Unveiling Building Façade Deterioration: A Drone-Powered Deep Learning Approach for Seamless Tile Peeling Detection

Ngoc-Mai Nguyen[1], Minh-Tu Cao[2] and Wei-Chih Wang[2]

[1]Department of Civil Engineering and Environmental Informatics, Minghsin University of Science and Technology, Taiwan
[2]Department of Civil Engineering, National Yang Ming Chiao Tong University, Taiwan
nnmai@must.edu.tw, mtcao@nycu.edu.tw, weichih@nycu.edu.tw

Abstract

The accurate segmentation of tile peeling on building facades holds considerable significance for effective building maintenance, particularly in regions like Taiwan, where tiles are the predominant facade protection. This research introduces YOLOM, a novel deep-learning-based segmentation model designed to address this challenge. YOLOM harnesses the capabilities of You Only Look Once version 7 (YOLOv7) and incorporates the BlendMask-based segmentation technique, further augmented by the Efficient Layer Aggregation Network (ELAN) to enhance feature discrimination and extraction capabilities specifically tailored for scenarios involving tile peeling. Employing a dataset comprising 400 images featuring 758 instances of peeling and 525 instances of sealed tiles observed during on-site surveys of public buildings, YOLOM exhibits outstanding segmentation performance. It outperforms the Resnet-BlendMask50 FPN with improvements of 7.1% of mean average percentage (mAP) and 0.4% of the average precision (AP) at the intersection over union (IoU) of 50%. Remarkably, YOLOM consistently surpasses other models, showcasing a 19.5% and 2.2% lead in AP for small and large objects, respectively. In a noteworthy advancement, YOLOM seamlessly integrates with drone technology, enhancing its capabilities for aerial surveying of building facades. This integrated approach proves invaluable for building maintenance teams, enabling proactive and cost-effective interventions. The study introduces a distinctive framework seamlessly integrating cutting-edge backbone and neck modules, particularly emphasizing the ALAN. The innovative YOLOM model establishes a new standard in artificial intelligence (AI) techniques for building maintenance, contributing significantly to academic discussions surrounding AI-enhanced image segmentation.

Keywords –

Tile peeling; Building façade; Building maintenance; Computer vision; YOLOv7; Deep learning; BlendMask technique; ALAN.

1 Introduction

While tile peeling may initially seem like a cosmetic concern in aging buildings, its ramifications are far-reaching. Beyond aesthetics, the detachment of tiles poses a direct threat to residents, risking falling accidents and compromising structural stability. The erosion of safety extends to the building's core, disrupting waterproofing and insulation capabilities. Exposed areas become susceptible to rainwater and humidity, accelerating material degradation and jeopardizing the adhesion between tiles and the structure. Recognizing tile peeling as more than a visual issue is crucial; it's a fundamental step in safeguarding both the safety and longevity of the built environment.

Routine inspections of building facades are imperative to address these risks. However, traditional inspection methods, involving manual surveys, photographic documentation, and physical condition recording, are labor-intensive and expensive and pose safety hazards for surveyors—particularly when navigating precarious sections of buildings such as high-rise rooftops and sidewalls [1]. Given the many buildings necessitating inspection, there is an urgent need to refine traditional methods to mitigate prolonged risks to structures and residents.

In response to the challenges inherent in infrastructure inspection, researchers and industry professionals are collaborating to explore innovative solutions. They are turning to advanced image analysis techniques, driven by the synergy of computer vision and artificial intelligence (AI), as a promising avenue for automating the assessment of damaged components in buildings and infrastructure [2-5]. Various models, including Faster R-CNN, SSD, SSD_Lite, and different

iterations of You Only Look Once (YOLO), have proven successful in detecting and categorizing damage on surfaces such as concrete structures, metro tunnels, bridges [6, 7] and road surfaces [8].

Despite numerous studies on defect survey work, a literature gap exists regarding advanced deep learning (DL) methods for identifying damage to architectural components, such as tile peeling on building facades. Existing methodologies lack practical modifications, potentially compromising accuracy or processing time. Additionally, the intricate context of captured images and constraints in drone-to-building surface access pose challenges in feature extraction for recognizing areas with peeling. The synergy between potent segmentation models and drone-powered technologies holds immense potential for revolutionizing building facade maintenance practices.

Addressing these literature gaps in problem-solving and methodology, this study introduces YOLOM, a pioneering segmentation model meticulously crafted to address the challenge of segmenting tile peeling areas on building facades. YOLOM leverages the strengths of YOLO version 7 [9] and integrates the segmentation framework of the BlendMask technique [10, augmented by efficient layer aggregation network (ELAN) blocks [10]. These ELAN blocks enhance feature discrimination and counteract gradual convergence deterioration, bolstering the model's performance in identifying and delineating tile peeling instances. Operating within a one-stage framework for pixel-wise segmentation, YOLOM capitalizes on the BlendMask-based segmentation mechanism, offering a robust solution to overcome identified limitations in the literature. By synergistically combining YOLOv7 with BlendMask, our aim is to establish a resilient segmentation model that significantly enhances the effectiveness and comprehensiveness of tile peeling inspection on building facades.

2 Literature review

In contemporary scholarly discourse, a discernible focus exists on harnessing AI and computer vision methodologies to facilitate scrutinizing structural components within buildings. These advancements have yielded significant benefits by furnishing tools identifying nuanced features such as subtle cracks, deformations, and structural irregularities. These imperceptible nuances might elude the human eye or escape manual inspections, making technological interventions indispensable for transforming the efficacy, precision, and inclusivity of inspections of building structures [11].

Despite the predominant emphasis on structural elements, a noticeable lack of attention has been directed towards architectural components, specifically facades and exterior wall cladding. This oversight is significant considering that, similar to their structural counterparts, architectural elements are vulnerable to wear, damage, and degradation as time progresses. The consequences of their decline extend beyond aesthetic considerations, influencing the overall functionality of a building and contributing to heightened maintenance expenses [12]. It becomes imperative to customize AI and computer vision techniques for architectural inspections, presenting a more comprehensive strategy to ensure the optimal condition of every aspect of a building, encompassing both structural and architectural facets.

Within the domain of computer vision applications for building inspections, especially in examining architectural components, enduring challenges persist despite recent advancements. A notable example is illustrated in the study undertaken by Kung, Pan [13], where a VGG-16 classifier [14] successfully classified damage on exterior wall tiles, attaining commendable accuracy. However, practical apprehensions regarding the viability of such a system emerge, particularly concerning capturing images at elevated heights and acquiring detailed images encompassing entire wall spans.

Expanding upon the initial research efforts, Guo, Wang [15] delved into applying a semi-supervised convolutional neural network (CNN) to classify façade damage, particularly under constraints of limited training datasets. Subsequent advancements were realized by Guo, Wang [11], who employed the Mask Region-based Convolutional Neural Network (Mask R-CNN) model to delineate plastered and painted façades. This application exhibited promising segmentation accuracy, with a mean average precision (mAP) of 58.4%. In a more comprehensive inquiry, Lee, Hong [16] scrutinized the efficacy of a Faster R-CNN architecture in the segmentation and categorization of defects on building facades. Notably, an average precision (AP) of 62.7% was achieved across all trained defects, employing an intersection over union (IoU) threshold of 0.5. Despite the laudable predictive performance, it is pertinent to acknowledge a significant limitation inherent in the Mask R-CNN model—its protracted inference time.

In recent research endeavors, Junior, Ferreira [17] made notable contributions by introducing the U-net, coupled with diverse Residual networks as the backbone architectures, to track crack lines in ceramic tiles. Extending the application of computer vision to address issues related to building facades, scholars have employed Faster R-CNN and Mask R-CNN models to identify and segment scratches on building glass panels [18]. The experimental results presented by Dais, Bal [19] compellingly support the effectiveness of DL on the crack segmentation on masonry surfaces. These findings, in conjunction with the previously mentioned studies,

underscore the growing potential of computer vision and DL in advancing methodologies for assessing facade defects.

The knowledge extracted from existing literature emphasizes that DL methods for detecting or segmenting defects on building facades largely align with conventional approaches. This observation underscores a notable gap, indicating an urgent necessity to innovate and enhance advanced DL models tailored to the intricate challenges of detecting building facade defects, including issues like tile peeling. Models prioritizing speed, robustness, and user-friendliness are essential to address practical concerns faced by building maintenance agencies. Beyond the immediacy of pragmatic considerations, the prospective trajectory and widespread integration of sophisticated DL-based computer vision models hold transformative potential for the building engineering sector. This paradigm shift can endow professionals with heightened levels of precision, efficiency, and comprehensiveness in the realm of facade inspections and the formulation of intervention strategies.

3 BlendMask-based YOLOv7 model

3.1 BlendMask-based image segmentation procedure

Introduced by Chen, Sun [20], BlendMask is a one-stage instance segmentation model within the Fully Convolutional One-Stage Object Detection (FCOS) framework. Its departure from the pre-defined anchor boxes employed in YOLO family models or region proposals like Mask R-CNN sets it apart, contributing to BlendMask's notably swift inference time. The architecture of the BlendMask model is composed of a feature extraction network and a mask prediction branch. The feature extraction network integrates a fusion of a residual network (Resnet) and a feature pyramid network (FPN). Concurrently, the mask branch incorporates three crucial components: 1) a bottom module determining the relative position of object instances, denoted as score maps; 2) a top layer generating specific attention maps for a detected region, concentrating on relationships between pixel pairs within the same instance by learning an embedding space, and 3) a blender module aligning the score maps with attention tensors (refer to Figure 1).

The bottom module, known as a "score map," predicts the location of a target object. Consequently, the output of the bottom module comprises bases (B) with a shape of $b \times n \times {}^{H}/_{S} \times {}^{W}/_{S}$, where b represents the batch size, n is the number of bases, and s is the output stride. The feature pyramid network output, including *P3*, *P4*, and *P5*, serves as the input for the bottom module. *P4* and *P5* undergo interpolation using the DeepLabV3+ decoder to match the size of *P3*, followed by concatenation through stacking. The loss function in this phase, termed semantic segmentation loss, is computed using the cross-entropy function.

Obtaining the feature pyramid network output (*P3* ~ *P7*) involves applying a convolutional layer to the tower's output. The tower is then expanded with a solitary convolutional layer, responsible for producing the bounding box size, center coordinates of the bounding box (center-nest), and determining the object class confined within the bounding box. Additionally, attention A is provided as the bounding box score, where the shape of this attention is $n \times M \times M$, with $M \times M$ denoting the resolution set at *14×14* in this study, and *n=4*. To finalize the bounding boxes for subsequent steps, the post-processing technique of FCOS [21] is applied to refine the bounding boxes $P = \{p_d \in \mathbb{R}_+^0 | d = 1, \dots, D|\}$ with the highest scores $A = \{a_d \in \mathbb{R}_+^{K \times M \times M} | d = 1, \dots, D|\}$. Two components contribute to the loss function in this phase, namely focal loss (L_{cls}) and IoU regression loss (L_{reg}).

$$
\begin{aligned}
&L(\{p_{x,y}\}, \{t_{x,y}\}) \\
&= \frac{1}{N_{pos}} \sum_{x,y} L_{cls}(p_{x,y}, c^*_{x,y}) \\
&+ \frac{\lambda}{N_{pos}} \sum_{x,y} \mathbb{I}_{\{c^*_{x,y} > 0\}} L_{reg}(t_{x,y}, t^*_{x,y})
\end{aligned}
\tag{1}
$$

where N_{pos} is positive samples, λ is the weight of regression loss (L_{reg}) term, $\mathbb{I} = 1 \; if \; c^*_{x,y} > 0$, and $\mathbb{I} = 0 \; if \; c^*_{x,y} \leq 0$

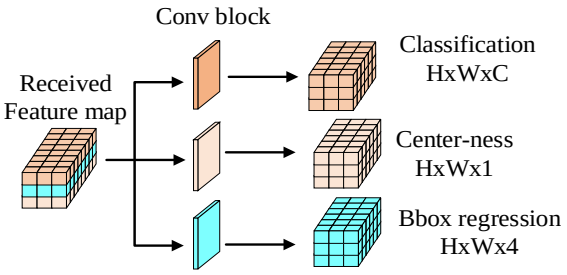

Figure 1. Content of tower block

The pivotal element within BlendMask is the blender module, which plays a crucial role in assimilating information from the bottom-level bases (B) and region proposals accompanied by corresponding top attention (A). During the training process, ground truth boxes serve as the region proposals, while in the prediction process, the bounding box is deduced. Given the varying sizes of the proposals, the Blender module employs the RoIPooler function to extract the area of the K bases associated with each proposal (p_d). Subsequently, this area is resized to a fixed size *(R×R)* with the feature shape denoted as r_d, as outlined in Equation (2). In executing this task, the RoIAlign technique was adopted,

implementing bilinear poolers.

$$r_d = RoIPool_{R\times R}(B, p_d), \quad \forall d \in \{1, \dots, D\} \qquad (2)$$

The attention resolution, denoted as $M \times M$, undergoes interpolation to match the size of the proposals ($R \times R$), forming a shape set $R = \{r_d | d = 1, \dots, D\}$. Subsequently, a'_d is subjected to normalization using the SoftMax function across the K bases, yielding the score map set $S = \{s_d | d = 1, \dots, D\}$. The next step involves the element-wise product between each entity r_d of the region proposal set R and the corresponding s_d of the score set S. This operation is performed for each of the K bases, and the results are summed to determine the mask logit (m_d), as outlined in Equation (3). The parameter K is consistently set at a value of 4, while R assumes values of 28 and 56, as proposed by Chen, Sun [20].

$$a'_d = interpolate_{M\times M \to R\times R}(a_d), \quad \forall d \in \{1, \dots, D\} \qquad (3)$$

$$s_d = softmax(a'_d), \quad \forall d \in \{1, \dots, D\} \qquad (4)$$

$$m_d = \sum_{k=1}^{K} s_d^k \circ r_d^k, \quad \forall d \in \{1, \dots, D\} \qquad (5)$$

3.2 ELAN-backbone and CSP-SPP + ELAN-PAN integration

Effective extraction and processing of features hold a central role in the analysis of image data. The advent of big data and advancements in convolutional neural networks (CNNs) and high-performance computers have facilitated the practicality of analyzing extensive image datasets. Using random trials is deemed impractical for developing efficient CNN networks tailored to extract specific task-related features. Therefore, a meticulous analysis of the particular task, incorporating intricate adjustments, becomes imperative.

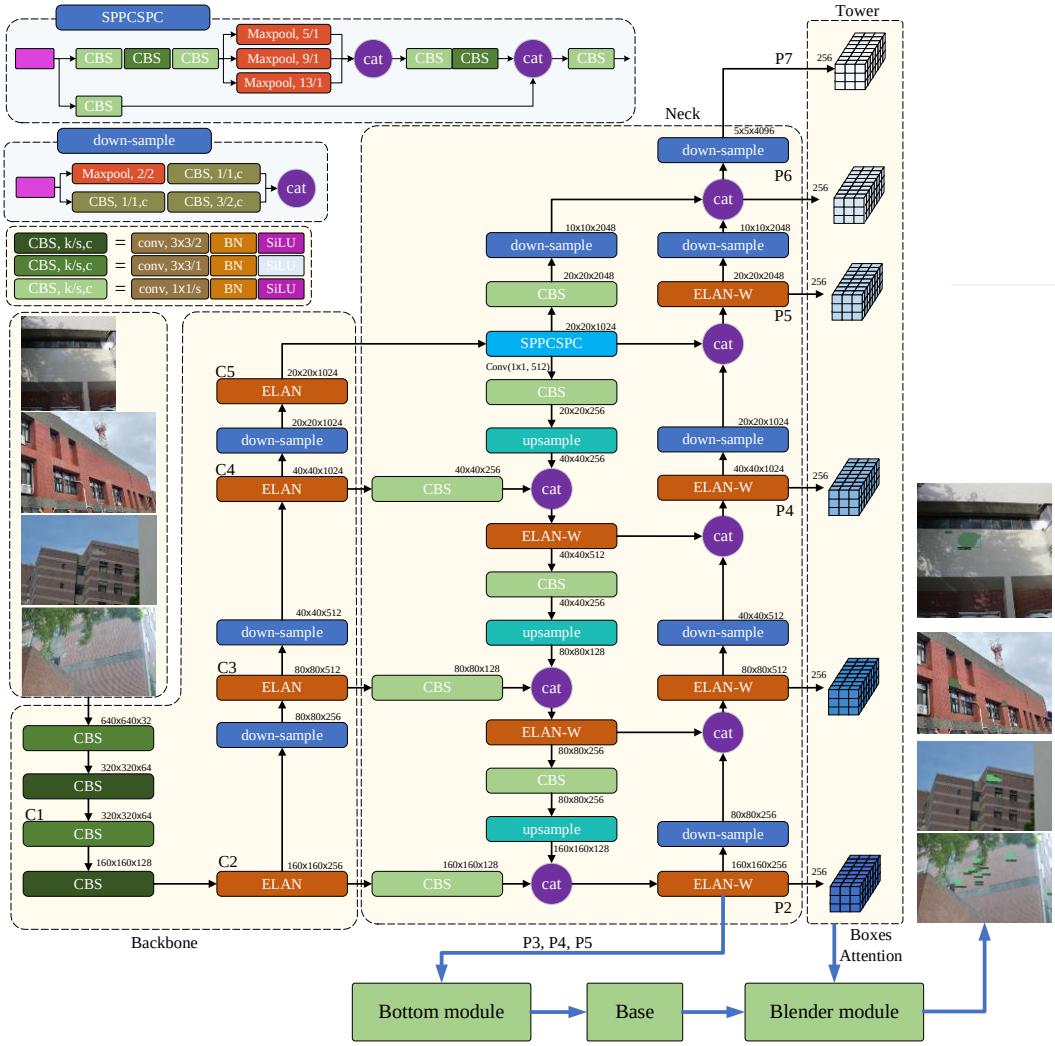

Figure 2. BlendMask with ELAN-base backbone and CSP-SPP+ELAN-PAN

After the YOLOv7 operations are completed, applying the SiLU activation function follows all batch normalizations. In a departure from conventional methods, this study introduces a unique network based on the ELAN [10]. ELAN, designed to counteract gradient deterioration in larger networks, combines elements from VoVNet [22] and CSPNet [23] to optimize the gradient length within a computational block. Notably, the layer aggregation network is trainable independently of the backbone network, facilitating faster training and experimentation, rendering it highly efficient for real-time operations.

Concerning the construction of the backbone, the ELAN is strategically incorporated between the downsampling blocks, as depicted in Figure 2. Diverging from YOLOv7, a deviation occurs by splitting two adjacent Cross-Stage Partial (CSP) blocks, and the output of all six CSP blocks in ELAN is directly concatenated, forming a structure referred to as ELAN-6. This ELAN-6 output, spanning C2 to C5, serves as the input at the forefront of the network. Moreover, ELAN-6 is employed at the head of the network to enhance feature extraction in CSP-SPP + ELAN-PAN, with PAN representing the Path Aggregation Network.

In this investigation, the Feature Pyramid Network (FPN) is substituted with the Pyramid Attention Network based on the Efficient Layer Aggregation Network (ELAN-PAN). ELAN is integrated into the layer scaling of PAN, processing backbone features before entering ELAN-PAN. In contrast to BlendMask, ELAN-PAN's scaling progresses from P2 to P7, as opposed to P7 to P2. ELAN is introduced into the transformation between the layers of ELAN-PAN. Due to a significant increase in the number of parameters in the P6 and P7 generations without substantially enhancing the segmentation model's inference power, ELAN avoids the downscaling task for P6 and P7. To align with the foundational structure of the BlendMask operating system, the channel number for each layer in ELAN-PAN is standardized. As a result, this study designates the proposed model as YOLOM.

4 Data collection and processing

Emphasizing the concern for public safety posed by peeling tiles, the surveyed buildings were strategically selected in high pedestrian-traffic zones. These encompassed various structures such as university campus buildings, apartment complexes, hospitals, government offices, and activity centers. Image data collection employed a Nikon D3200 digital camera, Autel Robotic EVO Lite+ unmanned aerial vehicle (UAV), Canon EOS M10, and iPhone 12 Pro, capturing photos across different seasons, times of day, and lighting conditions (*e.g.*, cloudy days, shadows, high and low light, and reflected light). Over a year, the survey team conducted fieldwork, ensuring diverse images with complex backgrounds to enhance the model's applicability in real-world scenarios. Each object was documented from various angles and within randomly sized rooms, contributing to the model's adaptability. The survey team utilized maximum zoom settings, especially for images capturing tile peeling at elevated heights. Following model training, an Autel Robotic EVO Lite+ unmanned aerial vehicle (UAV) was deployed to survey tile peeling in high-rise buildings.

A dataset comprising 400 surveyed images was employed in developing the tile peeling segmentation model, encompassing 758 instances of peeling and 529 instances of sealed tiles (refer to Figure 3 for representative samples). Upholding the quality of the dataset was a meticulous process involving labeling and verification by two additional members of the research team to annotate object instances. This rigorous approach ensured that the DL models were trained on a dataset of superior quality. Table 1 details the number of images and instances at each survey location for a comprehensive overview of the dataset. The primary survey locations across northern Taiwan are visually represented in Figure 4, accompanied by image samples collected from these regions.

Figure 3. Surveying locations in the northern Taiwan

Figure 4. Surveying locations in northern Taiwan

Table 1. Experimental dataset

Location	Images	Phase	instances	
			Peeling	Sealed
8	400	Training	638	410
		Test	120	119

5 Experimental results and discussions

In object segmentation, the assessment of model performance extends to metrics such as AP across varying IoU thresholds and mAP considering different object sizes. These metrics are pivotal for gaining a nuanced comprehension of the capabilities inherent in an object segmentation model. IoU is a critical gauge, elucidating the accuracy with which the predicted bounding box aligns with the actual object. The exploration of diverse thresholds, including AP50 (IoU > 0.5) and AP75 (IoU > 0.75), for small (APs), medium (APm), and large (APl) objects enables a nuanced evaluation of the model's precision across distinct levels.

The models in this comparative analysis underwent training on a sophisticated computing setup featuring an NVIDIA GeForce RTX 4090 24G GDDR6, RAM DDR5 5600MHz (2x32G), SSD Samsung 970 EVO Plus NVMe M.2, and an Intel CPU i7-13700-Core Processor. Hardware selection plays a substantial role in influencing training outcomes, particularly in DL models where GPU capabilities are paramount. A standardized image input size of 640x640 was maintained, and batch sizes of six were employed during the training process. Each model underwent an extensive training regimen encompassing 10,000 iterations, with meticulous fine-tuning to ensure optimal segmentation precision. These values were empirically chosen to yield the most favorable outcomes in the experiments, enhancing the model's proficiency in accurately segmenting tile peeling.

The experimental results of comparative models are presented in Table 2. Regarding segmentation, YOLOM stands out as the dominant model because it achieves the greatest values of all evaluation metrics.

The YOLOM obtained at least a 7.1% improvement compared with the second-best model, BlendMask–Resnet50 FPN, in terms of increasing the strictest AP value. The segmentation difference between the proposed model and other models incrementally increases as the complex challenge increases from large to small objects. As seen in Table 2, the YOLOM attained 2.2%, 11,9%, and 19.5% improvements in segmenting small, medium, and large objects compared with the remaining models, respectively.

YOLOM dominates the YOLOv7 mask that was published along with the YOLOv7 detection model by Wang et al. [25], boosting the AP50 and mAP values by 4.6% and 14.0% improvements. YOLOMASK is established by integrating CSP-SPP + ELAN-PAN and substituting the Resnet backbone with the YOLOv7 backbone while preserving the FPN found in BlendMask. This model structure was not streamlined, which is proven by the large segmentation accuracy drop of 16.7% and 14.0% of AP50 and mAP values. However, there is still an appraisal for its performance in segmenting small objects compared with BlendMask–Resnet50 FPN because it is supported by ELAN blocks. This study also uses BlendMask–CSPDarknet FPN to compare with the proposed model by substituting the YOLOv7 backbone with CSP Darknet. This model does not work efficiently, as proven by yielding much lower AP values than YOLOM and BlendMask-Resnet50 FPN with different IoU challenges.

Table 2. Experimental results of comparative models

	Model	*AP50*	*AP75*	mAP	AP^s	AP^m	AP^l
Segmentation results	BlendMask–Resnet50 FPN	0.799	0.537	0.496	0.181	0.399	0.652
	BlendMask–CSPDarknet FPN	0.639	0.386	0.367	0.107	0.385	0.471
	YOLOv7 mask by Wang, Bochkovskiy [24]	0.757	*NA*	0.427	*NA*	*NA*	*NA*
	YOLOMASK	0.636	0.348	0.352	0.125	0.307	0.473
	YOLOM	**0.803**	**0.699**	**0.567**	**0.376**	**0.518**	**0.674**
Bounding box results	BlendMask–Resnet50 FPN	**0.811**	0.611	0.543	0.334	0.475	0.662
	BlendMask–CSPDarknet FPN	0.650	0.433	0.402	0.187	0.431	0.475
	YOLOv7 mask by Wang, Bochkovskiy [24]	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
	YOLOMASK	0.650	0.420	0.387	0.162	0.366	0.476
	YOLOM	0.807	**0.736**	**0.622**	**0.540**	**0.583**	**0.702**

6 Conclusions

In the realm of building façade recognition, the early identification and precise delineation of issues emerge as pivotal factors. This study introduces a cutting-edge deep-learning-based segmentation tool named the YOLOM model, meticulously crafted to discern instances of tile peeling on building exteriors. Its application extends valuable support to building owners, aiding in proactive maintenance and resource conservation endeavors.

The YOLOM model has set a performance benchmark by incorporating state-of-the-art backbone and neck modules. A comparison with the BlendMask–Resnet50 FPN underscores its superiority, boasting a remarkable 7.1% increase in superior mean (mAP) and a notable 16.2% enhancement in AP75 values. Furthermore, it exhibits substantial leads in various precision metrics—APs, APm, and APl—with improvements of 19.5%, 11.9%, and 2.2%, respectively. These outcomes stem from a diligently curated dataset featuring 400 building façade images containing 1287 instances of peeling and sealed tiles. These validations affirm the model's robustness and propel academic and practical advancements.

Future endeavors will focus on integrating global building visuals and collecting tile-peeling images with diverse resolutions, enhancing the adaptability of YOLOM for improved training dynamics without compromising accuracy. Subsequent research will involve an ablation analysis to comprehensively assess the impact of various model components on the performance of the YOLOM, offering valuable insights for future enhancements.

Future studies could explore how variations in photographic datasets, including facade-distance, camera angle, lighting conditions, and image resolution, affect the performance of YOLOM, with a focus on strategies to mitigate challenges and optimize model performance across diverse real-world scenarios. Integration of augmented and synthetic data could augment dataset diversity and size, enhancing the model's generalization capabilities. Additionally, investigating transfer learning and domain adaptation techniques could improve YOLOM's adaptability to different datasets and mitigate domain shift issues. Evaluation of real-world deployment challenges should also be conducted to ensure successful implementation and adoption of YOLOM in building facade maintenance workflows, advancing its effectiveness and impact in practical applications.

Beyond its role as a mere model, this research contributes a robust framework to the academic community, accentuated by the ELAN-based structure. This foundation encourages researchers to explore advanced segmentation models. In conclusion, YOLOM is a guiding light in academic exploration, symbolizing AI's transformative potential in practical spheres, particularly building façade maintenance. As the journey forward unfolds with an array of improvements and expansions, collective efforts hold the potential to reshape façade maintenance narratives.

Acknowledgments

The authors would like to thank the National Science and Technology Council, Taiwan for financially supporting this research under contracts 111-2221-E-A49-189- and 112-2221-E-159-001-.

References

[1] Kim B, Yuvaraj N, Sri Preethaa KR, Arun Pandian R. Surface crack detection using deep learning with shallow CNN architecture for enhanced computation. *Neural Computing and Applications*, 33:9289-305, 2021.

[2] Dabetwar S, Padhye R, Kulkarni NN, Niezrecki C, Sabato A. Performance evaluation of deep learning algorithms for heat loss damage classification in buildings from UAV-borne infrared images. *Journal of Building Engineering*, 75:106948, 2023.

[3] Katsigiannis S, Seyedzadeh S, Agapiou A, Ramzan N. Deep learning for crack detection on masonry façades using limited data and transfer learning. *Journal of Building Engineering*, 76:107105, 2023.

[4] Xu Y, Qian W, Li N, Li H. Typical advances of artificial intelligence in civil engineering. *Advances in Structural Engineering*, 25:3405-24, 2022.

[5] Qiao W, Zhao Y, Xu Y, Lei Y, Wang Y, Yu S, et al. Deep learning-based pixel-level rock fragment recognition during tunnel excavation using instance segmentation model. *Tunnelling and Underground Space Technology*, 115:104072, 2021.

[6] Li D, Xie Q, Gong X, Yu Z, Xu J, Sun Y, et al. Automatic defect detection of metro tunnel surfaces using a vision-based inspection system. *Advanced Engineering Informatics*, 47:101206, 2021.

[7] Deng J, Lu Y, Lee VC-S. Concrete crack detection with handwriting script interferences using faster region-based convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, 35:373-88, 2020.

[8] Huyan J, Li W, Tighe S, Zhai J, Xu Z, Chen Y. Detection of sealed and unsealed cracks with complex backgrounds using deep convolutional neural network. *Automation in Construction*, 107:102946, 2019.

[9] Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:220702696*, 2022.

[10] Wang C-Y, Liao H-YM, Yeh I-H. Designing Network Design Strategies Through Gradient Path Analysis. *arXiv preprint arXiv:221104800*, 2022.
[11] Guo J, Wang Q, Li Y. Evaluation-oriented façade defects detection using rule-based deep learning method. *Automation in Construction*, 131:103910, 2021.
[12] Li J, Wang Q, Ma J, Guo J. Multi-defect segmentation from façade images using balanced copy–paste method. *Computer-Aided Civil and Infrastructure Engineering*, 37:1434-49, 2022.
[13] Kung R-Y, Pan N-H, Wang CCN, Lee P-C. Application of Deep Learning and Unmanned Aerial Vehicle on Building Maintenance. *Advances in Civil Engineering*, 2021:5598690, 2021.
[14] He K, Zang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*:770-8, 2016.
[15] Guo J, Wang Q, Li Y. Semi-supervised learning based on convolutional neural network and uncertainty filter for façade defects classification. *Computer-Aided Civil and Infrastructure Engineering*, 36:302-17, 2021.
[16] Lee K, Hong G, Sael L, Lee S, Kim HY. MultiDefectNet: Multi-Class Defect Detection of Building Façade Based on Deep Convolutional Neural Network. *Sustainability*, 12, 2020.
[17] Junior GS, Ferreira J, Millán-Arias C, Daniel R, Junior AC, Fernandes BJT. Ceramic Cracks Segmentation with Deep Learning. *Applied Sciences*, 11, 2021.
[18] Wang N, Zhao X, Zou Z, Zao P, Qi F. Autonomous damage segmentation and measurement of glazed tiles in historic buildings via deep learning. *Computer-Aided Civil and Infrastructure Engineering*, 35:277-91, 2020.
[19] Dais D, Bal İE, Smyrou E, Sarhosis V. Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning. *Automation in Construction*, 125:103606, 2021.
[20] Chen H, Sun K, Tian Z, Shen C, Huang Y, Yan Y. Blendmask: Top-down meets bottom-up for instance segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*:8573-81, 2020.
[21] Tian Z, Shen C, Chen H, He T. Fcos: Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF international conference on computer vision*:9627-36, 2019.
[22] Lee Y, Hwang J-w, Lee S, Bae Y, Park J. An energy and GPU-computation efficient backbone network for real-time object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*:0-, 2019.
[23] Wang C-Y, Liao H-YM, Wu Y-H, Chen P-Y, Hsieh J-W, Yeh I-H. CSPNet: A new backbone that can enhance learning capability of CNN. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*:390-1, 2020.
[24] Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*:7464-75, 2023.