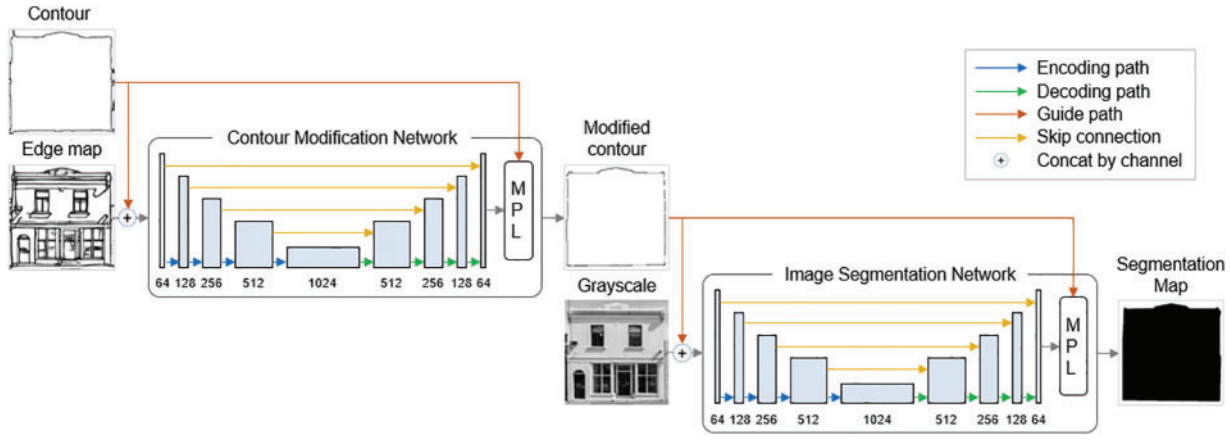

Figure 3: Cascade U-Net

This network fundamentally follows the U-Net model, taking an edge map as input and outputting a modified contour. The distinctive aspect proposed in this study is the use of the contour as guide data, adding it twice: once at the input part of the encoding path and again at the predict layer, as indicated by orange arrows in Fig. 3.

The second network, the image segmentation network, aims to produce an accurate segmentation map for the input object. Its overall structure including the application of guide data mirrors the Contour Refinement Network. However, the input for the network is changed to the grayscale image from the edge map. Additionally, the guide data here is the refined contours from the former network to ensure results are aligned with the segmentation objective.

In both networks, the guide data is concatenated with the primary input along the channel dimension at the encoder's input layer. This concatenation increases the number of input channels from one to two, allowing the network to process both the image intensity and contour information simultaneously.

In summary, the edge map and initial contour flow into the first U-Net, which outputs a refined contour. This refined contour then flows into the second U-Net along with the grayscale image to produce the segmentation map.

The process of incorporating guide data occurs twice, utilizing a layer named the Merging and Prediction Layer (MPL) before the prediction layer (Fig. 4). The traditional U-Net's prediction layer, indicated in blue (output of U-Net), carries a feature map of width $\times$ height $\times$ 64, which is processed through a 3×3 convolution layer and a 1×1 convolution layer (i.e., fully connected layer) to produce the final segmentation output. In this study, to effectively reflect the guide data, the feature map of width $\times$ height $\times$ 64 is reduced to two channels (width $\times$ height $\times$ 2) through a convolution layer, concatenated with guide data of width $\times$ height $\times$ 1, and then passed through two 3×3 convolution layers and a fully convolution layer in the MPL.

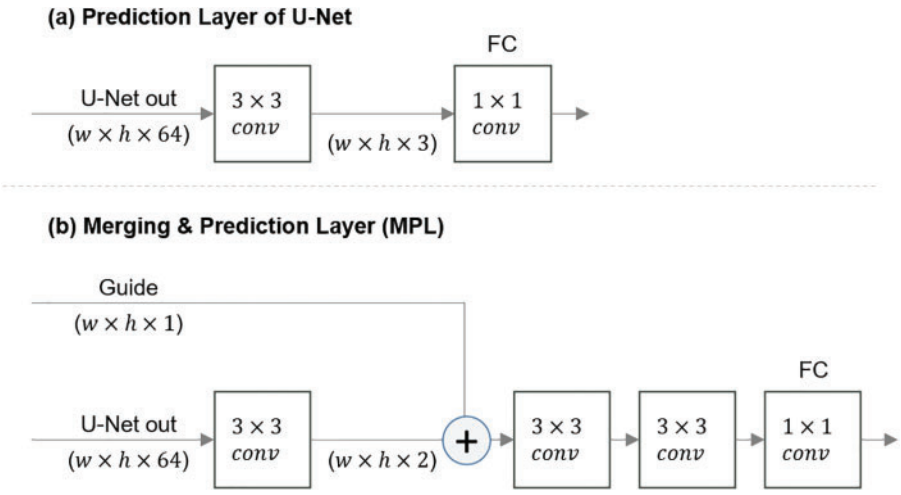

Figure 4: Merging and prediction layer

This MPL layer ensures that the guide data directly influences the final predictions, improving the network's ability to accurately segment object boundaries.

As the final step, output synthesis process is conducted to compile the final results. This stage reverses the process of object isolation. After passing through the Cascade U-Net, the results for each object, namely the segmentation map of the facade and the windows, are obtained. The segmented results for each object are restored to their original sizes, and then pasted back into their original positions, respectively. The sequence involves first attaching the exterior wall and then overlaying the windows onto them. The last segmentation map in Fig. 1 displays this integrated final output.

4 Experiment

4.1 Experimental Design

To demonstrate the effectiveness of the proposed methodology, we compare its performance against other methodologies using a dataset of Korean building images captured from the front.

4.1.1 Dataset

The CMP FACADE dataset [20] is a benchmark dataset frequently used for building image analysis. However, FACADE presented challenges for use in this study for three main reasons. Firstly, it is uncommon for the images within FACADE to capture the entire building, which is crucial for our study as it requires contour extraction and images showing the complete outline of the building facade. Secondly, the dataset predominantly features European-style buildings, which often include architectural elements not typically found in Korean buildings, such as column-like structures. Also, the dataset lacks features common in Korean buildings, such as external air conditioning units. Therefore, it was not suitable for out method which targets Korean building images. Lastly, inaccuracies in the segmentation ground truths within the FACADE dataset, particularly due to the ambiguous facades of older buildings could lead to incorrect identification of object types or areas on the segmentation maps. These factors made the dataset impractical for our purposes.

As a result, we collected our dataset, consisting of 586 images of Korean buildings, carefully selected to ensure each image included the complete outline of the buildings. Additionally, we included

a diverse range of building types commonly seen, such as commercial shops, villas, and office buildings, to ensure a comprehensive dataset. Fig. 5 showcases examples from the dataset we utilized.

Figure 5: Data examples

For the Cascade U-Net, we cropped images of facades or windows as inputs, using 586 cropped facade images and 1420 cropped window images. We employed the open-source tool labelIMG (https://github.com/tzutalin/labelImg (accessed on 08 August 2024)) for manually marking the locations of objects in a box shape on the training data. We applied a margin of 5 pixels around the bounding boxes during cropping to avoid accidentally cutting parts of the objects.

For data augmentation, we used only flip and rotate techniques. While blur [21], mixup [22], and cutmix [23] are popular in recent studies, they generate images in ways that could conflict with our study's requirement for detecting a single object within a box.

We split the dataset into training, validation, and test sets with an 8:1:1 ratio. For segmentation labeling, domain experts manually annotated the images to provide ground truth. For contours, we used the edges of the segmentation labels.

4.1.2 Evaluation Metric

To assess segmentation performance, we utilized widely recognized metrics: pixel-wise accuracy (PA) and intersection over union (IoU). Pixel-wise accuracy measures how correct each pixel is by comparing the predicted image with the ground truth, calculated as follows:

$$PA = \frac{\sum_i^k p_{ii}}{\sum_i^k \sum_j^k p_{ij}} \tag{3}$$

where, p_{ij} denotes the total count of pixels predicted as class j when they actually belong to class i.

IoU evaluates the proportion of overlap between the predicted and actual areas, calculated as follows:

$$IoU = \frac{|P \cap G|}{|P \cup G|} \tag{4}$$

Here, P represents the predicted area, and G is the ground truth area. Both metrics yield values between 0 and 1, with higher scores indicating either more accurately predicted pixels or greater overlap between the predicted and actual areas, thus denoting superior performance. Specifically, we differentiated our IoU analysis into mean IoU and wall IoU, where mean IoU averages the individual IoU scores for exterior walls and windows, and wall IoU exclusively focuses on the exterior wall's IoU score.

4.2 Results

4.2.1 Quantitative Analysis

In the quantitative analysis, we compared the segmentation map generated by each algorithm against the ground truth and evaluated performance. For baseline models, we selected end-to-end architectures: U-Net, DeepLab V3+, and HED-UNet. These models were given a grayscale version of the original images without object isolation (i.e., cropping into bounding boxes) as an input. For the training, U-Net and our proposed method were set to run for 100 epochs. In contrast, DeepLab V3+ and HED-UNet followed the training parameters recommended in their original publications.

The outcomes of these evaluations are summarized in Table 1. In the table, our method (the Cascade U-Net with object isolation) demonstrated the highest performance among the tested models. Compared to the U-Net, there was a significant improvement in performance. Specifically, it achieved a 2.90% increase over DeepLab V3+ and a 1.88% increase over HED-UNet based on the Wall IoU metric. These results clearly illustrate the superiority of the Cascade U-Net over traditional models.

Table 1: Performance comparison with baseline models

Method	Segmentation model	PA (%)	Wall IoU (%)	Mean IoU (%)
End-to-end segmentation	U-Net	63.4	65.0	65.6
	DeepLab V3+	68.7	66.1	70.7
	HED-UNet	70.8	67.1	70.8
Ours	Cascade U-Net	72.3	69.0	73.8

Here, DeepLab V3+ also outperformed the standard U-Net when applied in an end-to-end manner. However, for our proposed methodology, we selected U-Net to construct the Cascade U-Net architecture instead of cascading DeepLab V3+. This was because U-Net has a simpler and more straightforward architecture with fewer parameters compared to DeepLab V3+. When cascading networks for each object, computational complexity and resource demands increase significantly. Utilizing U-Net allows us to cascade two networks without rendering the model impractically large or slow.

To gain a deeper understanding of our methodology, we conducted experiments to evaluate the impact of each component of the methodology on performance. The results are summarized in Table 2, with descriptions for each setting as follows:

Table 2: Performance impact of methodology components (Overall) (OD: Object Detection, OI: Object Isolation, CE: Contour Extraction, OS: Output Synthesis, ISN: Image Segmentation Network)

Method	PA (%)	Wall IoU (%)	Mean IoU (%)
(1) OD	66.2	58.8	55.0
(2) OI, U-Net (grayscale)	69.2	63.6	67.6
(3) OI, CE, OS	66.3	53.8	60.0
(4) OI, CE, U-Net (edge map)	71.9	67.8	71.2
(5) OI, CE, ISN (contour, grayscale), OS	71.8	67.0	71.1
(6) OI, CE, Cascade U-Net without MPL, OS	72.0	67.7	71.4
(7) OI, CE, Cascade U-Net, OS	72.3	69.0	73.8

(1) Bounding boxes obtained through object detection in the Object Isolation stage were considered as the object areas themselves. (2) The part of the building's bounding box obtained through Object Isolation was converted to grayscale and fed into U-Net to obtain a segmentation map. (3) The internal areas of contours obtained for buildings and windows were considered as the areas of each object, and synthesizing these yielded a segmentation map. (4) An edge map of the building's bounding box obtained through Object Isolation was fed into U-Net to obtain a segmentation map. (5) Only the Image Segmentation Network part of Cascade U-Net was used to obtain and synthesize the segmentation map. (6) Cascade U-Net was applied, but without the additional application of MPL to guide contours. (7) Our methodology was fully applied.

Bounding boxes obtained through object detection showed poor performance (1), and contours alone were insufficient for segmentation (3). U-Net performs better when receiving an edge map rather than a grayscale image (2, 4), indicating the value of edge maps as additional information aiding segmentation. However, directly feeding contours extracted from the edge map into U-Net introduced noise, reducing performance (4, 5). Providing more accurate contours through contour refinement led to higher performance (6) and reflecting guidance in the last layer through MPL further improved performance (7).

The experimental results demonstrate that each component incrementally contributes to the overall performance improvement.

- Object Isolation (OI): By cropping images to focus on individual objects, we reduce background noise and allow the model to concentrate on relevant features, leading to an initial performance boost.
- Contour Extraction (CE): Extracting contours provides structural information about the objects, which is essential for accurate boundary delineation.
- Cascade U-Net Architecture: Introducing a two-stage U-Net allows for the refinement of contours before they are used as guide data in the segmentation process. This sequential processing ensures that the guidance provided to the segmentation network is accurate and reliable.
- Merging and Prediction Layer (MPL): The MPL enhances the model's ability to integrate guide data effectively. By merging the refined contours with the network's feature maps, it ensures that the final segmentation output benefits from precise boundary information.

By analyzing each component's contribution, we confirm that the combination of object-wise processing, contour refinement, and guided segmentation in the Cascade U-Net architecture leads to significant improvements.

4.2.2 Qualitative Analysis

For a qualitative comparison, we juxtaposed the segmentation ground truths with the results from DeepLab V3+ and our proposed method, as depicted in Fig. 6.

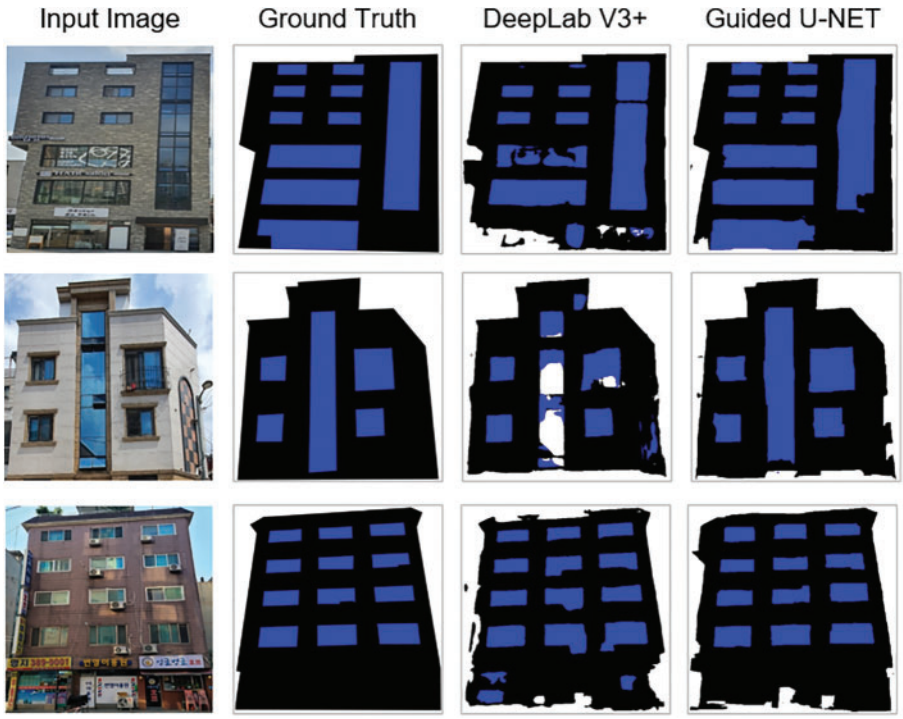

Figure 6: Qualitative analysis results comparing the ground truth with outcomes from DeepLab V3+ and our method

A notable observations from the qualitative analysis is the continuity within the predicted areas. Observing the second row of DeepLab V3+'s results in Fig. 6 reveals white spaces indicating the background surrounded by the building and windows. Since buildings typically lack openings in the middle, such discontinuities detract from the overall performance. In addition, when there are reflections on the windows, the performance gap between DeepLab V3+ and our methodology becomes more pronounced. Specifically, in cases where other buildings are reflected on the windows, DeepLab V3+ tends to misinterpret those areas as part of the building's exterior wall. Similarly, when the sky is reflected, it often mistakes those regions for the background outside the building. Additionally, using bounding boxes reduces unnecessary window detection, enhancing prediction precision. Thus, we conclude that our method achieves results that accurately reflect the common characteristics of building facade images.

4.2.3 Hyperparameter Optimization

Experiments were conducted on two key optimizable parameters: the input size of windows and the number of convolution layers within the MPL.

Due to the variability in bounding box sizes of objects, a standardized square input size is needed for the Cascade U-Net model. We experimented with window size w set to 32, 64, and 128 pixels, maintaining two layers in the MPL throughout. The experiments utilized these window sizes as inputs specifically for the window component of our proposed model. As shown in Fig. 7, the optimal performance was observed with $w = 64$, yielding the highest wall IoU across the model. Performance was significantly lower with $w = 32$, at 60.7%. At $w = 128$, there was a slight performance decrease to 68.6%.

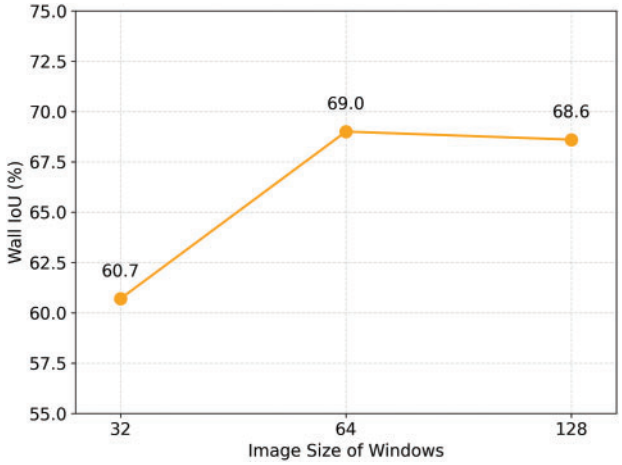

Figure 7: Performance relative to window image size

To ascertain why 64 × 64 offered the best performance, we checked the pixel count of window images cropped from the original images. The window images' average pixel count was 3339 pixels, closely aligning with the 64 × 64 size. This suggests that downsizing the window images to 32 × 32 compressed the information too much, leading to worse performance. This underscores the importance of tailoring the input bounding box size to accurately reflect the specific traits of the target dataset.

For the number of convolution layers, we explored the optimal number needed to merge the U-Net's standard output of $w \times h \times 64$ with the guide data of $w \times h \times 1$ for accurate segmentation. To avoid potential bias from directly concatenating the original U-Net's 64-channel output with the contour, we reduced it to 2 channels.

The experimental setup was designed to assess performance across one to four 3 × 3 convolution layers. The outcomes are depicted in Fig. 8. The scenario with a single layer, although closest to the original U-Net configuration, was inadequately trained with the addition of guide data, hence not shown in Fig. 8. The findings revealed that two convolution layers delivered the best performance, achieving a wall IoU of 69.01%. Introducing three or four layers led to performance reductions of approximately 2% and 7%, respectively. As the number of layers increased, performance declined, likely due to the unnecessary complexity causing information loss, consistent with insights from He et al. [24].

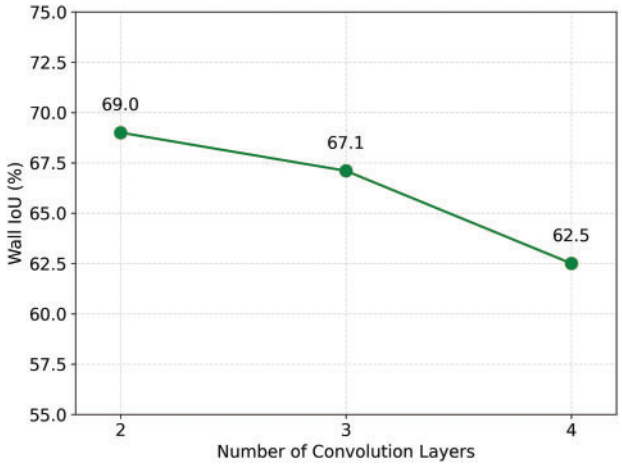

Figure 8: Performance with respect to number of convolution layers

4.2.4 Practical Scenario: Window Thermal Loss Classification

We explored the practical application of the proposed methodology by classifying windows based on their vulnerability to thermal loss. Fig. 9 illustrates an example of how we analyzed thermal loss in windows using the segmentation map generated from our research and infrared images.

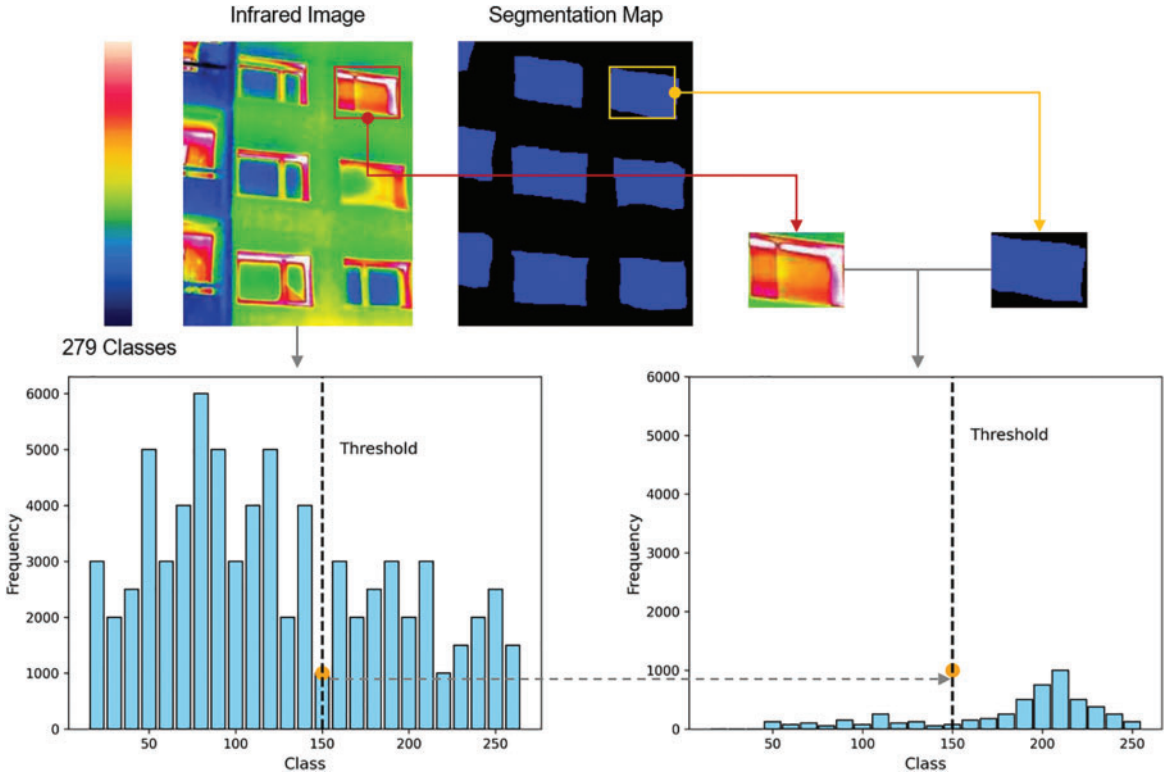

Figure 9: Window thermal loss classification based on infrared imaging and cascade U-Net segmentation

In this scenario, the infrared image was divided into 279 distinct classes representing relative temperature levels, and the pixel frequencies for each class were visualized in a histogram. Since there is no predefined threshold for thermal loss, we selected the midpoint of the observed bimodal distribution as the threshold. Using this threshold, we classified windows based on the proportion of pixels in the segmented window regions that exceeded the threshold. Specifically, if more than 20% of a window's pixels were classified above the threshold, the window was considered highly vulnerable to thermal loss. Windows with 5%–20% of pixels above the threshold were classified as moderately vulnerable, while those with fewer than 5% were classified as having minimal thermal loss. We applied this method to several buildings for which we had infrared images, confirming the potential of this approach for practical use. This demonstrates that our methodology can be used in real-world applications, such as identifying areas of concern in building energy management.

5 Discussion

As shown in Table 1, the proposed methodology demonstrated superior performance over other methodologies, including DeepLab V3+, on our constructed dataset of 586 images. Notably, the segmentation of building area exhibited high performance, with a mean IoU of over 73%.

The superior performance of our methodology over end-to-end models suggests that object-wise processing, which separates target objects into smaller bounding boxes rather than processing the original input image, is more effective. Furthermore, we explored the benefits of utilizing contours with the cascade U-Net. Contours, a key latent information embedded within the input image, act as a direct guide for the segmentation model, steering the segmentation process towards the desired outcomes.

Furthermore, our qualitative analysis revealed that the segmentation map produced by our model exhibited greater continuity compared to those from other models. Once again, this success can be attributed to the use of object-wise processing, generating cropped images by bounding boxes in the object isolation step.

It is important to note that the dataset used in this study is relatively small compared to standard benchmark datasets. Future efforts should focus on expanding the dataset for further testing. It is deemed valuable to not only increase the count of building facade images but also to examine the performance on non-frontal images, including side view images and partially obscured building images.

Despite achieving high performance, our methodology carries certain drawbacks. The use of two interconnected U-Net structures for the segmentation of walls and windows introduces increased complexity in the model architecture, a greater number of parameters to train, and extended training durations. While these may be minor issues when prioritizing performance, future studies should explore solutions to alleviate these challenges.

6 Conclusion

In this study, we proposed a methodology enhancing the segmentation of the building and the windows within building facade images designed to facilitate the analysis of thermal loss in buildings. Our method has two key characteristics.

Firstly, instead of directly extracting the final segmentation map from the input image, we utilized a method that separately segments the objects comprising the facade image before synthesizing them. This object-wise processing confines the input image to bounding boxes, allowing the image

segmentation model to focus on the target object's features. It also aids in maintaining the continuity of areas in the segmentation output.

Secondly, we proposed the Cascade U-Net, composed of two consecutive U-Nets. It treats contours as a key latent factor in image segmentation and uses them as guide data. This approach steers the model towards producing the desired outcomes.

Our experiments on the custom-built dataset of Korean building images demonstrated that these features contributed to achieving higher segmentation performance. If this approach enables better segmentation of building facade images, it is anticipated to greatly assist in building energy management when linked with thermal loss detection.

The main advantage of our methodology is its improved segmentation accuracy, outperforming other end-to-end methodologies including DeepLab V3+. By object-wise processing, each model was able to concentrate on specific objects, leading to further improvements in segmentation performance. Additionally, the incorporation of contours as guide data significantly enhanced the model's ability to accurately delineate object boundaries, a critical aspect of image segmentation.

However, our approach also presents some disadvantages. Although U-Net is considered a relatively lightweight model compared to more recent architectures, the use of multiple Cascade U-Net structures for different objects increases the model's complexity, resulting in a larger number of parameters to train and higher computational demands. This added complexity also results in longer training and inference times compared to simpler models, which may limit its use in scenarios requiring fast processing. Moreover, the success of our method relies on accurate contour extraction, meaning that it may require larger and more diverse datasets to generalize effectively across different building types and conditions.

For future research directions, we are considering the diversification of input images and the variety of objects to be detected. The input images in this study were building facade images that fully encompass the front of the buildings, but there's no guarantee that buildings will always be photographed in this manner. Moreover, since there are various objects constituting the building envelope beyond just windows, we plan to research methods capable of segmenting into multiple classes based on a wider variety of inputs. Finally, to ensure the practicality of this approach in real-world scenarios, we plan to evaluate the model's performance on edge devices. While we select U-Net due to its lightweight nature, we did not conduct specific experiments on such devices. This evaluation will be crucial in determining its efficiency and feasibility in real-time applications in resource-constrained environments.

Acknowledgement: The authors would like to express sincere gratitude to Smart Systems Lab for their invaluable support and guidance throughout this research.

Funding Statement: This work was supported by Korea Institute for Advancement of Technology (KIAT): P0017123, the Competency Development Program for Industry Specialist.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization: Haemin Jung, Heesung Park and Kwangyon Lee; Analysis and interpretation of results: Haemin Jung and Kwangyon Lee; Draft manuscript preparation: Haemin Jung, Heesung Park, Hae Sun Jung and Kwangyon Lee. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] U.S. Department of Energy, "Chapter 5—Increasing efficiency of building systems and technologies," Sep. 2015. Accessed: Aug. 8, 2024. [Online]. Available: https://www.energy.gov/articles/chapter-5-increasing-efficiency-buildings-systems-and-technologies

[2] United Nations Environment Programme, "Global status report for buildings and construction—Beyond foundations: Mainstreaming sustainable solutions to cut emissions from the buildings sector," in *Technical Reports*. Nairobi: United Nations Environment Programme, 2024. doi: 10.59117/20.500.11822/45095.

[3] H. Liu, J. Zhang, J. Zhu, and S. C. Hoi, "DeepFacade: A deep learning approach to facade parsing," presented at the 26th Int. Joint Conf. Artif. Intell. (IJCAI-17), Melbourne, Australia, Aug. 19–25, 2017, pp. 2301–2307.

[4] G. Park, M. Lee, H. Jang, and C. Kim, "Thermal anomaly detection in walls via CNN-based segmentation," *Autom. Constr.*, vol. 125, 2021, Art. no. 103627. doi: 10.1016/j.autcon.2021.103627.

[5] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," presented at the Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sep. 8–14, 2018, pp. 801–818.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," presented at the Med. Image Comput. Comput.-Assist. Interv. (MICCAI 2015), Munich, Germany, Oct. 5–9, 2015, vol. 9351, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[7] K. Chen, G. Reichard, X. Xu, and A. Akanmu, "Automated crack segmentation in close-range building façade inspection images using deep learning techniques," *J. Build. Eng.*, vol. 43, 2021, Art. no. 102913. doi: 10.1016/j.jobe.2021.102913.

[8] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[9] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.

[10] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah, "BB-UNet: U-Net with bounding box prior," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 6, pp. 1189–1198, 2020. doi: 10.1109/JSTSP.2020.3001502.

[11] K. Heidler, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: Combined segmentation and edge detection for monitoring the Antarctic coastline," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021. doi: 10.1109/TGRS.2021.3064606.

[12] A. Vaswani *et al.*, "Attention is all you need," presented at the 31st Conf. Neural Inform. Process. Syst. (NeurIPS 2017), Long Beach, CA, USA, Dec. 4–9, 2017, vol. 30.

[13] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, 1986. doi: 10.1109/TPAMI.1986.4767851.

[14] S. Xie and Z. Tu, "Holistically-nested edge detection," presented at the IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 11–18, 2015, pp. 1395–1403. doi: 10.1109/ICCV.2015.164.

[15] X. S. Poma, A. Sappa, P. Humanante, and A. Arbarinia, "Dense extreme inception network for edge detection," 2021, *arXiv:2112.02250*.

[16] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2010. doi: 10.1109/TPAMI.2010.161.

[17] H. Kervadec, J. Dolz, S. Wang, E. Granger, and I. B. Ayed, "Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision," presented at the Med. Imaging Deep Learn. (MIDL), Montreal, QC, Canada, Jul. 6–8, 2020, pp. 365–381.

[18] J. Lee, J. Yi, C. Shin, and S. Yoon, "BBAM: Bounding box attribution map for weakly supervised semantic and instance segmentation," presented at the IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, Jun. 19–25, 2021, pp. 2643–2652.
[19] G. Bradski, "The OpenCV library," *Dr. Dobb's J. Softw. Tools*, vol. 25, no. 11, pp. 120–123, 2000.
[20] R. Tyleček and R. Šára, "Spatial pattern templates for recognition of objects with regular structure," presented at the 35th German Conf. Pattern Recognit. (GCPR), Saarbrücken, Germany, Sep. 3–6, 2013, pp. 364–374.
[21] R. Liu, Z. Li, and J. Jia, "Image partial blur detection and classification," presented at the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Anchorage, AK, USA, Jun. 23–28, 2008, pp. 1–8. doi: 10.1109/CVPR.2008.4587465.
[22] H. Inoue, "Data augmentation by pairing samples for images classification," 2018, *arXiv:1801.02929*.
[23] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," presented at the IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, Republic of Korea, Oct. 27–Nov. 2, 2019, pp. 6023–6032.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," presented at the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 26–30, 2016, pp. 770–778.