A comparative study of attention mechanism and generative adversarial network in facade damage segmentation

Fangzheng Lin*1,3 | Jiesheng Yang1 | Jiangpeng Shu2 | Raimar J. Scherer1

1 Institute of Construction Informatics, Dresden University of Technology, Dresden, Germany
2 Collage of Civil Engineering and Architecture, Zhejiang University, Hangzhou, China
3 Deep Learning Center, Changzhou Microintelligence Co., Ltd., Changzhou, China

*Correspondence
Fangzheng Lin, Ph.D. Candidate, Institute of Construction Informatics, Technische Universität Dresden, Dresden 01069, Germany.
Email: fangzheng.lin@tu-dresden.de

Abstract
Semantic segmentation profits from deep learning and has shown its possibilities in handling the graphical data from the on-site inspection. As a result, visual damage in the facade images should be detected. Attention mechanism and generative adversarial networks are two of the most popular strategies to improve the quality of semantic segmentation. With specific focuses on these two strategies, this paper adopts U-net, a representative convolutional neural network, as the primary network and presents a comparative study in two steps. First, cell images are utilized to respectively determine the most effective networks among the U-nets with attention mechanism or generative adversarial networks. Subsequently, selected networks from the first test and their combination are applied for facade damage segmentation to investigate the performances of these networks. Besides, the combined effect of the attention mechanism and the generative adversarial network is discovered and discussed.

KEYWORDS:
U-net, attention mechanism, generative adversarial network, semantic segmentation, facade damage

1 INTRODUCTION
The deep learning based computer version (CV) concept implies the ultimate expectation of technology development, i.e., to completely replace or even exceed the human eyes' functionality. Nowadays, state-of-the-art is still far from the goal yet, but CV technology develops by handling tasks step by step. The current CV technology can bring advantages in processing complex images and massive graphical data. Generally, CV usage can be categorized into four categories: classification object detection, semantic segmentation, and instance segmentation. The detected object can be rendered in a specific color in the prediction results from semantic segmentation. This kind of labeling presents more precise information of objects’ sizes positions than object detection. Instance segmentation is a more advanced application to detect each object instance in a photo.

Due to the outstanding effects of CV, the Architecture, Engineering, and Construction (AEC) industry explores the possibilities of the related applications, especially in the field of Structural Health Monitoring. The visible damage in buildings and constructions is supposed to be detected, recognized, and classified by means of the deep learning based CV.

As the first step, on-site inspection is executed to collect a structure’s one-hand real-time data. Hence, visual inspection, a conventional, laborious but practical approach, is unavoidable. Nevertheless, the manual inspection approach still has to face several obstacles. The amount of manual data collection could be restricted. The inspectors’ life might be under threat, for instance, if they step into a building damaged in an earthquake or they stand under a sea-crossing bridge. Broader utilization of drones breaks through the limitation of humans and yields an efficient approach to damage data acquisition. A sufficient amount of image data can be acquired from on-site inspection using drones. This convenient and efficient approach improves the productivity of data collection productivity and changes the conventional processing methods of inspection results.

Abbreviations: CV, Computer Version; AEC, Architecture, Engineering and Construction; GAN, Generative Adversarial Network; CNN, Convolutional Neural Network; ANN, Artificial Neural Networks; FCN, Fully Convolution Network
This paper presents a comparative study on the U-net based semantic segmentation using attention mechanism and Generative Adversarial Networks (GANs). In section 2, literature is collected and reviewed to clarify the state-of-the-art Convolutional Neural Network (CNNs), semantic segmentation, and the corresponding application in the AEC industry. The newly proposed combinations of U-net and attention mechanism are elaborated in section 3. Afterward, GANs are employed to enhance the semantic segmentation, and the proposed discriminators are depicted in section 4. In section 5, a lightweight test is executed to select the CNNs with optimal performances, and the selected CNNs are applied for facade damage segmentation. Section 6 specifically discussed the combined effect of attention mechanism and GANs. Finally, section 7 concludes the entire research work, and suggests the next research steps in the futures.

2 | STATE OF THE ART

As a kind of modern computer science technology, Deep Learning is directly subject to the rapid development of computing capability. The corn of deep learning lies in the intensive training of Artificial Neural Networks (ANNs). Compared with other derived ANN, e.g., Recurrent Neural Networks and novel Graph Convolutional Networks, CNN are distinguished by their strength in image processing. Initial functions i.e. recognition and classification, cannot fulfill the requirement from the industry. Advanced approaches and the corresponding application are proposed, studied and implemented. This section retrospects the growth of CNN and particularly reviewed the related work in segmentation.

2.1 | The convolutional networks

Autonomous vehicles have become one of the most famous AI products in recent years. Its development is directly subject to one vital technology, computer vision, which experienced a development journal in the past years. The first work on modern CNNs occurred in the 1990s, inspired by the recognition. LeCun, Bottou, Bengio, & Haffner (1998) demonstrated that a CNN model which aggregates more specific features into progressively more complicated features could be successfully used for handwritten character recognition. MNIST (LeCun, Cortes, & Burges, 1998) is a now-famous dataset that includes images of handwritten digits paired with their true label of 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. A CNN model is trained on MNIST by giving it an example image, asking it to predict what digit is shown in the image, and then updating the model’s settings based on whether it predicted the digit identity correctly or not.

Around 2012 the CNN enjoyed a massive surge in popularity after Krizhevsky, Sutskever, & Hinton (2017) published AlexNet and achieved state-of-the-art performance labeling pictures in the ImageNet challenge. Similar to MNIST, ImageNet (Stanford Vision Lab, 2020) is a public, freely available dataset of images and corresponding true labels. However, instead of focusing on handwritten digits labeled 0-9. ImageNet focuses on “natural images” or pictures of the world, labeled with a variety of descriptors including “amphibian,” “furniture,” and “person.” Simonyan & Zisserman (2015) trained their proposed CNN VGG-16 for weeks and achieved top 5 accuracies in the ImageNet test. The massive data in the ImageNet were manually labeled and prepared. However, the laborious work guarantees the comprehensiveness and diversity of this dataset. Thus, it keeps encouraging innovative work of CNNs and also validates them. He, Zhang, Ren, & Sun (2015) conceived the residual learning and constructed the ResNet, which is much deeper than VGG but has a lower complexity.

All the CNNs mentioned above were initially developed as classifiers. Nevertheless, they have become the representative foundation of more advanced CV technology, e.g., object detection, semantic segmentation, and instance segmentation. The popular Fast R-CNN (Girshick, 2015) took the VGG-16 as the backbone for object detection. Ou et al. (2019) extended the ResNet-18 with an encoder-decoder structure. The developed method is capable of segmenting objects from dynamic scenes. Instance segmentation requires both object detection and semantic segmentation. Zhao, Shadabfar, Zhang, Chen, & Huang (2021) adopted ResNet-101 to classify leakage-areas of shield tunnel linings and utilized the Mask R-CNN (He, Gkioxari, Dollár, & Girshick, 2018) to label these different leakage-areas.

2.2 | Semantic segmentation

An image is nothing but a collection of pixels. Image segmentation is the process of classifying each pixel in an image belonging to a particular class and hence can be thought of as a classification problem per pixel. There are two kinds of image segmentation, namely semantic segmentation, and instance segmentation. The former is the process of classifying each pixel belonging to a particular label. It does not distinguish different instances of the same object. Instance segmentation differs from semantic segmentation in the sense that it gives a unique label to every instance of a particular object in the image.

Semantic segmentation is a classic topic, which has been studied for decades. Dating back to 1992, Beek, Reinders, Sankur, & Lubbe (1992) faced the head-and-shoulder scene in video telephony and developed a segmentation system to
locate the speaker, the face, and the eyes automatically and sequentially in a video. Conventional approaches and mathematical modes were adopted in that system, such as model-based coding and fuzzy membership functions. This paper, which reveals the state of semantic segmentation at the early stage in a certain aspect, is a representative but non-ephemeral one. Csurka & Perronnin (2011) inherited that thought and conceived an efficient but straightforward approach enhanced by the Fischer kernel. Based on open-source datasets, the results provided a comparable performance on semantic segmentation, for which CNNs are nowadays prevailing deducted.

2.3 Fully Convolution Networks and the U-net

Numerous pieces of research work, which present diverse convolutional networks, endorse that CNNs improve the quality of semantic segmentation significantly with respect to efficiency and accuracy. In 2014, Long, Shelhamer, & Darrell (2015) published the Fully Convolution Networks (FCNs). They interchanged the fully connected layers of a CNN classifier, e.g., LeNet (Lecun et al., 1998), into convolution layers. In experiments, FCN was tested and delivered satisfactory prediction results. FCN is a concept rather than a specific permanent convolutional net. It can be implemented with the existing CNNs that serve as the backbone. Mou, Hua, & Zhu (2019) adopted the FCN, took the VGG-16 (Simonyan & Zisserman, 2015) net as the backbone, and inserted the relation modules for semantic segmentation in aerial scenes. Tong, Xu, & DenÅux (2021) composed an FCN and a Dempster-Shafer layer so that the pixel-wise feature maps are extra extracted from an input image to promote the prediction quality. FCN might be one of the most popular CNNs since its appearance. A large number of papers contributed to enhancing FCNs in various ways. However, most of them do not think outside the box of improving FCNs by adding a component.

Beyond that, novel architectures are proposed to explore other approaches to semantic segmentation. U-net [Ronneberger, Fischer, & Brox, 2015] is a typical one; it is established still on the basis of the encoding-decoding principle. Nevertheless, utilizing the skip connection to leverage the output from down-sampling can be interpreted as an innovative strategy in the U-net in order to reduce the required training data amount. Since an original CNN is usually simply designed. Thus there is much potential for further development. With focuses on the functionality of skip connections, U-net++ (Zhou, Siddiquee, Tajbakhsh, & Liang, 2018) and U-net 3+ (H. Huang et al., 2020) share a similar idea of enhancement. In the U-net++, the encoder and decoder sub-networks are nests to replace the original skip connections. This network can be interpreted as a more dense and complicated version of the U-net. H. Huang et al. (2020) proposed the U-net 3+, which abandons the nested skip connections and adopts full-scale skip connections. By means of that, the output of an encoding layer is transmitted to each sub-decoding layer. In addition, the labels are used in the decoding layers to adjust the loss calculation. Experiments in both papers were executed with the state-of-the-art dataset, and the consequence is that these kinds of reinforcement positively influence semantic segmentation. Changing skip connection is one direction to obtaining a more powerful U-net. Another direction would be transformation or accumulation. W-net (Xia & Kulis, 2017) is not a pure concatenation of two U-nets; two specific loss functions maximize semantic segmentation. Mehta & Sivaswamy (2017) added a left leg and a right one to a U-net named M-net. An effective semantic segmentation materializes through concatenating the initial input and the output from each encoding layer and concatenating the output of the bottom layer and the output from each decoding layer.

2.4 Generative adversarial networks

Since the generative adversarial network was introduced to the world by Goodfellow et al. (2014), its related developments and applications present a tendency of exploration. This design opens a new window of CNN application. In the definition, the training of a GAN should be proceeded by two steps in each epoch (see Eq. 1). When an epoch E starts, the discriminator D(x) is trained firstly with the generator’s G(x) initial output and the labels z. As long as D(x) reaches the maximum maxD, the training of discriminator is completed. Subsequently, the trained discriminator feedbacks information to the generator, the generator’s training is then launched by feeding the input data x until the generator’s minimum minGD is obtained. This ingenious orchestration between a generator and a discriminator encourages researchers to mine its potential in semantic segmentation.

\[
\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))]
\]  

(1)

Various domains have benefited from the semantic segmentation using GANs. Decourt & Duong (2020) combined the ResNet-101 (He, Zhang, Ren, & Sun, 2016) as the generator and an FCN as the discriminator to build a GAN for MRI image segmentation. Object identification in medical images is indeed an issue that challenges researchers. Brain tumors in MRI pictures can be segmented through a pre-trained GAN (Ghassemi, Shoeibi, & Rouhani, 2020). Y. Zhang, Miao, Mansi, & Liao (2020) accomplished multi-organ segmentation for X-ray images. Apart from medical images, remote sensing pictures are often requested to be segmented. J. Huang.
Liu, Tang, & Zhang (2021) connected the U-net as the generator and an FCN as a discriminator to identify the ships in remote sensing images of a harbor. Xiong, He, Liu, & Liao (2020) proposed a more complicated GAN architecture to promote semantic segmentation. In this combination, the FCN segments the buildings, roads, and forests in the input remote sensing images, and the GAN is trained to optimize the first segmentation.

### 2.5 Application in the AEC industry

While computer scientists developed new CNN models or CV-based approaches for a generic and holistic application, the practitioners in the AEC industry adore optimizing these models or approaches to address specific issues. Dais, Bal, Smyrou, & Sarhosis (2021) proposed a comprehensive assessment system for future landscapes. In this system, an open-source toolbox was adopted to segment the landscape images. The predictions were visualized with live videos in a mixed reality system for landscape assessment. In a single building project, semantic segmentation can be applied for building component classification. Czerniawski & Leite (2020) considered the features of color and depth in training a deep convolutional network for this objective. The authors also adopted two trained CNNs to label the object in point clouds of building interiors (Ma, Czerniawski, & Leite, 2020). Objects in infrastructures can also be identified via semantic segmentation. Chen et al. (2020) optimized the deep convolutional network to detect the weak interlayers in tunnel faces and evaluate them quantitatively. Compared with the FCN, the proposed network shows a higher accuracy and less training.

Generally, the encoding-decoding principle denotes the essential thought behind semantic segmentation using CNNs. No matter FCNs, U-nets, or GANs, all of them obey this principle. Although copious distinct networks (Milletari, Navab, & Ahmadi, 2016; Shim & Cho, 2020; Zheng et al., 2021) are conceived to address semantic segmentation, the deja-vu about the three networks always exists in these publications. In terms of the AEC industry, visible or superficial damage identification has become a specific vital domain, such as detecting cracks, spall, or corrosion. Some papers emphasize the networks themselves; meanwhile, some propose a workflow that encompasses the function of semantic segmentation and the strategy of data acquisition. For example, drones promote the efficiency and safety of taking images of the damage in, e.g., bridge super-/substructures (Seo, Duque, & Wacker, 2018) or building roofs (Inoguchi, Tamura, & Hamamoto, 2019). The majority of related publications propose novel CNN architectures to compare it with other networks. Generally, only the calculated metrics are compared to prove the feasibility and advancement of the proposed network. However, few papers analyzed the relation between the metrics and the design of CNNs. This research work proposes novel CNNs and simultaneously investigates the effects from different combinations of networks through comparative study.

### 3 THE U-NET BASED ARCHITECTURES FOR SEMANTIC SEGMENTATION

In this section, the original U-net and the attention mechanism are initially manifested. The redistributions of attention mechanisms are studied. Inspired by the attention U-net (Schlemper et al., 2019), two novel CNNs are proposed to upgrade the usage of input data during training and test.

#### 3.1 U-net

The U-net (Ronneberger et al., 2015) is designed as an encoder-decoder structure with four layers of down- and upsampling. The default input image should have one channel. In the case of colorful inputs with three channels, the phase of data preparation can switch them into mono-colored images. The first layer of downsampling is composed of two convolution operations. The former convolution reshapes the input data, while outputs of the latter one remain the size of the inputs. Unlike the first layer in the encoder, a max-pooling stage is added to the other encoder layers. Along with the downsampling process, the defined input and output channels change. The bottom layer is treated as the extension of the encoder layer. Hence, it is built with the same structure and corresponding channel adaptation. Decoder layers are constructed strictly identically to the encoders apart from the pooling stage. Up pooling serves to concatenate the skip connection and the output of the downer layer and subsequently scale on the processed data. The softmax function is recommended by Ronneberger et al. (2015) to be applied in the last stage of the topside decoder layer. However, other functions having equivalent effects can be employed as well. The skip connection, the output from each encoder layer, is a feature, which distinguishes the U-net from other FCN based architectures. By using skip connections in the U-net, fewer images can be fed for training and achieve comparable semantic segmentation results.

#### 3.2 Attention mechanism

Psychology defines attention as the cognitive process of selectively focusing on a few objects while disregarding others. A neural network is regarded as a reduced effort to imitate human brain activities. Attention Mechanism is also an effort to accomplish the same action by concentrating selectively on
FIGURE 1 The U-net and the studied U-net based networks
certain essential items while disregarding other things in deep neural networks.

Bahdanau, Cho, & Bengio (2015) implemented the attention mechanism in an RNN for neural machine translation. It was the very first appearance of an attention mechanism in artificial neural networks. The RNNs, oriented to language translation, are usually developed in the encoder-decoder. Generally, updating translation quality entails a relatively complicated RNN, which means multiple layers and a more extended structure. Consequently, the vital features in the input data could be reduced through convolutional layers in the decoder. Therefore, the attention mechanism, which was applied initially to address the issue, is dedicated to improving the possibility of the vital feature extraction in the decoder. For example, Usama et al., 2020) combined an RNN and a CNN for the sentiment classification of short text. As the RNN works as the encoder, the attention within assists in focusing on essential text features.

The performance of the attention mechanism in RNNs inspires researchers to explore its potential in CNNs. Tian et al. (2020) inserted the attention mechanism in a CNN to denoise the images automatically. In that CNN, the attention block exploits the two steps to implement the attention mechanism. The first step uses convolution of size 1 × 1 to compress the obtained features into a vector as the weights for adjusting the previous stage. The second step utilizes obtained weights to multiply the previous layer’s output for extracting more prominent noise features. Hang, Li, Liu, Ghamisi, & Bhattacharyya (2021) employed attention mechanism in image detection. In the frame of that research, a dual-CNN group was proposed. A spectral attention module and a spatial one, integrated into the CNN group respectively, tried to extract the features from different dimensions. The former focuses the input along the direction of channels, and the latter evokes the feature vertically to the channels. The experiment verified the positive effects of two novel attention modules’ collaboration in image classification.

The attention gate (Schlemper et al., 2019) is constructed with more stages, as Fig. 2 illustrates. The previous attention implementations combined the data and that after attention processing once. On the contrary, the attention gate leverages skip connections, the unique feature of a U-net, named the attention U-net in Fig. 1. The output from the previous stage is firstly concatenated with a skip connection and subsequently serves as the input. After the attention processing, the input is combined with the outputs and exported out of the attention gate. The corresponding upgrade of prediction quality was validated in the experiment of medical image segmentation. The attention mechanism is merely an abstract concept; hence, it can be implemented in various forms based on different usages. A practical attention module has not to be that sophisticated. The vital issue lies in feature extraction. The more features can be filtered out from the inputs, the better prediction results can be obtained.

To the author’s best knowledge, Schlemper et al. (2019) contributed firstly to the combination of the U-net and attention mechanism. Attention gates inherit the conception of the attention mechanism and adjust it for the U-net through creating an interface to the skip connections. Attention gates are located before the concatenation operation in each decoding layer. The value g denotes the skip connection produced by the encoder layer on the same level, and the value x' represents the upsampled data from the lower decoder layer. Though these two sets of data, which actually are two tensors, present with different sizes. The output of the attention gate should have the same size of g before the concatenation with x'. The idea behind attention gets leis that the upsampled data is reused to upgrade the prediction of the possible image regions of interest, which is merely realized by the skip connection in the U-net.

3.3 The advanced attention U-net

Schlemper et al. (2019) focused mainly on the decoder in the attention U-net. However, the encoder was not touched. Since the attention U-net proves that the accumulation of the outputs from encoders is capable of facilitating the U-net’s performance in semantic segmentation, it is reasonable to expect even better results from a more comprehensive accumulation of encoders’ outputs. The advanced attention U-net (see Fig. 2) is constructed to leverage the data in skip connections. Zhou et al. (2018) noticed this issue as well and proposed the U-net++ with more en-/decoder layers and sophisticated skip connection network between the layers. H. Huang et al. (2020) proposed the U-net 3+, in which each decoder layer adopts the outputs from encoder layers on the same or lower levels, and related decoder layers also apply the output of the bottom layer. Aiming at a higher accuracy in semantic segmentation, the U-net 3+ combined several loss functions to present a correct gradient descent. Analog to these two networks, the advanced attention U-net proposed in this paper emphasizes
the inherent structure of the U-net but is built as a novel architecture to find a compromise solution. By means of that, the output of an encoder layer is convoluted and transmitted to all the lower layers. For example, the topside output is operated three times so that the singly or repetitively convoluted output of the topside layer stands in each layer beneath. All the outputs on one level are concatenated as a tensor, which moves to the attention gate as the parameter $g$ in Fig. 2. The advanced attention U-net is designed to adopt typical loss functions, and no novel loss function has to be assembled or invented.

### 3.4 The full attention U-net

While the advanced attention U-net obeys the order concatenation-to-attention, the full attention U-net is designed to explore a reverse way for the comparative study. Deriving from the advanced attention U-net, the full attention U-net switches the order of concatenation and attention. Except for the topside layer, encoded data on a level and one on the adjacent lower level are transported into an attention gate respectively as the $g$ and $x'$ in Fig. 1 d. Afterward, all the tensors produced by attention gates are concatenated as one tensor for upsampling in the decoder. The study of the order of concatenation and attention is barely to be seen in the related publications. Therefore, the comparative study in the paper is expected to investigate the effects caused by certain operations in two completely reverse orders.

### 4 ENHANCEMENT WITH GENERATIVE ADVERSARIAL NETWORKS

This section elaborates on the solution to enhance the U-net utilizing GANs. Therefore, in the frame of GANs, the aforementioned CNNs are to be adopted as the generator. And the discriminators are supposed to be studied. Thus, in order to investigate the segmentation capability of discriminators, four discriminators are proposed with the consideration of dual-dimensions of complexity, namely the layer number (length) of a discriminator and the stage number (thickness) of a single layer.

#### 4.1 The principle of GANs

GANs are initially proposed as a generic architecture. For distinct objectives, e.g., classification, self-adaption, or semantic segmentation, the roles of generators and discriminators vary accordingly. As known, a GAN is composed of a generator and a discriminator (see Eq. 1); hence the training process should encompass both. First, the generator learns to generate plausible data. The instances that are created serve as negative training examples for the discriminator. Second, the discriminator learns to differentiate between the generator’s actual and false data generated. The generator is penalized by the discriminator for providing improbable results. The generator learns to generate fake data by integrating discriminator feedback. It learns to manipulate the discriminator to identify its output as real. Generator training necessitates a closer integration of the generator and the discriminator than discriminator training necessitates. Random input, generator network, discriminator network, discriminator output, and generator loss are all items of the GAN that train the generator. In a GAN constructed for semantic segmentation, the generator can be interpreted as a network to produce the segmented images as the fake data; and the discriminator should evaluate the difference between the segmented images and the ground truths, i.e., the real data. The forward- and back-propagation are processed through both the generator and discriminator as a whole network, while only the generator, a partial GAN, addresses semantic segmentation. Therefore, it is possible to exclusively design a discriminator with a complex network structure for training, and the parameters in the generator can be optimized simultaneously.

The U-net is adopted as the backbone of the generator in this research work. Approaches to enhancing the feature extraction of the U-net are explored. In such circumstances, the classifier is not requested to recognize massive sorts of objects but only the limited objects labeled in the ground truths. Hence, finding out a light but effective discriminator is the ultimate goal of this research. In addition, the implementation impedes this research work from creating a highly complicated discriminator as well. A GAN is a combination of two networks, which entails an enormous computing power in training and thus challenges the hardware configuration of a research institution. Due to the primary reasons, four discriminators are proposed for test and validation. These four discriminators are composed of fully connected layers to reduce the data processing load in training. The contingent impacts caused by different lengths of a discriminator and thickness of a single layer in discriminators are considered in the model establishment and will be investigated in the experiments.

#### 4.2 The proposed discriminators

Fig. 3 a illustrates a D-4 GAN, in which the U-net serves as the generator, and it is combined with a discriminator with five fully connected layers (D-5). Each layer in the forepart encompasses three stages, a linear transformation to the incoming data, a ReLU activation function, and a dropout operation. After the image of ground truth and the segmented image are concatenated, the linear transformation is applied to extract
features by scaling the data according to the given size requirements, such as the 512, 1024, 256, 128, 64, 1 in Fig. 3 a. The transformed data should be activated by the ReLU function to maintenance the nonlinearity during data processing. The dropout operation is a practical technique to avoid the over-fitting problem. Given an empirical dropout rate, some elements of the input tensor are randomly zeroed with the probability of the dropout rate. The rear layer contains a sigmoid activation function and a linear transformer to switch the tensor with length 1. The D-6 GAN in Fig. 3 b is named by its discriminator with seven layers. This discriminator is extended on the basis of the D-4 GAN and has an identical rear layer. In the forepart of the D-6 GAN, two fully connected layers are inserted with the linear transformation (256 to 128 and 128 to 64) for a more gentle descent of the tensor size.

Inspired by the structure of the serial of VGG network series, the G-5 GAN is varied by duplication of linear and activation stages in the forepart of the discriminator, thus the variant is called the G-5V GAN. As Fig. 3 c demonstrates, each layer is inserted by a linear transformation stage with same feature size of in- and out-put, and a ReLU activation function. The D-5V GAN (see Fig. 3 d) is not a simple extension of the D-4V GAN. An extra complete fully connected layer is installed right before the rear part of the D-4V GAN. Moreover, the second stage in the second and the third layer in the D-5V GAN should be repetitively deducted.

Researchers prefer to propose a novel architecture or a network and examine its better prediction results than other existing proposals. In this research, novel architectures and networks are proposed firstly for comparison to determine the best ones and applied for validation afterward. Thus, the design of the four proposed discriminators is oriented to the comparative study. The segmentation results of these four GANs can be cross-referenced to insight into the impacts from lengths and thicknesses.

5 | IMPLEMENTATION FOR THE COMPARATIVE STUDY

The implementation of the proposed CNNs is executed through two steps. Initially, cell images are applied to compare the different combination of the U-net and attention mechanism in order to determine the CNNs owning the best performances. These selected networks are further employed for damage segmentation.
5.1 Datasets

The eventual objective of this paper is to determine the CNN models with the optimal performance in semantic segmentation and employ them to process the image of structural damage. As a matter of fact, images of structural damage (see Fig. 7) usually involve more noise objects, e.g. walls, windows, backgrounds. Therefore, before the images of structural damage are fed into the CNN models, it makes sense to deduct a lightweight test, in which simple data can be applied to investigate the preceding models initially with less time consumption, however, the prediction quality still can be well examined. Cell images (see Fig. 4 and 5) with similar graphical features like facade damage but less interference in images fulfill the data requirement of a lightweight test since these graphic data contain simple interference noise such as organelles but distinct cytoderm for segmentation. The lightweight test is essentially a comparative study to investigate the performances respectively of attention mechanism and the discriminator in cell image segmentation.

In the lightweight test, 24 images are adopted as the input images, the corresponding 24 images of ground truth are labels. The other five images are utilized as the test data; the prediction results of segmentation will be compared with these five images of ground truth via PA and mIoU. The four architectures, i.e., the U-net, the attention U-net, the advanced attention U-net, and the full attention U-net (see Fig. 1-a, 1-b, 1-c and 1-d) are trained with the same configuration of hyper-parameters. Such configuration does not incur these models to reach their own highest performance. However, on the basis of the unified standard of training hyper-parameters, the segmentation results are sufficiently plausible to reflect the strengths and weaknesses distinctly.

The dataset for facade damage segmentation applied the images of physics-based graphics models. These models are generated in a virtual digital environment and annotated with spalling areas, cracks, and exposed reinforcement bars to present the damage in facades after an earthquake or long-term weathering. The procedure of image acquisition also imitates the view angle and the movement trajectory of UAVs. This implementation focuses on the semantic segmentation of spall in the building facade. Therefore, the dataset is composed a vast number of images and the corresponding labels of spalling areas (see Fig. 7). The large amount of the original images have a resolution of 1980x1080. For an efficient implementation, 1,000 representative images and the corresponding ground truths are selected and resized into 512x512.

5.2 Training configuration

Since the configuration for training is not a research issue in this paper, it is hence empirical to select a loss function, an optimization algorithm, and metrics. Hereby, the RMSprop algorithm is chosen for the optimizer, the Binary Cross Entropy (BCE) with logits loss works as the loss function. Pixel Accuracy and the mean Intersection over Union (mIoU) are employed to evaluated the results of semantic segmentation.

5.2.1 Optimization algorithm

Gradient descent is one of the most prominent optimization methods, and it is by far the most frequent method for optimizing neural networks. Nowadays, every available Deep Learning framework (e.g., Tansorflow [Abadi et al. 2015], Pytorch [Paszke et al. 2019], Keras [Chollet et al. 2015]) includes numerous methods for optimizing gradient descent. However, since practical descriptions of their benefits and limitations are difficult to come by, these algorithms are frequently employed as black-box optimizers. Gradient descent is a method for minimizing an objective function parameterized by model parameters by updating the parameters in the opposite direction as the gradient of the objective function. The learning rate determines the magnitude of the steps taken to attain a minimum.

RMSprop in Eq. 2 to Eq. 5 is a gradient-based optimization method used in neural network training. Gradients of extremely complicated functions, such as neural networks, have a propensity to disappear or explode as data passes through the function. RMSprop is a stochastic approach for mini-batch learning. RMSprop addresses the aforementioned issue by normalizing the gradient using a moving average of squared gradients. This normalization balances the step size, reducing the step size for big gradients to prevent bursting and increasing the step size for small gradients to avoid disappearing.

\[
v_t = \beta_1 v_{t-1} - (1 - \beta_1) g_t
\]

(2)

\[
s_t = \beta_2 s_{t-1} - (1 - \beta_2) g_t^2
\]

(3)

\[
\Delta \omega_t = -\eta \frac{v_t}{\sqrt{s_t + \epsilon}} g_t
\]

(4)

\[
\omega_{t+1} = \omega_t + \Delta \omega_t
\]

(5)
\( \eta \) : initial learning rate  
\( g_t \) : gradient at time \( t \) along \( \omega_j \)  
\( v_t \) : exponential average of gradients along \( \omega_j \)  
\( s_t \) : exponential average of squares of gradients along \( \omega_j \)  
\( \beta_1, \beta_2 \) : hyperparameters

### 5.2.2 Loss function

During training a CNN, the degree of deviation between the predicted and input data is represented by loss calculated by loss functions. The calculated loss affects the backpropagation over training epochs. A larger loss indicates the more significant difference between the inputs and predictions. The loss should be minimized epoch by epoch in training iteration as much as possible. Thus a minor loss is always preferred in model training. The BCE with logits loss is adopted in this research work. This loss represented in Eq. 6 combines a sigmoid layer and the BCE loss (Torch Contributors, 2019) in one single class.

\[
l_n = -\omega_n \left\{ t_n \cdot \log (x_n) + (1 - t_n) \cdot \log (1 - \sigma (x_n)) \right\}
\]

\( l_n \) : calculated loss in the \( n^{th} \) loop  
\( \omega_n \) : weight parameter in the \( n^{th} \) loop  
\( x_n \) : precision of correct prediction in the \( n^{th} \) loop  
\( t_n \) : target in the \( n^{th} \) loop  
\( \sigma () \) : sigmoid activation function in the \( n^{th} \) loop

### 5.2.3 Metrics

PA in Eq. 7 provides the percentage of pixels in the images identified adequately as a metric for evaluating semantic segmentation. This metric is usually provided independently for each class as well as globally across all classes. In the assessment of per-class pixel accuracy, a binary mask is evaluated. A true positive indicates a properly projected pixel to belong to the given class according to the target mask. In contrast, a true negative represents a correctly detected as not belonging to the given class. However, when the class representation is minimal inside the image, this metric can often produce deceptive results since the measure is skewed in reporting how effectively you detect negative cases.

\[
PA = \frac{TP + TN}{TP + TN + FP + FN}
\]

\( TP \) : true positive  
\( TN \) : true negative  
\( FP \) : false positive  
\( FN \) : false negative

The IoU metric in Eq. 8 is a way for quantifying the percentage overlap between the target mask and our prediction output. The IoU metric is calculated by dividing the number of pixels shared by the target and prediction masks by the total number of pixels contained in both masks.

\[
IoU = \frac{\text{target} \cap \text{prediction}}{\text{target} \cup \text{prediction}}
\]

The intersection target \( \cap \) prediction is made up of pixels from both the prediction and ground truth masks, whereas the union target \( \cup \) prediction is made up of pixels from either the prediction or target mask. The IoU score is generated individually for each class and averaged across all classes to produce the semantic segmentation prediction’s global, namely mean IoU (mIoU) score.

### 5.3 A lightweight test using cell images

Tab. 1 displays the result evaluation of the lightweight test. PA emphasizes the correct location of the pixels in the prediction. According to the metric, the full attention U-net and advanced attention exhibit approximately 0.3% to 0.4% higher than the U-net and the attention U-net. The mIoU is a more comprehensive evaluation considering the locations of right and wrong pixels in the prediction results. Though the prediction enhancement caused by the attention mechanism is distinctly presented, it is uncertain that the prediction quality is positively correlated with the distribution density of the attention mechanism. The mIoU of advanced U-net, 78.22%, shows a significant advantage, as the full attention U-net with more attention gate and more complex combination of skip connections has merely the mIoU of 77.97%. The examples in Fig. 4 endorse the data in Tab. 1 as well. Fewer and lighter noise spots appear in Fig. 4 e and 4 f than those in Fig. 4 c and 4 d. The cross-comparison of Fig. 4 e and 4 f indicates that the advanced attention U-net is able to predict the cytoderm with higher completeness. Therefore, consequently, the advanced attention U-net performs best due to its appropriate distribution of attention gates and utilization of skip connections.

The attached discriminator in a GAN can alert the loss function’s tendency to convergence. The four proposed discriminators are coupled with the U-net for training and testing. The discriminator comparative study investigates the four discriminators constructed with different layer numbers and thicknesses. The two metrics of the D-6 GAN are 0.11% and 0.29%
TABLE 1 Result evaluation of the lightweight test

| Method               | PA (%) | Δ (%) | mIoU (%) | Δ (%) |
|----------------------|--------|-------|----------|-------|
| U-net                | 91.55  | -     | 77.56    | -     |
| Attn. U-net         | 91.50  | -0.055| 77.73    | +0.22 |
| Adv. attn. U-net    | 91.88  | +0.36 | 78.22    | +0.85 |
| Full attn. U-net    | 91.85  | +0.33 | 77.97    | +0.53 |

PA: Pixel Accuracy
Δ: the increment compared with the results from the U-net
mIoU: the mean Intersection over Union

* Attn. U-net: the attention U-net
* Adv. attn. U-net: the advanced attention U-net
* Full attn. U-net: the full attention U-net

(see Tab. 1) higher than those two values of the D-4 GAN. Moreover, the results of the D-5V GAN are relatively better than the D-4V GAN, even though the layers in the two architectures are not completely conceived. In the condition that two discriminators are established with the same layer thickness, the discriminator with more layers segments the images with higher accuracy in PA and mIoU. Two discriminators (e.g., the D-4 GAN and the D-4V GAN) with the same length can deliver different results for the different layer designs. Thus, the more operations are inserted in a discriminator’s layer, the worse metrics of the discriminator can be obtained. The illustrations in Fig. 5 reflect the calculated metrics. The predictions of the D-6 and D-5V GANs include fewer interference objects in cells and more details of cell walls. According to the lightweight test, the strategy of advanced attention materializes a significant promotion in semantic segmentation; the generative adversarial network with discriminator including six fully connected layers wins the competition with the other three proposed ones. Therefore, these two architectures (see Fig. 1 c and 3 b) are employed in the experiment. Besides, a complex U-net, which is composed of the U-net, the advanced attention mechanism, and the discriminator with six layers, is constructed and experimented with the images of structural damage for validation to explore the combined effects in semantic segmentation. The U-net (see Fig. 1 a) participates in the experiment, too, as the reference for result analysis.

5.4 Damage segmentation for on-site inspection

The evaluation of the segmentation quality is executed through two steps. The PA and mIoU of the best ten results are firstly calculated (see Tab. 2). After that, the entire 50 segmentation images are evaluated by the two same metrics (see Tab. 2). The original images corresponding to the selected results have several common features, such as larger areas, distinct edges,
and higher color contrast between the spall and other objects (see Fig. 7).

As the values in Tab. 2 indicate, three enhanced architectures materialise the prominent improvement in semantic segmentation with different significance. The advanced attention U-net results in a better evaluation in 50 images than the selected ten. The U-net attached by the D-6 presents a robust capability of semantic segmentation, the quality of input images on that is much less than the advanced attention U-net. Though the D-6 U-net provides a higher PA and mIoU than the advanced attention U-net in segmenting the ten selected images, both architectures show comparable performance in facing a more generic dataset. The combined network shows stable results, which are also better than the U-net. Nevertheless, the results appear to have adverse effects from combining the attention mechanism and a discriminator in a U-net because the evaluations of the combined network are apparently lower than the other two enhanced U-nets.

The semantic segmentation results reveal another angle of observation. Fig. 7 depicts that the strength of the attention mechanism is prone to detect the edges of an object. The combined network performs with a weaker edge identification, and the U-net + D-6 is the worst one. This observation seems does not endorse the calculated metrics. However, as a matter of fact, the metrics, PA and mIoU, are widely applied for a more comprehensive evaluation to examine the exact locations of each pixel area. Visual observation is exclusively able to discover few aspects in result assessment.

### TABLE 2 Result evaluation of the 10 and 50 selected images in validation

|                  | 10 images |                  | 50 images |
|------------------|-----------|-----------------|-----------|
|                  | PA(%)     | Δ(%)            | mIoU(%)   | Δ (%)    |
| U-net            | 96.06     | -               | 75.29     | -        |
| Adv. attn. U-net | 97.19     | +1.16           | 81.15     | +7.78    |
| U-net + D-6      | **97.38** | **+1.37**       | **82.50** | **+9.58**|
| Com. network     | 96.98     | +0.96           | 80.17     | +6.48    |
|                  |           |                 |           |          |
|                  | PA(%)     | Δ(%)            | mIoU(%)   | Δ (%)    |
| U-net            | 96.02     | -               | 68.64     | -        |
| Adv. attn. U-net | 97.98     | **+2.04**       | 74.97     | **+9.22**|
| U-net + D-6      | **98.02** | **+2.06**       | **74.80** | **+8.97**|
| Com. network     | 97.80     | +1.85           | 73.04     | +6.41    |

a PA: Pixel Accuracy  
b Δ: the increment compared with the results from the U-net  
c mIoU: the mean Intersection over Union  
d Adv. attn. U-net: the advanced attention U-net  
e Com. network: the network combined with the advanced attention U-net and the discriminator D-6

### 6 DISCUSSION

The attention mechanism and the discriminator are supposed to be two individual components in the frame of a GAN. The former extracts the features from the inputs in a generator’s different stages; the latter evaluates the generator’s outputs and feeds back for parameter adjustment during training. Incorporating the attention mechanism into a GAN is primitively expected to combine these two components’ strengths. On the contrary, Tab. 2 proves this kind of combination does not...
function as the expectation. The segmentation quality of the combined network is worse than the other two enhanced U-nets rather than reaches the best values. To the author's best knowledge, this inharmonious orchestration between the attention mechanism and a discriminator was the very first to be discovered and discussed.

In order to investigate whether the inharmonious orchestration is a contingency or not, another group of architectures is chosen from a comparative study, i.e., the full attention U-net, the D-4 U-net, and their combination (see Fig. 8). These three models are trained with the identical training configuration to the previous group in the lightweight test, and tested by the same dataset with 50 images. According to the assessment in Tab. 3, this combination does work better than the other two U-nets. However, none of the tested architecture delivers more optimal values of PA and mIoU.

After all, this test manifests that the collaboration of the attention mechanism and a discriminator could negatively impact semantic segmentation, but exclusively in certain circumstances. One possible explanation could be that the feedback information transmitted from the discriminator to the generator interferes with the attention operation on the decoder's side. The full attention mechanism is not or barely influenced because most of the attention gates are distributed

FIGURE 7 Data and semantic segmentation results for the on-site inspection
on the decoder’s side. This analysis is partially endorsed by the GAN model developed by G. Zhang, Pan, & Zhang (2021). That proposed GAN is integrated with the attention mechanism as well. However, the attention block is placed in front of the U-net. In other words, the data fed to the inputs have been previously attended outside of the U-net. The mathematical analysis could be agnostic and heterogeneous so that further study should focus on this phenomenon.

7 | CONCLUSION

This paper presents a comparative study for semantic segmentation focusing on on-site inspection. Inspired by the attention mechanism and the design of the U-net, two novel architectures integrated with attention gates are proposed, i.e., the advanced attention U-net and the full attention U-net. Furthermore, four novel architectures of discriminators in U-net based GANs are proposed as well, namely the D-4 U-net, the D-6 U-net, the D-4V U-net, and the D-5V U-net. A lightweight experiment is executed to compare these enhanced U-net architectures in two groups. The U-net, the attention U-net and the two brand-newly proposed U-nets with attention mechanisms are marked as the first group, in which the advanced attention U-net achieves the best results of semantic segmentation. Four newly proposed GANs are compared with the U-net in the second group. Consequently, the D-6 U-net shows the highest segmentation accuracy. Hence, the advanced U-net, the D-6 U-net, and the U-net combined of both are selected for the further implementation, where the graphical data acquired from a pseudo on-site inspection. The results prove the improved capability of the tested architectures in semantic segmentation. PA and mIoU are calculated to designate the strengths of the advanced U-net and the D-6 U-net in segmenting ten selected images and the whole dataset.

Meanwhile, the adverse effects occasioned by the combination of a discriminator and attention mechanism are discovered. For further investigation, another group is trained and tested in the experiment, including the full attention U-net, the D-4 U-net, and the U-net combined with both. These test results demonstrate positive combination effects and thus stand on the opposite of the previous test. The comparison with all the groups ascribes this phenomenon to the different distribution of attention gates in the proposed enhanced U-nets.

It is to be noticed that the inherent shortcoming of the U-net lies in segmenting slim objects. Though the spall is well recognized in this research work, semantic segmentation of slim damage such as cracks can not be easily segmented. Therefore, other strategies, e.g., featured pyramid network and residual operation, could be integrated in the future to promote the semantic segmentation quality of the proposed CNNs.
ACKNOWLEDGMENTS

The authors would like to acknowledge the organization committee of The 2nd International Competition for Structural Health Monitoring (IC-SHM, 2021) that shared the graphical data for training and test, and the hardware support from Microintelligent Inc.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). Tensorflow: Large-scale machine learning on heterogeneous systems. Retrieved from http://tensorflow.org/ Software available from tensorflow.org.

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In The proceedings of 3rd International Conference on Learning Representations. San Diego. Retrieved 2021-09-17, from http://www.scopus.com/inward/record.uri?scp=85062889504&partnerID=8YFLogxK

Beek, P. J. L. v., Reinders, M. J. T., Sankur, B., & Lubbe, J. C. A. v. d. (1992). Semantic segmentation of videophone image sequences. In Visual Communications and Image Processing ’92 (Vol. 1818, pp. 1182–1193). International Society for Optics and Photonics. doi:

Chen, J., Zhang, D., Huang, H., Shadafar, M., Zhou, M., & Yang, T. (2020). Image-based segmentation and quantification of weak interlayers in rock tunnel face via deep learning. Automation in Construction, 120, 103371. doi:

Chollet, F., et al. (2015). Keras. https://keras.io

Csurka, G., & Perronnin, F. (2011). An efficient approach to semantic segmentation. International Journal of Computer Vision, 95(2), 198–212. doi:

Czerniawski, T., & Leite, F. (2020). Automated segmentation of RGB-D images into a comprehensive set of building components using deep learning. Advanced Engineering Informatics, 45, 101131. doi:

Dais, D., Bal, A. E., Smyrou, E., & Sarhosis, V. (2021). Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning. Automation in Construction, 125, 103606. doi:

Decourt, C., & Duong, L. (2020). Semi-supervised generative adversarial networks for the segmentation of the left ventricle in pediatric MRI. Computers in Biology and Medicine, 123, 103884. doi:

Ghassemi, N., Shoebi, A., & Rouhani, M. (2020). Deep neural network with generative adversarial networks pre-training for brain tumor classification based on MR images. Biomedical Signal Processing and Control, 57, 101678. doi:

Girshick, R. (2015). Fast R-CNN. arXiv:1504.08083 [cs].

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial networks. ArXiv, abs/1406.2661.

Hang, R., Li, Z., Liu, Q., Ghamisi, P., & Bhattacharyya, S. S. (2021). Hyperspectral image classification with attention-aided CNNs. IEEE Transactions on Geoscience and Remote Sensing, 59(3), 2281–2293. Conference Name: IEEE Transactions on Geoscience and Remote Sensing. doi:

He, K., Gkioxari, G., Dollâr, P., & Girshick, R. (2018). Mask R-CNN. arXiv:1706.03787 [cs].

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv:1512.03385 [cs].

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. doi:

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., … Wu, J. (2020). UNet 3+: A full-scale connected UNet for medical image segmentation. arXiv:2004.08790 [cs, eess].

Huang, J., Liu, S., Tang, Y., & Zhang, X. (2021). Object-Level Remote Sensing Image Augmentation Using U-Net-Based Generative Adversarial Networks. Wireless Communications and Mobile Computing, 2021, e1230279. doi:

Inoguchi, M., Tamura, K., & Hamamoto, R. (2019). Establishment of workflow for roof damage detection utilizing drones, human and AI based on human-in-the-loop framework. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4618–4623). doi:

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. doi:

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. doi:

LeCun, Y., Cortes, C., & Burges, C. J. (1998). MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. Retrieved 2021-07-30, from http://yann. lecun.com/exdb/mnist/index.html

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. arXiv:1411.4038 [cs].
Ma, J. W., Czerniawski, T., & Leite, F. (2020). Semantic segmentation of point clouds of building interiors with deep learning: Augmenting training datasets with synthetic BIM-based point clouds. Automation in Construction, 113, 103144. doi:

Mehta, R., & Sivaswamy, J. (2017). M-net: A convolutional neural network for deep brain structure segmentation. In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017) (pp. 437–440). doi:

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. arXiv:1606.04797 [cs].

Mou, L., Hua, Y., & Zhu, X. X. (2019). A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In (pp. 12416–12425).

Ou, X., Yan, P., Zhang, Y., Tu, B., Zhang, G., Wu, J., & Li, W. (2019). Moving object detection method via ResNet-18 with encoder-decoder structure in complex scenes. IEEE Access, 7, 108152–108160. doi:

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, & R. Garnett (Eds.), Advances in neural information processing systems 32 (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. arXiv:1505.04597 [cs].

Schlemper, J., Oktay, O., Schap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis, 53, 197–207. doi:

Seo, J., Duque, L., & Wacker, J. (2018). Drone-enabled bridge inspection methodology and application. Automation in Construction, 94, 112–126. doi:

Shim, S., & Cho, G.-C. (2020). Lightweight semantic segmentation for road-surface damage recognition based on multiscale learning. IEEE Access, 8, 102680–102690. doi:

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs].

Tian, C., Xu, Y., Li, Z., Zuo, W., Fei, L., & Liu, H. (2020). Attention-guided CNN for image denoising. Neural Networks, 124, 117–129. doi:

Tong, Z., Xu, P., & DenAṣux, T. (2021). Evidential fully convolutional network for semantic segmentation. Applied Intelligence. doi:

Torch Contributors. (2019). BCELoss â/uni0102/uni0164 PyTorch 1.9.1 documentation. Retrieved 2021-10-11, from https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html

Usama, M., Ahmad, B., Song, E., Hussain, M. S., Alrashoud, M., & Muhammad, G. (2020). Attention-based sentiment analysis using convolutional and recurrent neural network. Future Generation Computer Systems, 113, 571–578. doi:

Xia, X., & Kulis, B. (2017). W-Net: A deep model for fully unsupervised image segmentation. arXiv:1711.08506 [cs].

Xiong, D., He, C., Liu, X., & Liao, M. (2020). An end-to-end bayesian segmentation network based on a generative adversarial network for remote sensing images. Remote Sensing, 12(2), 216. doi:

Zhang, G., Pan, Y., & Zhang, L. (2021). Semi-supervised learning with GAN for automatic defect detection from images. Automation in Construction, 128, 103764. doi:

Zhang, Y., Miao, S., Mansi, T., & Liao, R. (2020). Unsupervised X-ray image segmentation with task driven generative adversarial networks. Medical Image Analysis, 62, 101664. doi:

Zhao, S., Shadabfar, M., Zhang, D., Chen, J., & Huang, H. (2021). Deep learning-based classification and instance segmentation of leakage-area and scaling images of shield tunnel linings. Structural Control and Health Monitoring, 28(6), e2732. doi:

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., … Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12196–12205). Retrieved 2021-10-09, from http://arxiv.org/abs/2012.15840 arXiv:2012.15840.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-net architecture for medical image segmentation. arXiv:1807.10165 [cs, eess, stat].

How to cite this article: Lin F., Yang J., Shu J., Scherer R. J., Guo J., A comparative study of attention mechanism and generative adversarial network in facade damage segmentation. Comput Aided Civ Inf., 20xx;xx:x–x.