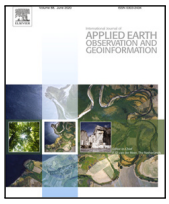
International Journal of Applied Earth Observation and Geoinformation 136 (2025) 104259

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



# An enhanced network for extracting tunnel lining defects using transformer encoder and aggregate decoder

Bo Guo [a,g], Zhihai Huang [b], Haitao Luo [c], Perpetual Hope Akwensi [d], Ruisheng Wang [a,d,g,*], Bo Huang [e], Tsz Nam Chan [f]

[a] School of Architecture and Urban Planning, Shenzhen University, Shenzhen, 518060, Guangdong, China
[b] School of Civil and Transportation Engineering, Guangdong University of Technology, Guangzhou, 510006, Guangdong, China
[c] Guangzhou Metro Design & Research Institute Co., Ltd, Guangzhou, 510010, Guangdong, China
[d] Department of Geomatics Engineering, University of Calgary, Calgary, T2N 1N4, Alberta, Canada
[e] Department of Geography, The University of Hong Kong, Pokfulam, Hong Kong
[f] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, Guangdong, China
[g] State Key Laboratory of Subtropical Building and Urban Science, Shenzhen, 518060, Guangdong, China

ARTICLE INFO

*Keywords:*
Deep learning
Transformer
Attention mechanism
Defect detection
Tunnel lining

ABSTRACT

The tunnel environment is characterized by insufficient ambient light, obstructed view, and complex inner lining construction conditions. These factors frequently result in limited anti-interference capability, reduced recognition accuracy, and suboptimal segmentation results for defect extraction. We propose a deep network model utilizing an encoder–decoder framework that integrates Transformer and convolution for comprehensive defect extraction. The proposed model utilizes an encoder that integrates a hierarchical Transformer backbone with an efficient attention mechanism to fully explore complete information at multi-scale granularities. In the decoder, multi-scale information is initially aggregated using a Multi-Layer Perceptron (MLP) module. Additionally, the Stacking Filters with Atrous Convolutions (SFAC) module are implemented to enhance the perception of the complete defect scope. Furthermore, a Boundary-aware Attention Module (BAM) is implemented to enhance edge information to improve the detection of defects. With this well-designed decoder, the multi-scale information from the encoder can be fully aggregated and exploited for complete defect detection. Experimental findings illustrate the effectiveness of our proposed approach in addressing tunnel lining defects within the image dataset. The outcomes reveal that our proposed network achieves an accuracy (Acc) of 94.4% and a mean intersection over union (mIoU) of 78.14%. Compared to state-of-the-art segmentation networks, our model improves the accuracy of tunnel lining defect extraction, showcasing enhanced extraction effectiveness and anti-interference capability, thus meeting the engineering requirements for defect detection in complex environments of tunnels.

## 1. Introduction

Convenient and rapid transportation is crucial for enhancing productivity and quality of life. As an indispensable component of contemporary transportation, the construction of tunnels has substantially shortened the distance between regions and greatly saved transportation time. The widespread adoption of tunnel-based subways has become pivotal in urban transportation, effectively alleviating road traffic congestion. However, affected by factors such as groundwater, ground settlement, and seismic activity, subway tunnels may suffer from various issues such as cracking, leakage, and lining detachment (Peng, 2016). These defects disrupt the normal operation of subway tunnels. Hence, the timely detection and repair of lining defects in subway tunnels are crucial.

Current methods for detecting defects in tunnel linings primarily involve manual and instrumental inspection. With manual inspection, skilled inspectors must make numerous decisions regarding a variety of potential defects. The accuracy of manual inspection has a negative correlation with the complexity of the defects inspected. The oversight of small but serious defects can lead to disaster. Instrumental inspection often requires specialized equipment, limiting its applicability for the quick detection of defects (Deng, 2015). Thus, the development of efficient and effective techniques for defect detection is urgently needed.

* Corresponding author at: School of Architecture and Urban Planning, Shenzhen University, Shenzhen, 518060, Guangdong, China.
*E-mail address:* ruiswang@szu.edu.cn (R. Wang).

https://doi.org/10.1016/j.jag.2024.104259
Received 1 February 2024; Received in revised form 21 October 2024; Accepted 5 November 2024
Available online 16 December 2024

Advances in digital image processing technology have led to the emergence of diverse automated approaches for detecting defects in tunnel lining images. Utilizing edge detection, morphology, and region-growing algorithms, conventional methods can approximately identify most defect areas. However, these methods face challenges in adapting to changes in image collection conditions, such as changes in illumination, shooting angles, or distances, leading to poor generalization and performance drawbacks (Rui Gong and Chaohua, 2020; Weibo Xu, 2017; Diao et al., 2010). Recently, image processing has advanced significantly, with the fast growth of machine learning technologies, including region growing and clustering algorithms (Qu et al., 2010; Lei and Zuo, 2009). However, their robustness for detecting defects remains insufficient, especially in complex tunnel environments. Deep learning, known for its ultra-high-level feature representation and robust learning ability derived from extensive training data, offers advantages such as rapid recognition speed and high segmentation accuracy. The Deep learning-based method enables superior object detection and target extraction (Liujie Sun and Wenju, 2022). However, existing deep neural networks face challenges in fully extracting defects from tunnel lining images, owing to the varying scales of defects and poor lighting conditions in tunnels.

We consider two primary challenges in applying deep neural networks for the task of defect extraction. First challenge involves extracting sufficient information from poorly lit images captured in the dark environment of tunnels. Convolutional Neural Networks (CNNs) are widely used end-to-end models that can make label predictions at every pixel (Bi Yang, 2016). To enhance the feature perception range, it is necessary to concatenate an adequate number of convolutional layers with local receptive field to get a global one (Simonyan and Zisserman, 2015). However, an excessively deep network will lead to accuracy saturation and model degradation, posing a bottleneck in the continued development of CNNs (He et al., 2016; Dosovitskiy et al., 2020; Zheng et al., 2021). The utilization of the attention mechanism in Transformer demonstrates a globally enhanced receptive field, thereby helping to avoid information bias (Cao et al., 2022; Xie et al., 2021a; Chen et al., 2021). The advantage of modeling long-range feature allows Transformers to better process global information in images. Even in low-light conditions, where many local details may be unclear, Transformers can still extract valuable features from the overall structure. We adopted a novel Transformer backbone to fully extract contextual information. Compared to using only local features extracted with CNN, our proposed model enables the network to better address various challenges in object recognition with poor or highly variable lighting conditions.

Another challenge arises from the presence of defects with multiple granularities. In a tunnel environment, small but serious defects can occur, leading to catastrophic accidents. Hence, defect regions at all granularities need to be comprehensively extracted. To address this issue and obtain sharp segmentation results, we propose a module that gradually increases the perceptual range of features by implementing a series-parallel structure with multiple dilated convolutions to fully extract defects (Chen et al., 2017a; Yu and Koltun, 2015). Tunnels typically contain multiple pipelines or traffic facilities that may obstruct defects and thus interfere with defect detection. Furthermore, due to the complexity of the tunnel environment, accurately determining defect boundaries is often challenging. Local fine-grained features with rich edge information, extracted from the shallow layers of the encoder, should be fully exploited in the decoder to generate accurate defect boundaries. Therefore, in the decoder, we leverage features from the shallow layers of the encoder to enhance boundary information.

To conquer the challenges to fully extract defects with multiple granularities in non-uniform lighting environments, we propose a hierarchical deep network with a reliable Transformer-based encoder and a well-designed decoder to comprehensively extract lining defect regions. Specifically, we use a sequence reduction attention (SRA) mechanism to reduce attention computational complexity. Building upon efficient attention, a Transformer backbone is introduced as the encoder to output multi-layer features containing local fine-grained and global contextual information. For the decoder, we use a simple yet efficient MLP concatenation module to connect all individual branches from the encoder to aggregate information at different granularities. Then, stacking filters with atrous convolutions are implemented to enhance the perceptual range, facilitating the extraction of complete defect regions. Meanwhile, we introduce a boundary-aware attention module to fully explore boundary information of low-level features within shallow layers in the encoder. This information is directly connected to the decoder for accurate edge boundary extraction. Finally, we utilize an improved loss function by combining the Dice function and cross-entropy function to achieve robust segmentation results and fast network convergence. This paper makes the following primary contributions:

1. To address the issue of insufficient information extraction caused by the dark tunnel environment, a hierarchical Transformer backbone enhances the capability of extracting useful context features through incorporating attention mechanism and overlap embedding and merging to output multi-layer feature maps to facilitate accurate segmentation.
2. To comprehensively extract defects of various granularities, we enhance the decoder's capability by implementing a series-parallel structure with multiple dilated convolutions to continuously enlarge the perceptual range, which ensures a comprehensive detection of defect areas with multiple granularities in tunnel lining.

To conclusively showcase the effectiveness and efficiency of our proposed network in identifying defects within subway tunnel linings, we carried out a comprehensive set of experiments. Meanwhile, ablation experiments validate the efficacy of well-designed modules. The results affirm that our proposed model surpasses existing approaches and establishes a new baseline performance in tunnel lining defect detection.

## 2. Related work

### 2.1. Tunnel lining defect detection

Common approaches for detecting tunnel lining defects include inspections using specific equipment such as knocking hammers, and infrared thermal cameras to assess the tunnel lining stability (Huang et al., 2021; Afshani et al., 2019; Zhang et al., 2019). Following this, hand-crafted recognition-based methods are employed to extract the locations of defects. However, in the majority of implementations, the use of powerful hardware devices and significant human resources is required, often leading to missed or false detections of defects (Attard et al., 2018).

Recently, laser scanners and cameras have become commonplace for fast and high-precision large-scene data collection, rendering them suitable for tunnel surveillance (Menendez et al., 2018; Kim and Lee, 2018; Xu and Yang, 2019; Cao et al., 2019). Typically, relying on simple feature extraction algorithms, image and point cloud data collected using these devices can be processed to focus primarily on detecting crack-related defects (Cha et al., 2017; Dung and Anh, 2019; Xue and Li, 2018). Cao et al. utilized laser scanning point clouds to construct a tunnel lining model and successfully implemented model-based tunnel defect monitoring by identifying specific bolts used for assembly connections in 3D tunnel blocks (Cao et al., 2021).

Alidoost et al. developed a visual-based system for damage and object detection tasks in roadway tunnels using deep learning (Alidoost et al., 2023). However, their convolution algorithm is not optimized for the tunnel environment, resulting in fragmented segmentation outcomes, which requires further improvement. Xue et al. collected defect information using a charge-coupled device (CCD) camera and employed a region-based CNN to identify possible defect location boxes (Xue and

Li, 2018). However, this approach fails to capture the specific morphological distribution of defects, which is not conducive to subsequent maintenance guidance. Huang et al. proposed a network model based on Mask R-CNN, augmented with an artificially designed morphological closing operation to detect the complete shape of defects (Huang et al., 2022). However, the limited efficacy of artificial design methods indicates the potential for enhancement by incorporating more advanced approaches. Liu et al. utilized a laser scanner to perform round-trip measurements to detect water leakage using intensity images (Liu et al., 2022a). They enhanced Mask R-CNN by incorporating Res2Net into the backbone, thus proposing a multi-scale instance segmentation framework with a cascade structure. Their work demonstrated that enlarging the receptive field can improve tunnel defect segmentation capabilities. However, they overlooked the issue of pipeline occlusion in defect extraction. Zhou et al. addressed the challenge of detecting long cracks and linear seams by leveraging attention mechanisms to embed channel and positional information (Zhou et al., 2022). They introduced an efficient multi-scale feature fusion technology in the Tunnel Crack Detection Network (TCDNet), which can effectively capture the long-range dependencies of crack characteristics.

It is crucial to optimize the network according to specific target characteristics. Therefore, we have developed an advanced defect detection method that takes into account environmental interference and the distribution characteristics of defects. Meanwhile, expanding the receptive field is conductive to comprehensive perception of potential defects. We enhance the perception extraction capability in the proposed architecture.

### 2.2. Deep neural networks for image segmentation

In this section, we delineate relevant studies and works about image segmentation for defect detection using deep network.

#### 2.2.1. Convolution based models

CNN is a popular deep learning-based image processing technique that can automatically learn and extract features through connecting multiple convolutions, fully connected layers, and pooling layers (Bi Yang, 2016). The component modules of CNN significantly impacts the network's capabilities. Long et al. initially introduced the concept of Fully Convolutional Networks (FCNs) by substituting fully connected layers with convolutional layers, thereby accomplishing pixel-level semantic segmentation tasks (Long et al., 2017). Reasonable network architecture is another aspect of improving the network. The encoder–decoder structure first downsamples the image and then gradually decodes and restores the target information to the original scale. This design makes the model well-suited for recovering multi-scale details of target objects (Zhou et al., 2018; Lin et al., 2017; He, 2020; Erdem et al., 2023). The U-Net for medical image segmentation, leverages skip connections to integrate comprehensive information with multiple granularities Ronneberger et al. (2015). Bhowmick et al. further modified U-Net for bridge crack detection by using color image segmentation (Bhowmick et al., 2020). However, the segmentation performance calls for further improvements owing to the limited training samples. Dian et al. introduced TCS-Net by incorporating an attention compensation module in skip connections, which effectively enhances target focus weights for better detection of fine cracks on the safety shell (Dian et al., 2022). Optimizing neural networks entails the minimization of a loss function. The integration of the loss function is crucial for governing the reverse learning process, enabling optimal target localization (Chen and Qi, 2019). Feng et al. proposed SOLOv2-TL, which utilizes ResNeXt-50 as a backbone, incorporating deformable convolutions and a path-enhanced feature pyramid network (PAFPN) for liner leakage detection (Feng et al., 2023). The bottleneck structure helps mitigate information loss and enhances the accuracy of detecting leaks of various sizes within layers.

Taking the encoder–decoder framework into full consideration, and fully fusing shallow and deep features, DeepLabV3+ achieves significant semantic segmentation accuracy (Chen et al., 2018). It has been implemented to detect various objects. Ji et al. developed pixel-level fracture identification in asphalt pavement using the DeepLabv3+ (Ji et al., 2020), enabling effective detection and quantitative computation of the cracks. However, their work lacks optimized improvements for the detection of non-linear targets. Bhakti et al. optimized the dilation rates in Atrous Spatial Pyramid Pooling (ASPP) to adapt to the characteristics of the traffic environment for scene recognition (Baheti et al., 2020). However, the ASPP design requires customization to align with the unique characteristics of different application scenarios.

Leveraging an efficient encoder–decoder segmentation architecture, our proposed network incorporates a combination of serial and parallel connections with dilated convolutions. This progressive increment in the perceptual range enhances the network's ability to comprehensively extract defects at diverse granularities, ensuring a more detailed detection.

#### 2.2.2. Transformer and attention mechanism

While convolutional networks showcase robust feature extraction capabilities, particularly in capturing local features within an image, their upper limit is distinctly defined. Traditional deep convolution architectures rely on the stacking of multiple local convolutions to achieve depth in their architecture, posing challenges in maintaining identity mapping between layers, resulting in inductive bias (He et al., 2016). Simultaneously, the convolutional network architecture maintains a relatively fixed receptive field, unable to adapt to various sizes of the target areas, resulting in sporadic segmentation block issues. In response to these shortcomings, an attention mechanism can be used to effectively augment the network's capacity to capture essential global context. This capacity addresses limitations in adaptability to diverse target area sizes.

The attention mechanism has played a crucial role in driving recent advancements across various recognition tasks. It enriches the representation capability of features through the modeling of distant connections (Liu et al., 2024; Agrawal et al., 2024). Following the remarkable success in Natural Language Processing (NLP) with the Transformer (Vaswani et al., 2017), Dosovitskiy et al. extended its application to image processing (Dosovitskiy et al., 2020).

There have been many efficient Transformer models for image processing. The Swin-Transformer reduces the enormous computational cost by leveraging the fusion of sliding windows (Liu et al., 2021). Zhou et al. integrated the Swin-Transformer and CNN to propose a hybrid semantic segmentation network, SCDeepLab, for detecting tunnel lining cracks (Zhou et al., 2023). It extracts both deep and shallow features using a joint backbone, achieving effective segmentation of lining cracks. However, the substantial computational load is a notable drawback. Chen et al. proposed TransUNet, a model explicitly designed for medical image segmentation (Chen et al., 2021). By integrating Transformers with UNet, TransUNet enhances the extraction of global information, albeit with an increase in computation costs. Cao et al. introduced the Swin-Unet, which constructs a framework solely based on the Transformer mechanism. Their work demonstrates that preserving only skip connections can still significantly improve segmentation accuracy without the need for convolution or interpolation (Cao et al., 2022). However, as a result of employing window-partitioned feature blocks, the intricate nature of information exchange between windows becomes apparent. While the performance of these models continues to improve, persistent challenges still need to be addressed. Spatial Reduction Attention (SRA) is an efficient method to reduce complexity (Wang et al., 2021). Feng et al. combined CNN with ViT to develop a tunnel lining defect classification network (Feng et al., 2024). Following this, they utilized a You Only Look Once (YOLO) network with a multi-class head to detect multiple defects from the classified images. This two-step approach realizes multi-defect recognition in tunnel lining.

Our proposed network is built based on Transformer. By utilizing the powerful contextual relationship extraction mechanism inherent in the Transformer architecture, and combining it with a decoder that systematically expands the perceptual range of features, our carefully optimized network model ensures effective performance in handling diverse lining defect shapes in complex tunnel environments.

## 3. Method

We propose an encoder–decoder architecture designed to efficiently and effectively detect defects in tunnel linings. To address the issue of insufficient lighting in tunnels, which hampers the comprehensive extraction of defect features, we employ a Transformer backbone in the encoder. The backbone integrates a hierarchical Transformer with an efficient attention mechanism to fully explore complete information at multi-scale granularities.

To tackle the complexity of defect features and the frequent occlusion by pipelines, we enhance the perception extraction capability in the decoder. We use SFAC module to expand the perceptual field, thereby aggregating multi-scale information. Enhanced boundary information is then fused to ensure sharp segmentation results. Notably, we improve computational efficiency by utilizing depthwise separable convolutions in place of traditional convolutions in a specific structure, thereby enhancing the overall performance of the network.

### 3.1. Overview of existing related models

#### 3.1.1. Attention mechanism and transformer

In a self-attention, for given an input sequence $z \in \mathbb{R}^{N\times D}$, and with the provided mapping matrices $W_q$, $W_k$, $W_v \in \mathbb{R}^{D\times D_h}$, the matrices $q$(query), $k$(key), $v$(value) $\in \mathbb{R}^{N\times D_h}$ can be computed. We calculate the self-attention $SA(\cdot)$ of $z$, as shown in the formula:

$$[q,\ k,\ v] = z \cdot Wqkv$$

$$A = softmax\left(\frac{q \cdot k^T}{\sqrt{D_h}}\right), A \in \mathbb{R}^{N\times N} \tag{1}$$

$$SA(z) = A \cdot v$$

where $W_{qkv} \in \mathbb{R}^{D\times 3D_h}$ represents the mapping matrix. $D_h$ serves as the scaling factor for multi-head attention, which is typically set to $\frac{D}{h}$ based on the number of heads $h$. $N$ is the length of the embedding sequence.

It is typical to compose $h$ heads of self-attention in a multi-head attention. After concatenating the final vectors obtained from multiple heads, the resulting vector is mapped back to the original scale, represented as:

$$MSA(z) = [SA_1(z); SA_2(z); \cdots; SA_h(z)] \cdot W^0 \tag{2}$$

where $W^0 \in \mathbb{R}^{h\cdot D_h\times D}$, $SA_i(\cdot)$ represents the $i$th single-head self-attention, and $W^0$ is the concatenated mapping matrix. We obtain $MSA(z) \in \mathbb{R}^{N\times D}$.

Comprising self-attention as the core module, the Transformer encoder is structured with $L$ layers, as depicted in Fig. 2. The stacking of $L$ Transformer layers can be formulated as follows:

$$\begin{aligned} z'_l &= MSA\left(LN\left(z_{l-1}\right)\right) + z_{l-1} \\ z_l &= FNN\left(LN\left(z'_l\right)\right) + z'_l \end{aligned} \tag{3}$$

where $z_l$ represents sequence encoding of the $l$th layer Transformer, $MSA(\cdot)$ denotes multi-head attention, $LN(\cdot)$ represents layer normalization, $FNN(\cdot)$ represents the Feed-forward Neural Network. Unlike the standard Transformer encoder, we have made several improvements and discussed them in Section 3.3.

#### 3.1.2. Depthwise separable convolution

To reduce redundant parameter usage, depthwise separable convolution breaks down a standard convolution into a depthwise component and a pointwise component. Each feature channel corresponds to a unique kernel in depthwise convolution, ensuring that the convolution operates independently on this channel. Subsequently, with a $1 \times 1$ pointwise convolution, information from different channels is fused into a fixed number of channels, facilitating efficient information exchange and integration among different channels, and the utilization of parameters becomes more efficient Howard et al. (2017) and Chollet (2017).

### 3.2. Proposed network architecture

This paper proposes an enhanced method to the traditional encoder–decoder structure by incorporating both Transformer and convolution (Fig. 1). We aim to create a model better suited for tunnel scenes than general segmentation methods, as defect detection in tunnels is particularly challenging. In the encoding stage, we adopt a multi-scale framework constructed solely with a Transformer encoder as the backbone (Section 3.3). The Transformer, equipped with an efficient attention mechanism, effectively compensates for the limitations of local convolution, enhancing the receptive field globally and enabling a more comprehensive understanding of the scene.

During the network decoding stage, multi-scale information from the encoder is initially aggregated using an MLP Module (Section 3.4.1). Then, the SFAC module is implemented to enhance the perceptual range, facilitating the extraction of complete defect regions (Section 3.4.2). Following this, an effective attention module is used to enhance the weights of low-level features, focusing on extracting crucial boundary information (Section 3.4.3). Subsequently, a $3 \times 3$ convolution followed by a 4x bilinear upsampling process is employed to restore the feature map to its initial spatial resolution, producing the final segmentation output.

### 3.3. Encoder backbone based on hierarchical transformer

The Transformer backbone dramatically pushes the performance boundary of image segmentation. We implement a hierarchical Transformer with four stages as the encoder backbone. Each stage consists of an ***overlap patch embedding and merging*** and ***Transformer encoder*** (FNN head included), as illustrated in Fig. 3.

#### 3.3.1. Overlap patch embedding and merging

Instead of directly performing self-attention on global feature, we utilize overlapping patch embedding and merging modules to traverse overlapping patches and generate feature maps before calculating self-attention. This approach emphasizes the importance of local relationships while ensuring global relationship modeling, which is crucial for comprehensive perception of tunnel lining defects. Specifically, for stage $i$, when provided with the input feature map from stage $i-1$ of size $H_{i-1}\times W_{i-1}\times C_{i-1}$, we use convolution with appropriate padding to obtain an output feature map with dimensions $H_i\times W_i\times C_i$. We use stride $s_i$ and kernel size $k_i$ to control the overlap range and construct patches with overlapping relationships, maintaining local continuity around the feature blocks. The feature map's downsampling and merging are completed using convolution with stride $s_i > 1$. In general, for stage $i$, each patch consists of pixels with size $k_i \times k_i$; The output size of the feature map is $\frac{H_{i-1}}{s_i} \times \frac{W_{i-1}}{s_i}$; The embedding dimension can be easily derived from the output channel $C_i$ using convolution operations. The specific parameters for each stage are detailed in Table 1.

For simplicity, we adopt output stride (OS) as a metric, which signifies the proportion of the spatial resolution of the input image to the resolution of the final output, quantifying downsampling during the encoding process. After the encoding process, the original image with size $H\times W$ is embedded and merged into the feature map at stage $i$ with

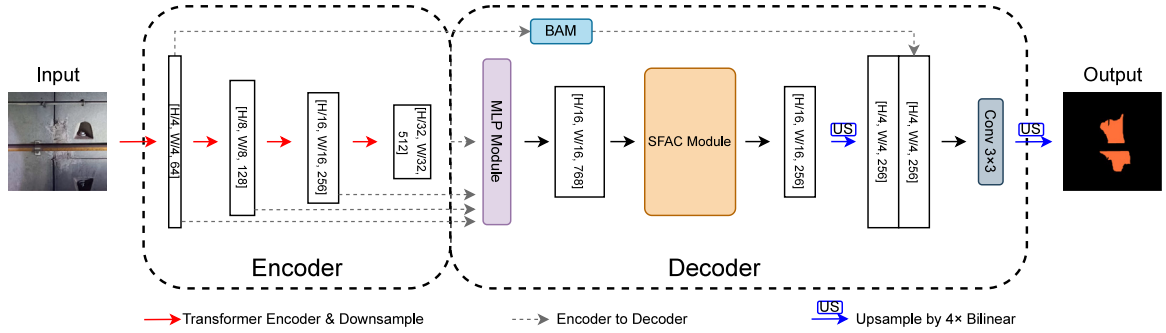



**Fig. 1.** Our proposed network architecture.

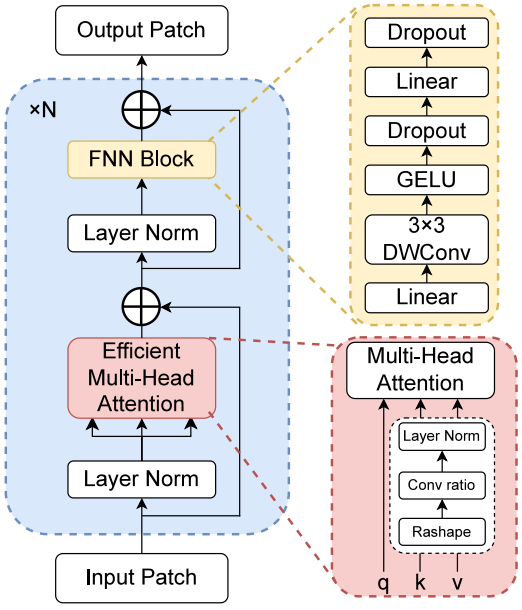



**Fig. 2.** Transformer encoder.

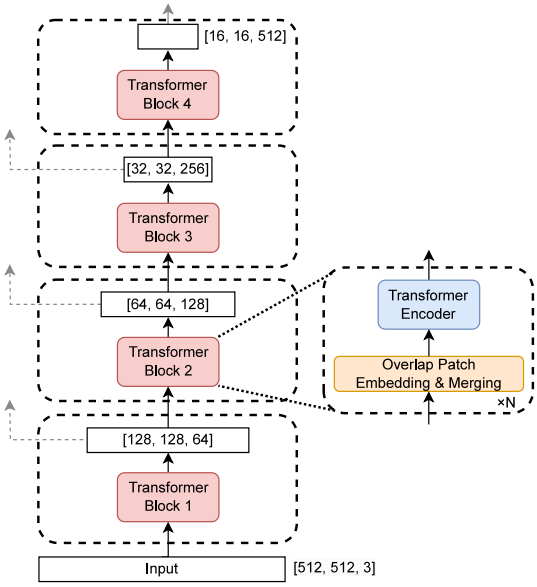



**Fig. 3.** Transformer backbone in encoder.

size $\left[\frac{H}{OS_i}, \frac{W}{OS_i}\right]$. Multi-scale feature maps are produced at resolutions of {1/4, 1/8, 1/16, 1/32} by considering $OS_i$, which represents the output stride for stage $i$ (see Table 1 for details).

### 3.3.2. Transformer encoder

**Efficient attention mechanism** to reduce computational costs. Directly using self-attention imposes a substantial computational burden, particularly when it comes to segmenting higher-resolution images (Ji et al., 2020). Therefore, the progressive SRA is implemented to ensure the model's efficiency (Wang et al., 2021). As shown in Fig. 1, this process involves selectively reducing the key and value matrices, mitigating the computational burden while maintaining the essential aspects of self-attention. Taking the key as an example, this can be represented as follows:

$$\begin{aligned} W_k' &= ReShape\left(\frac{N}{R},\ D_h \cdot R\right)(W_k) \\ \hat{W}_k &= Linear\left(D_h \cdot R,\ D_h\right)(W_k') \end{aligned} \tag{4}$$

where $R$ denotes the reduction ratio. $W_k \in \mathbb{R}^{N \times D_h}$ represents the original key matrix, $\hat{W}_k \in \mathbb{R}^{\frac{N}{R} \times D_h}$ signifies the reduced matrix, $ReShape$ refers to the operation of reshaping dimensions, and $Linear$ refers to the operation of dimension mapping. The computational complexity of the self-attention mechanism is decreased from $\mathcal{O}\left(N^2\right)$ to $\mathcal{O}\left(\frac{N^2}{R}\right)$.

In each encoding stage of the Transformer backbone, we employ the multi-head self-attention mechanism, which allows the input features to be mapped into different spaces for learning, thereby enhancing the network's fitting capability. The relevant parameters are specified and detailed in Table 1.

**FNN module** to preserve the non-linearity of global context information extraction, ensuring that the network can capture complex relationships. The FNN module can be represented as:

$$FNN(z) = MLP\left(GeLU\left(DWConv_{3\times3}\left(MLP(z)\right)\right)\right) + z \tag{5}$$

where MLP represents a multi-layer perceptron, GeLU is an activation function. In traditional image Transformers, the calculation retains the position information of patches and typically requires position embeddings, as seen in models like ViT and PVT (Dosovitskiy et al., 2020; Wang et al., 2021). However, position encoding involves interpolation when dealing with inconsistent feature sizes, which may introduce position bias. We excluded position encoding in the embedding stage. In contrast, we employ an FNN module with zero padding position encoding to incorporate positional information efficiently. Specifically, we incorporate positional information by adding a 3 × 3 depth-wise convolution with zero padding, offering an effective way to capture spatial relationships (Islam et al., 2019).

### 3.4. Effective decoder

Facing the challenges of multi-scale distribution of tunnel lining defects and background interference, it is crucial to refine features at multiple granularities after encoder extraction. Therefore, in our

**Table 1**
Parameters of transformer backbone in encoder.

| Stage | Overlap patch embedding & Merging | | | Efficient multi-head attention | | |
|---|---|---|---|---|---|---|
| | Overlap patch embedding (k, s, p) | Output Stride (OS) | Embedding dimension | Multi-head number | Depth | Reduction ratio |
| Stage 1 | (7, 4, 3) | 4 | 64 | 1 | 3 | 8 |
| Stage 2 | (3, 2, 1) | 8 | 128 | 2 | 4 | 4 |
| Stage 3 | (3, 2, 1) | 16 | 320 | 5 | 6 | 2 |
| Stage 4 | (3, 2, 1) | 32 | 512 | 8 | 3 | 1 |

decoder, we fully integrate the potential feature extraction information and employ SFAC to expand the receptive field to ensure comprehensive defect perception. Finally, we fuse these with shallow features that contain rich positional information, ensuring that defect edge details are captured.

### 3.4.1. MLP module for aggregating multi-scale features

The encoder adopts a hierarchical feature attention extraction structure, where shallow features emphasize fine-grained relationships, while deep features capture global contextual relationships. Therefore, a straightforward MLP module (as shown in Fig. 1) aggregates multi-scale information from the encoder. In the encoder, four feature maps are obtained through the Transformer backbone, whose height and width are downsampled 4, 8, 16, and 32 times, respectively, relative to the original size. These feature maps are first mapped to a unified channel dimension through the MLP layer to balance the contribution level of each feature map. In this module, our proposed network downsamples the feature map to 16 times smaller than the original image resolution, providing masks for the decoder. This operation achieves a balanced trade-off between segmentation quality and memory usage. The feature map downsampled by 32 times is upsampled to 16 times using bilinear upsampling, while the feature maps downsampled by 4 and 8 times are further downsampled to 16 times through $3 \times 3$ convolutions.

### 3.4.2. Stacking filters with atrous convolutions

Facing the complexity and multi-scale variability of lining defects, it is crucial to achieve detailed perception at each scale. This challenge can be effectively addressed by gradually expanding the receptive field through a pyramid structure. We employ SFAC module to enhance the perceptual range for complete defect region extraction. SFAC connects and fuses features extracted from the MLP module, achieving improved sensitivity to defect regions with multiple granularities. The essence of SFAC lies in constructing dilated convolutions with varying dilation rates, enabling the extraction of multi-scale target information to achieve sharp segmentation results.

Drawing inspiration from Xception, we initially utilize DSConv as a substitute for regular convolution to reduce computational costs (Howard et al., 2017). Furthermore, drawing inspiration from DenseASPP (Yang et al., 2018), we incorporate multiple $3 \times 3$ convolutions with dilation rates ranging from 3 to 24 and a $1 \times 1$ convolution with dilation rate 1, to form a parallel structure of stacking dilated convolution branches with dilation rates {1, 3, 6, 12, 18, 24}. Moreover, we introduce a pooling branch extracting global information to expand the receptive field further. Beyond parallel structure, we further enhanced it with a straightforward yet effective method: the outputs of the dilated convolution blocks, each with dilation rates of 1, 3, 6, 12, and 18, are aligned and stacked with the input of the next convolution branch through a concatenation module. This cascaded dilated convolution approach integrates perceptual receptive fields and enhances the ability of dense and wide spatial aggregation. Our designed SFAC module also addresses the issue of insufficient local feature extraction under low-light conditions. SFAC is depicted in Fig. 4.

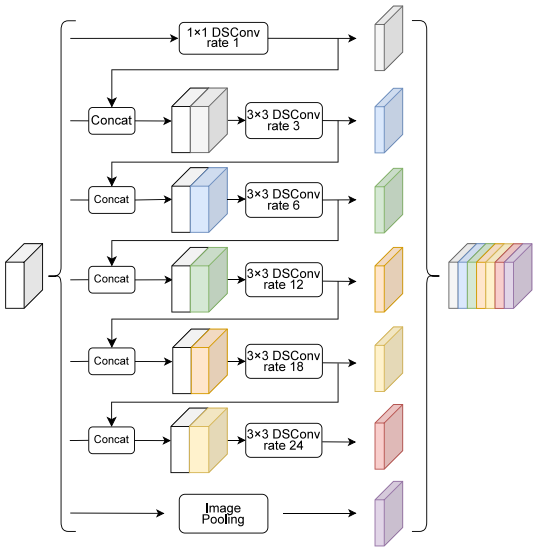



**Fig. 4.** Enhanced SFAC structure.

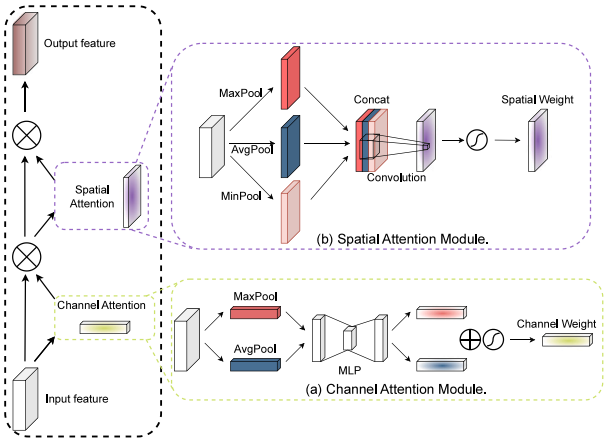



**Fig. 5.** Boundary-aware attention module in decoder.

### 3.4.3. Boundary-aware attention module

Inspired by the fact that shallow features retain more edge information, we establish a shortcut to allow shallow features to directly provide key edge features to the decoder, enhancing defect detection by preserving boundary details to address the blurred edge extraction caused by the dark environment of the tunnel. Meanwhile, we introduce BAM in the shortcut to achieve spatial and channel attention. BAM enhances the weight of shallow local target features and reduces the influence of occlusion noise, which will be proven to be highly effective for extracting tunnel lining defects in the experimental part.

In the BAM module (Fig. 5), the input features $F$ first undergo the channel-wise attention module. We first extract important channels with defect features using both average pooling and maximum pooling. Specifically, maximum pooling captures the most prominent information, while average pooling captures the average information, which is

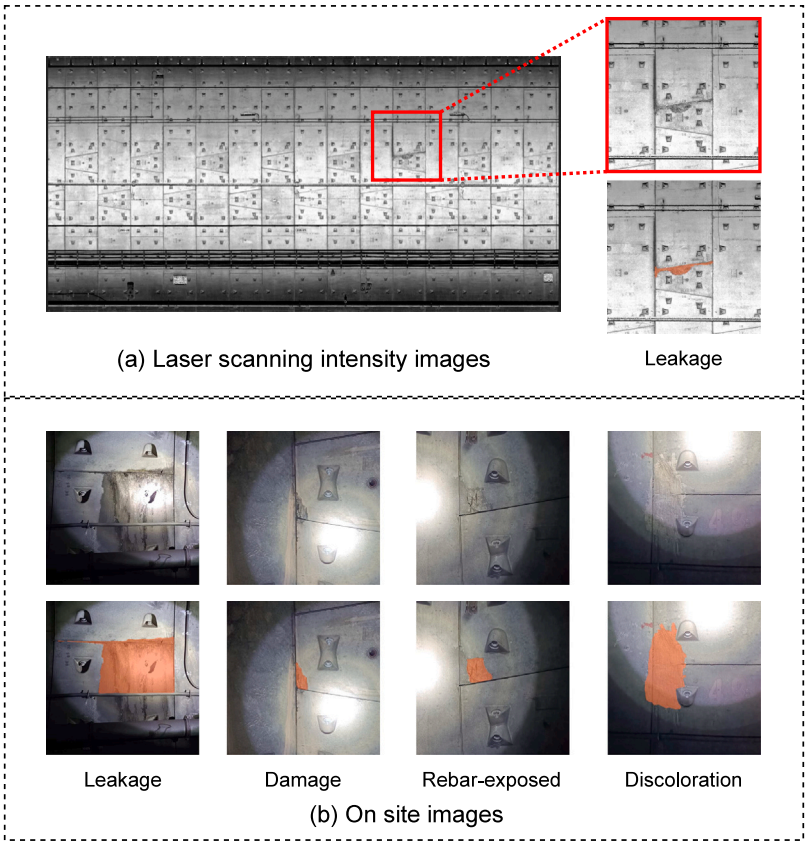



**Fig. 6.** Defect dataset of tunnel lining.

beneficial under uneven illumination conditions. The MLP module is added to improve the ability to distinguish nonlinearly separable data. Therefore, channel-wise attention can be represented as:

$$M_C(F) = \sigma(MLP(AvgPool(F))) + \sigma(MLP(MaxPool(F))) = \sigma\left(W_1\left(ReLU\left(W_0\left(F^c_{avg}\right)\right)\right) + W_1\left(ReLU\left(W_0\left(F^c_{max}\right)\right)\right)\right) \quad (6)$$

where $\sigma$ represents the sigmoid activation function, $AvgPool(\cdot)$ denotes global average pooling, $MaxPool(\cdot)$ represents global max pooling, $W_0 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_1 \in \mathbb{R}^{C\times\frac{C}{r}}$ are MLP weights. Here, $r$ serves as the channel reduction ratio to alleviate computational complexity.

The channel-enhanced features $F'$ are obtained by multiplying the output channel weights with the original input features. The feature map $F'$ is then enhanced with spatial-wise information. Specifically, we use max pooling, average pooling, and min pooling to obtain features containing defect information in spatial-wise distribution. Max pooling extracts the strongest response feature at each spatial location, which often corresponds to key defect features in our images. Average pooling computes the average response of all channels at each spatial location, providing global contextual information for that location. This helps the module understand the overall characteristics of a specific area, thereby highlighting the differences between the defects and background in the image. The addition of min pooling is particularly crucial as it captures the most subtle information caused by shadows or occlusions under low light conditions. This improvement will be shown to be effective in the less-than-ideal tunnel environment in the experimental part. Spatial-wise attention can be represented as:

$$M_S(F') = \sigma\left(F^{7\times7}\left(\left[AvgPool\left(F'\right); MaxPool\left(F'\right); MinPool\left(F'\right)\right]\right)\right) = \sigma\left(F^{7\times7}\left(\left[F^s_{avg}; F^s_{max}; F^s_{min}\right]\right)\right) \quad (7)$$

where $F^{7\times7}$ denotes the $7 \times 7$ convolution.

The overall representation of BAM can be expressed as follows:

$$BAM(F) = M_S \bigotimes F' = M_S \bigotimes M_C \bigotimes F \quad (8)$$

where $F$ is the input features, with $M_C$ representing the channel attention weight and $M_S$ as the spatial attention weight, and $\bigotimes$ as the element-wise multiplication operation.

## 4. Experiment

### 4.1. Datasets

The experiment utilized an RGB image dataset collected by on-site inspection provided by the Guangzhou Metro, and a grayscale image dataset generated from LiDAR scanning intensities. The datasets were collected within subway tunnels which are affected by groundwater and ground settlement factors. The defects caused by such factors may result in cracking and lining detachment issues. The dataset collection environment in subway tunnels is characterized by insufficient or uneven ambient light, the obstructed view occlusion by pipelines or traffic facilities. Therefore, the collected images are of low quality, with blurry details and dim colors. We manually annotated the ground truth of the defects for supervising network training. In addition, we also used additional RGB images collected on site as testing data. The regions where defects appeared were uniformly cropped to a size of $617 \times 617$, which then was processed to be $512 \times 512$ through augmentation for network input. Annotation software was used to mark areas of defects, including but not limited to leakage, damage, exposed rebar, and discoloration regions, as illustrated in Fig. 6.

Data augmentation encompassing both physical and color changes is used to increase the diversity of input shapes, as outlined in Table 2. The initial training dataset's image count expanded from 109 to 730

**Table 2**
Dataset augmentation.

| | Physical change | | | | | | |
|---|---|---|---|---|---|---|---|
| | Rotate | Flip | Zoom | Distortion | Shear | Crop | Scale |
| Probability (%) | 80 | 50 | 30 | 80 | 100 | 100 | 100 |
| Range | ±25° | – | ±15% | ±10 | 10% | 80% | ±1.3 |
| | Color change | | | | | | |
| | Brightness | | | Discolor | | Contrast | |
| Probability (%) | 100 | | | 100 | | 100 | |
| Range | ±50% | | | ±100% | | ±50% | |

through augmentation. Subsequently, we divide the dataset into training and validation sets in a ratio of 8:2, including 584 training images and 146 validation images.

### 4.2. Loss function

The cross-entropy loss is frequently employed to exploit the maximum probability of the predicted values:

$$\ell_{ce} = -\sum_{i=1}^{n} p\left(x_i\right) \times ln\left(q\left(x_i\right)\right) \tag{9}$$

where $x_i$ represents the $i$th sample, $n$ represents the total number of classes, $p(\cdot)$ denotes the true probability, and $q(\cdot)$ denotes the predicted probability. With an unbalanced distribution of tunnel lining defects in images, the negative samples (background) account for the majority. Merely using cross-entropy loss function tends to favor negative classes, which is not conducive to the network's ability to detect diseases in tunnels.

The Dice function is a set similarity metric commonly utilized to weigh the similarity between two samples. Its value falls within the range of 0 to 1, with a score closer to 1 denoting the two samples are more comparable. Adoption of the Dice loss function helps the model concentrate more on foreground targets:

$$\ell_{Dice} = 1 - \frac{2\left|y \cap y'\right|}{|y| + |y'|} \tag{10}$$

where $y$ denotes the ground truth value, $y'$ represents the predicted value, and $|\cdot|$ denotes the count of values.

Incorporating a weighted combination to balance the contributions from positive and negative samples, the adopted final composite loss function is represented as follows:

$$\ell_{all} = \alpha \ell_{ce} + \beta \ell_{Dice} \tag{11}$$

where $\alpha$ and $\beta$ are balancing weights.

### 4.3. Evaluation metrics

Image segmentation involves the classification of all pixels, assigning them different label values. In our study, we assess the effectiveness of our segmentation approach using two key evaluation metrics: accuracy (Acc) and mean Intersection over Union (mIoU).

1. Acc measures the proportion of correctly classified pixels in the entire image, which can be expressed as:
$$I_{acc} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{FN} + N_{TN}} \tag{12}$$
where $N_{TP}$ as the number of true positives, $N_{FP}$ as the number of false positives, $N_{FN}$ as the number of false negatives, and $N_{TN}$ as the number of true negative.
2. mIoU denotes the mean Intersection over Union (IoU) between the predicted values and the ground truth. The mIoU value falls within the range of 0 to 1, with a score nearer to 1 signifying superior segmentation performance. The formula can be expressed as follows:
$$mIoU = \frac{1}{n+1} \sum_{i=0}^{n} \frac{p_{ii}}{\sum_{j=0}^{n} p_{ij} + \sum_{j=0}^{n} p_{ji} - p_{ii}} \tag{13}$$
where $n$ represents the number of classes, $i$ as the ground truth, $j$ represents the predicted value, $p_{ii}$ represents true positives, $p_{ji}$ represents false positives, and $p_{ij}$ indicates false negatives.

### 4.4. Setup and implement details

Training efficient deep-learning models with Transformer involves considerable computational and memory expenses. We conducted our model's training on a workstation equipped with two Nvidia 3090 GPUs. Parallel computing was leveraged for the distribution of training examples across multiple GPUs. The Adam optimizer with momentum was employed to expedite convergence. The base models were trained for 100 epochs.

The momentum and weight decay of the Adam optimizer were configured to be 0.9 and $5\times10^{-4}$, respectively. We employed a polynomial exponential decay, starting with an initial learning rate of 0.007. SyncBatchNorm was implemented for parallel training across multiple GPUs, with a batch size of 16. The experimental procedure involved pretraining the backbone on the ImageNet-1K dataset to speed up the training convergence. After sufficient pre-training, the Transformer architecture has richer background knowledge, enabling it to perform effectively with minimal preprocessing of the target low-light images.

## 5. Experimental results

### 5.1. Results and compare to state-of-the-art models

To assess the effectiveness of our proposed method in detecting tunnel lining defects, a series of experiments were conducted. The performance of our method was evaluated through comparisons with state-of-the-art models, employing evaluate metrics such as mIoU and Acc scores. Several sample methods, i.e., FCN, U-Net, SegNet, PSPNet, DeepLabv3, DeepLabv3+, DenseASPP, DANet, SegFormer, Swin-Unet, and BiFormer, were chosen as reliable baselines. Table 3 illustrates our proposed method outperformed the comparison methods in detecting defects in tunnel linings. Among the baseline methods, DANet with ResNet101 achieved the highest mIoU and Acc scores among the baselines before the introduction of our proposed method. Our proposed method demonstrates improved segmentation performance, with a 0.18% increase in mIoU and a 0.05% increase in Acc over DANet with ResNet101. Moreover, compared with DANet using the high-stacked backbone ResNet101, our proposed method achieves a parameter reduction of 20.92M and a computational reduction of 238.63 GFLOPs, indicating a more efficient model. The DANet with ResNet101 needs more dense feature sampling, which results in the neglect of edge information and leads to poor defect segmentation integrity. Our proposed model addresses this limitation by preserving edge information, achieving a mIoU of 78.14%.

Our proposed method proves to be highly effective and efficient in detecting defects in tunnel linings, as demonstrated by its superior performance in various evaluation metrics. In comparison to models with a similar parameter count, such as PSPNet, DeepLabv3 with ResNet50, DeepLabv3+ with MobileNet-V2, and SegFomer. Our proposed method exhibits a significant improvement in detecting defects in tunnel linings, achieving a minimum 4.4% increase in mIoU. Notably, the DeepLabv3+ model with MobileNet exhibits a parameter count comparable to that of DeepLabv3 with ResNet50. However, it achieves a remarkable 7.64% increase in mIoU. This highlights the significant performance improvement achieved using a well-designed decoder within DeepLabv3+ to refine multi-scale information fully.

**Table 3**
Comparison with state-of-the-art methods.

| Model architecture | GFlops | Params (M) | Evaluation | | |
|---|---|---|---|---|---|
| | | | F1 (%) | Acc (%) | mIoU (%) |
| U-Net (Ronneberger et al., 2015) | 124.50 | 13.9 | 47.16 | 71.36 | 63.64 |
| FCN (VGG-16) (Long et al., 2017) | 148.53 | 35.31 | 49.17 | 90.91 | 66.03 |
| SegNet (VGG-16) (Badrinarayanan et al., 2017) | 160.68 | 29.44 | 65.64 | 91.53 | 70.32 |
| PSPNet (ResNet50) (Zhao et al., 2017) | 59.21 | 46.71 | 68.35 | 93.86 | 72.06 |
| DeepLabv3 (ResNet50) (Chen et al., 2017b) | 173.79 | 41.99 | 50.38 | 91.10 | 66.10 |
| DeepLabv3+ (MobileNet-V2) (Chen et al., 2018) | 33.75 | 41.83 | 71.14 | 92.66 | 73.74 |
| DeepLabv3+ (ResNet101) (Chen et al., 2018) | 82.88 | 74.87 | 73.93 | 94.25 | 77.69 |
| DenseASPP (densenet121) (Yang et al., 2018) | 66.32 | 18.19 | 71.02 | 93.90 | 75.12 |
| DANet (ResNet101) (Fu et al., 2019) | 296.85 | 68.61 | 74.16 | 94.35 | 77.96 |
| SegFormer (Xie et al., 2021b) | 71.36 | 47.22 | 69.89 | 92.97 | 72.44 |
| Swin-Unet (Cao et al., 2022) | 5.92 | 27.17 | 70.72 | 93.19 | 73.65 |
| Bi-Former (Zhu et al., 2023) | 91.10 | 56.81 | 66.95 | 92.61 | 71.17 |
| ViT-CoMer (UperNet) (Xia et al., 2024) | 1194.16 | 60.50 | 55.42 | 90.60 | 69.37 |
| Our proposed method | 58.22 | 47.69 | **75.01** | **94.40** | **78.14** |

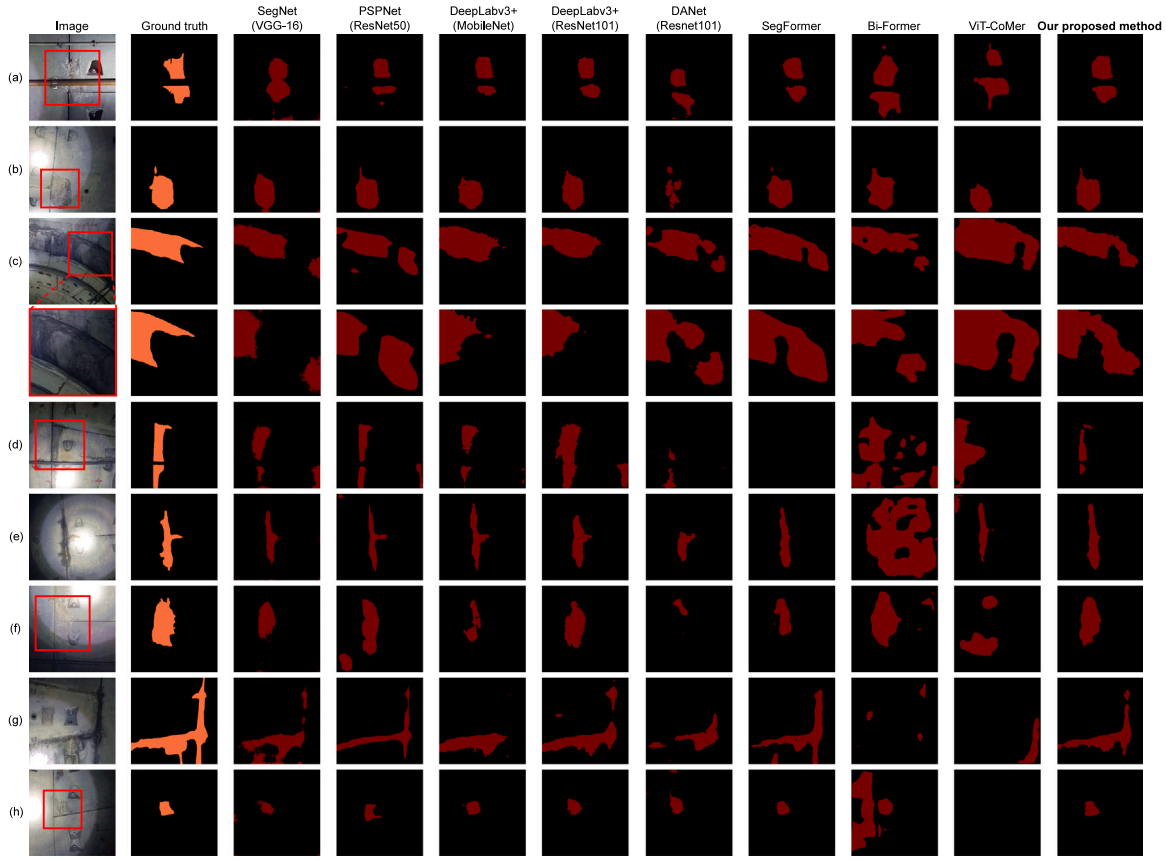



**Fig. 7.** Defect detection results compared to state-of-the-art models.

We can come to the same conclusion by making a comparison with SegFormer. Through the meticulous design of the decoder, aiming to fully utilize and refine both local fine-grained and global contextual information, our proposed method outperforms the comparison methods in terms of defect detection. This design choice enhances the model's capability to capture intricate details and overall context, contributing to its enhanced performance in defect detection.

Early general segmentation networks performed poorly for extracting defects in tunnel linings. U-Net exhibits the smallest parameter count as it lacks a deep backbone for feature extraction, which consequently results in lower segmentation accuracy. This limitation arises from the network's shallow architecture, which hinders its ability to capture complex features and details necessary for accurate defect segmentation. FCN and SegNet exhibit limitations in effectively extracting significant defects, primarily attributed to their shallow network stacking depth. Their segmentation accuracy did not achieve satisfactory scores during validation, indicating their challenges in accurately delineating defects in tunnel linings. PSPNet demonstrates improved precision compared to SegNet; however, the parameter count increases accordingly due to the utilization of ResNet-50 as a backbone. This trade-off between precision and model complexity is important in choosing segmentation architectures.

The results of the detected defects are visualized in Fig. 7, where specific areas demonstrating typical differences in segmentation performance are highlighted to illustrate the details of the detected defects. Meanwhile, we intentionally select defect images captured in complex environments with low or uneven lighting conditions to illustrate the robustness of our proposed approach. These scenarios highlight the ability of the model to perform effectively under challenging conditions. PSPNet and SegNet exhibit regions of missegmentation due to

**Table 4**
Ablation study.

| Model | MLP | SFAC | BAM | Evaluation | |
|---|---|---|---|---|---|
| | | | | Acc (%) | mIoU (%) |
| 1 | – | – | – | 93.32 | 75.16 |
| 2 | ✓ | – | – | 93.32 | 76.42 |
| 3 | – | ✓ | – | 93.91 | 76.14 |
| 4 | – | – | ✓ | 93.95 | 77.82 |
| 5 | ✓ | ✓ | – | 93.67 | 77.37 |
| 6 | ✓ | ✓ | ✓ | **94.40** | **78.14** |

their insufficient network depth, making them struggle to cope with complex environments, as illustrated in Fig. 7(a), (c), (d), and (f). SegNet displays unclear edge segmentation with numerous scattered segmentation points, as highlighted in Fig. 7(g). While maintaining complete edges, PSPNet tends to exhibit excessive false positives. BiFormer struggles to cope effectively with scenes exhibiting significant changes in lighting, and SegFormer faces difficulties in extracting fine-grained information, because it lacks a well-designed decoder, as evident in Fig. 7(d), (e), and (h). ViT-CoMer clearly shows incorrect defect localization and struggles with fine-grained requirements, as illustrated in Fig. 7(g). DeepLabv3+ with MobileNet and DeepLabv3+ with ResNet101 exhibit comparable performance. Although the backbone of ResNet101 can extract more representative features than that of MobileNet, the Atrous Spatial Pyramid Pooling (ASPP) module used in DeepLabv3+, with its relatively independent receptive field, faces challenges in expanding the perception of multi-scale target defect areas. DeepLabv3+ with MobileNet demonstrates strong segmentation capabilities in Fig. 7(a) and (c) but exhibits incomplete segmentation areas in Fig. 7(d), (f), and (g). DeepLabv3+ with ResNet101 achieves the most accurate segmentation in Fig. 7(g) but still shows some false positives.

Our proposed network leverages the powerful Transformer backbone to meticulously explore complete information at multiple granularities, thereby enhancing the comprehensive extraction of defect regions. Simultaneously, the reinforced SFAC module ensures the integrity of defect areas by establishing reliable boundaries. As illustrated in Fig. 7(a), (b), (f), and (h), the segmentation areas generated by our proposed network are the most complete, and the edges are sharply defined. While the performance may not be as satisfactory as other baseline methods in Fig. 7(d), it is notable that our proposed method still identifies clear traces of attempted segmentation in locations obscured by pipes. Fig. 7(c) appears to exhibit missegmentation, where certain details may not have been accurately delineated. Some defects were neglected during the manual annotation, contributing to potential discrepancies in the evaluation. Therefore, the proposed network effectively mitigates annotation errors, providing representations that are more likely to reflect the true positions of defects. In most cases, our segmentation accurately captures the shape of defects completely (Fig. 7(g)). To summarize, the proposed network establishes a new state-of-the-art performance in tunnel lining defect segmentation tasks, showcasing its effectiveness in capturing comprehensive defect information.

### 5.2. Ablation study

In this experiment, we introduced ablation studies using the MLP module (Section 3.4.1) and the SFAC module (Section 3.4.2) as ablation units. The goal is to investigate whether the inclusion of these modules contributes to the improvement of the network. Subsequent validation experiments were carried out by integrating the BAM to investigate whether boundary-aware information can effectively enhance defect detection.

From Table 4, it is clearly shown that Model 1, without the multi-scale information aggregation through the MLP module and the strengthened SFAC, exhibited the lowest segmentation accuracy, with a 3% decrease in mIoU compared to Model 6. Model 2, which exclusively integrates the MLP module, exhibited a 1.3% increase in mIoU. Simultaneously, Model 3, incorporating solely the SFAC, demonstrated a 1% increase in mIoU compared to Model 1. The effectiveness of each proposed module in the network is evident. The BAM proves to be more impactful than the other two modules. Model 4, incorporating only BAM, slightly improved the accuracy compared to the combination of MLP and SFAC modules in Model 5. Combining all three proposed modules in the network achieved the highest mIoU score of 78.14% (Model 6). It can be seen that each proposed module is indispensable in the network.

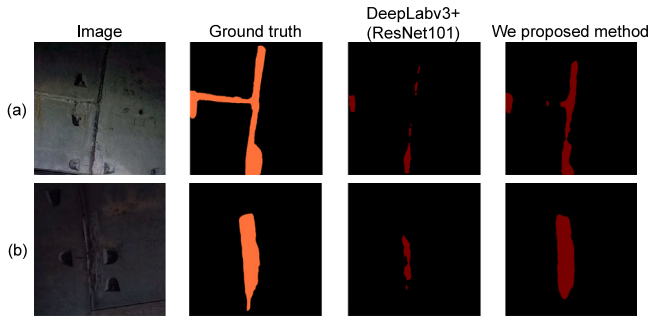



**Fig. 8.** Ours strengths.

### 5.3. Strengths and weaknesses

To thoroughly illustrate the strengths and weaknesses of our proposed network, we deliberately selected challenging scenarios to differentiate the models' segmentation capabilities, as depicted in Fig. 8. We specifically chose to compare our results with DeepLabV3+ (ResNet101), a model that achieves promising segmentation performance, as indicated in Table 3.

Our proposed network successfully segments the areas of defects in Fig. 8(a), even under challenging non-uniform illumination conditions. Furthermore, leveraging the complete information extraction and aggregation mechanisms, our model accurately segments the defect areas without any breaks, a task incorrectly performed by DeepLabV3+ (ResNet101). Furthermore, due to the enhanced receptive field coverage in our decoder, our model can fully extract defect areas even in extremely dark scenes, as demonstrated in Fig. 8(b).

Our proposed method may encounter challenges and fail when dealing with specific extreme environments. In the scenario depicted in Fig. 9(a), our proposed model encounters difficulty in effectively addressing the intricate challenge posed by the coexistence of pipe occlusion and shadow interference, owing to the visual similarity between defects and shadows. Our model exhibits shortcomings in scenarios where all granularity information needs to be preserved, as illustrated in Fig. 9(b), where the regions extracted by our model are notably smaller than the ground truth.

## 6. Analysis

Various experiments were conducted in this section to investigate the contributions of the suggested Transformer backbone, the enhanced SFAC, the BAM and the weight assignment in the composite loss function.

**Table 5**
Comparison with different backbones.

| Backbone | GFlops | Params (M) | Evaluation | | |
|---|---|---|---|---|---|
| | | | F1 (%) | Acc (%) | mIoU (%) |
| ResNeXt-101(32 × 8d) (Xie et al., 2017) | 363.59 | 86.74 | 75.98 | 94.57 | 78.72 |
| Xception (Chen et al., 2018) | 46.67 | 37.87 | 32.63 | 88.51 | 55.33 |
| MobileNet-V2 (Woo et al., 2018) | 10.69 | 15.4 | 71.37 | 92.66 | 73.74 |
| Swin Transformer-T (Liu et al., 2021) | 24.76 | 27.52 | 73.14 | 93.62 | 74.70 |
| Swin Transformer-L (Liu et al., 2021) | 191.45 | 194.90 | 74.05 | 94.13 | 76.81 |
| Pyramid Vision Transformer-S (Wang et al., 2021) | 22.83 | 27.12 | 68.17 | 92.69 | 71.90 |
| Pyramid Vision Transformer-L (Wang et al., 2021) | 53.37 | 64.01 | 73.94 | 94.03 | 76.52 |
| Swin TransformerV2-T (Liu et al., 2022b) | 24.76 | 27.58 | 46.48 | 89.97 | 64.00 |
| Pyramid Vision TransformerV2-B2(Wang et al., 2022) | 24.00 | 28.39 | 70.97 | 93.36 | 73.89 |
| Pyramid Vision TransformerV2-B4 (Wang et al., 2022) | 54.93 | 65.58 | 95.00 | 94.34 | 76.91 |
| Efficient ViT-L2 (Cai et al., 2022) | 27.02 | 38.07 | 48.79 | 90.51 | 65.25 |
| ConvNeXtV2-B (Woo et al., 2023) | 80.22 | 88.72 | 44.95 | 89.56 | 60.35 |
| ViT-CoMer-S (Xia et al., 2024) | 79.29 | 31.19 | 73.36 | 93.84 | 75.62 |
| ViT-CoMer-B (Xia et al., 2024) | 283.25 | 117.51 | 74.98 | 94.42 | 78.20 |
| Proposed Transformer | 28.64 | 24.2 | 75.01 | 94.4 | 78.14 |

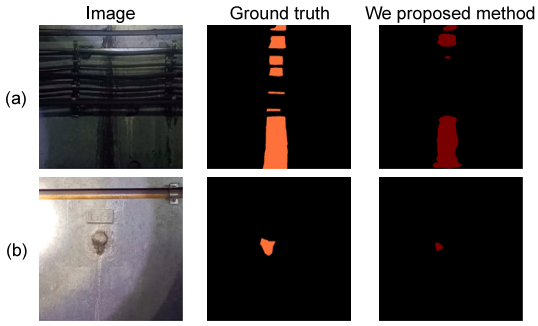



**Fig. 9.** Ours weaknesses.

### 6.1. Impact of transformer backbone

To evaluate the performance and computational complexity of our proposed Transformer backbone, we conducted experiments with many state-of-the-art backbones, including Xception, MobileNet-V2, ResNet-101, ResNeXt-101, Swin-Transformer, Pyramid Vision Transformer (PVT), Efficient ViT, ConvNeXt and ViT-CoMer. The goal is to compare the parameter amount and segmentation accuracy. In this experiment, we used the default parameter settings for each compared backbone by referencing published articles or utilizing the provided codes. For ResNeXt-101, we employed 32 kernel groups in the bottleneck structure convolution, each containing 8 depth-wise separable convolution kernels. Other remaining models are configured according to their official code libraries. The results of the experiments are presented in Table 5.

Compared to MobileNet, with the smallest number of parameters, our proposed Transformer exhibits a 4.4% improvement in mIoU. Benefiting from the robust design of multi-scale feature output, our proposed Transformer not only achieves an improvement of 3.44% in mIoU but also decreases the parameter number compared to the single-output backbone of Swin-Transformer. Compared to the powerful and bulky ResNeXt-101, which has 3.6 times the number of parameters than our proposed backbone, our proposed Transformer exhibits a slight 0.6% reduction in mIoU. The number of parameters is significantly reduced, which benefits network optimization. Due to the more efficient integration of positional information in the FNN, our proposed method achieves an improvement of approximately 4% in mIoU compared to the PVT (B2) model with similar computational complexity. Furthermore, compared to the more complex PVT (B4) model, our approach reduces GFLOPs by approximately 48% while achieving higher F1 score. Compared to the state-of-the-art model ViT-CoMer, our model demonstrates superior performance in terms of F1 score, although it lags slightly behind in terms of mIoU. However, it is noteworthy that our model reduces GLOPs by approximately 90% and model size by 80%. Hence, our proposed Transformer backbone stands out as the most cost-effective choice.

### 6.2. Impact of SFAC

Experiments were conducted to evaluate the impact of the SFAC on accuracy and to analyze its computational cost by calculating the number of parameters. Our SFAC model (Model A) incorporates multiple 3 × 3 convolutions with dilation rates ranging from 3 to 24, a 1 × 1 convolution with dilation rate 1, and a pooling branch, forming a series-parallel structure. This design aims to enhance the perceptual range for the complete extraction of defect regions. We verified the efficacy of our proposed structure by examining two simplified versions, namely Model B and Model C. Model B includes a merely parallel structure with multiple 3 × 3 convolutions having dilation rates ranging from 3 to 24. Model C adopts a series-parallel structure by concatenating convolutions with dilation rates ranging from 6 to 18. Furthermore, we compared our proposed structure with the state-of-the-art Atrous Spatial Pyramid Pooling (ASPP) structure (Chen et al., 2018) (Model D) to demonstrate our advancements.

Table 6 reveals a noticeable advancement when comparing the SFAC module with the existing ASPP. Our SFAC module achieved a 2% increase in mIoU and a 0.69% increase in Acc than ASPP, reaching an mIoU of 78.14%. Furthermore, simplifying some convolution branches and the concatenation module (Model B and Model C) demonstrated a notable decrease in mIoU compared to the SFAC. In addition, we compared the receptive fields in the SFAC module with ASPP and DenseASPP. As shown in Fig. 10, our SFAC module not only retains the central perception focus through the final concatenation but also expands the receptive field to cover a larger area. This enables the network to fully perceive the potential features. Consequently, the conclusion can be drawn that combining the proposed convolution branches for enhancing the perceptual range leads to superior segmentation outcomes.

In the computational cost analyzing experiment for the SFAC module, we compared standard convolution with depth-wise separable convolution by counting the number of parameters. The results of this comparison are presented in Table 7.

### 6.3. Impact of BAM

We conducted two experiments to demonstrate the effectiveness of the BAM module in occlusion or shadow situations. The first experiment tested the impact of adding min pooling to the BAM module, as shown in Table 8. We also visualized the attention distribution in some evaluation cases with occlusion, as illustrated in Fig. 11. The second

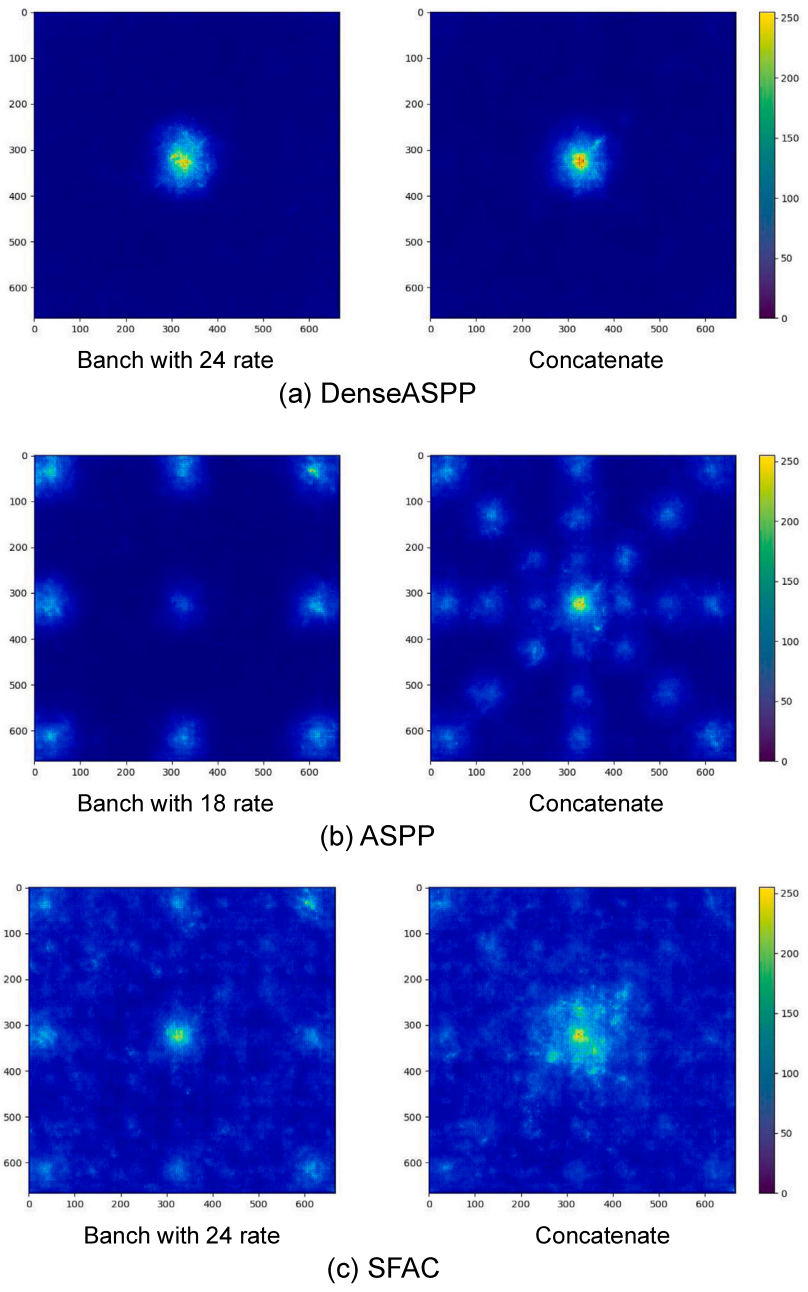



**Fig. 10.** Comparison of receptive fields. The weight of attention is represented by a color bar. The closer the color is to yellow, the higher the attention weight. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 6**
Comparison with different variants of SFAC.

| Experiment | GFlops | Params (M) | Evaluation | |
|---|---|---|---|---|
| | | | Acc (%) | mIoU (%) |
| Model A | 29.93 | 6.18 | **94.40** | **78.14** |
| Model B | 13.62 | 3.71 | 94.09 | 76.86 |
| Model C | 19.04 | 4.24 | 94.02 | 76.73 |
| Model D | 9.25 | 2.65 | 93.71 | 76.01 |

**Table 7**
Comparison of SFAC parameters using different convolutions.

| Convolution used in SFAC | GFlops | Params (M) |
|---|---|---|
| Standard convolution | 158.49 | 37.49 |
| Depth-Wise separable convolution | **29.93** | **6.18** |

**Table 8**
Comparison with different spatial attention structures of BAM.

| Spatial structure | Evaluation | |
|---|---|---|
| | Acc (%) | mIoU (%) |
| Max&Avg&Min | **94.40** | **78.14** |
| Max&Avg | 94.26 | 77.21 |

experiment focused on the size of the convolution kernel used in the BAM module, which is closely related to its efficiency, as shown in Table 9.

As shown in Table 8, it is particularly effective to enhance the occlusion or shadow information in the spatial-wise dimension by incorporating additional min pooling. As demonstrated in Fig. 11, our

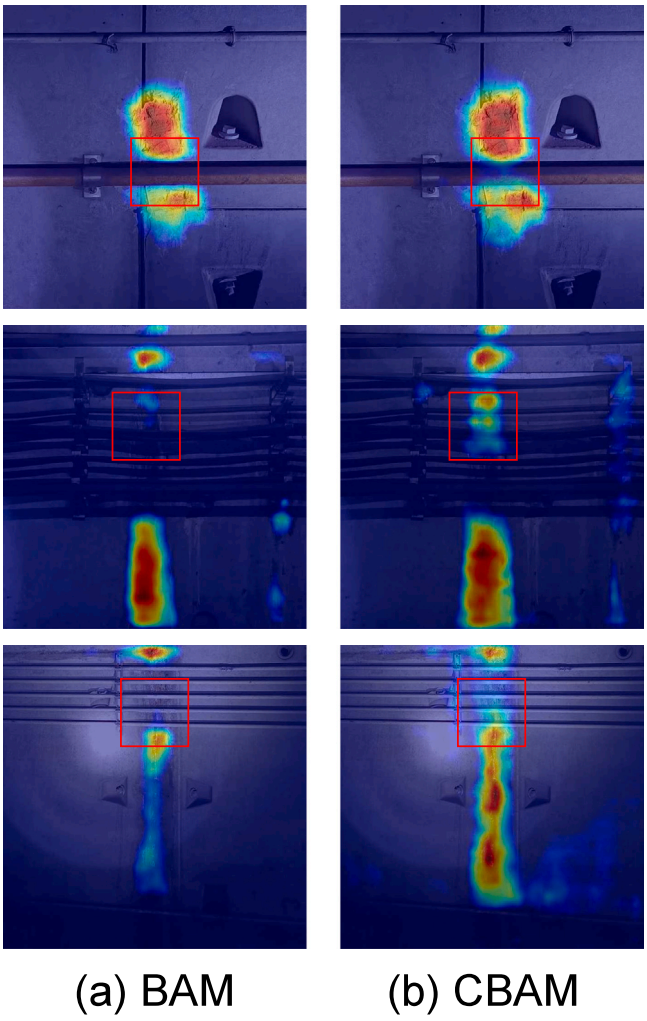



**Fig. 11.** BAM vs. CBAM. Attention is represented in the form of a heat map, where red indicates areas of high attention.

**Table 9**
Comparison with different convolutional kernel size of BAM.

| Kernel size | MFlops | Params (M) | Evaluation | |
|---|---|---|---|---|
| | | | Acc (%) | mIoU (%) |
| 9 × 9 | 584 | 16.19 | 93.97 | 75.55 |
| **7 × 7** | 488 | 9.89 | **94.40** | **78.14** |
| 5 × 5 | 416 | 5.18 | 94.11 | 76.82 |
| 3 × 3 | 368 | 2.03 | 93.46 | 75.78 |

proposed module can avoid the influence of pipeline occlusion when the reflectivity of the pipeline is usually lower than that of the tunnel lining. Therefore, our proposed BAM has proven to be more effective than CBAM in dealing with pipeline interference. In addition, as shown in Table 9, we demonstrate that the selected convolution kernel size can efficiently aggregate the information obtained from maximum pooling, average pooling and minimum pooling.

### 6.4. Impact of composite loss function

To quantify the disparity between the network predictions and ground truth values, we utilized a composite loss function that combines the cross-entropy and Dice loss functions. This experimental investigation aimed to determine the optimal weights for the composite loss function. We compared two weight combinations represented by

**Table 10**
Accuracy comparison based on different loss functions.

| Loss function | | Evaluation | |
|---|---|---|---|
| $\alpha$ | $\beta$ | Acc (%) | mIoU (%) |
| 1 | 0 | 93.85 | 75.94 |
| 0.8 | 0.2 | 93.74 | 76.39 |
| 0.6 | 0.4 | 93.81 | 77.00 |
| 0.5 | 0.5 | **94.40** | **78.14** |
| 0.4 | 0.6 | 93.96 | 77.69 |
| 0.2 | 0.8 | 94.06 | 77.74 |
| 0 | 1 | 93.44 | 75.38 |

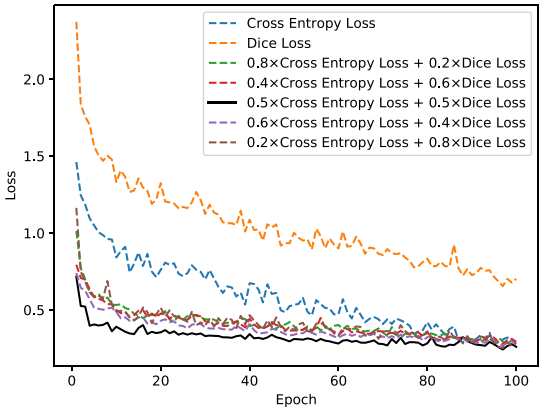



**Fig. 12.** Training progresses with different composite loss functions.

$\alpha$ and $\beta$ to identify the configurations that would better facilitate the training of the network, resulting in more accurate predictions.

As depicted in Table 10, incorporating the Dice loss alleviates the adverse effects of imbalanced training samples, particularly those with a substantial number of background samples. This operation leads to a notable increase of 2.43% in mIoU compared to using only the cross-entropy loss. Examining the training progress illustrated in Fig. 12, it is evident that the composite loss function with $\alpha = 0.5$ and $\beta = 0.5$ yielded the optimal optimization results and demonstrated the fastest convergence speed when compared to other weights.

## 7. Conclusion

This paper introduces a deep network model built an encoder–decoder structure that combines Transformer and convolutional elements to comprehensively address the extraction of tunnel lining defects. The proposed model employs a pure Transformer backbone as the encoder, leveraging the attention mechanism's powerful context information extraction capability. Our proposed method enables effective capturing and outputting of multi-layer features containing both fine-grained and global contextual information. An MLP module is implemented in the decoder to aggregate multi-scale information from the encoder. Subsequently, the stacking filters with atrous convolutions module further enhance the perception of interpreting information about defects at different granularities. Additionally, we introduced a boundary-aware attention module in the decoder to fully extract defect boundaries by effectively suppressing noise. Furthermore, the utilization of progressive spatial reduction attention and depth-wise separable convolution reduces the model's parameters, improving computational efficiency.

The proposed network has demonstrated state-of-the-art performance in experiments focused on detecting tunnel lining defects using on-site captured images. The dataset utilized in this study comprises typical lining defect images captured in complex tunnel environments.

Comparative experimental results emphasize that our proposed network achieves more accurate and complete segmentation compared to baseline models, including U-Net, DeepLabv3, and Swin-Transformer. The achieved mIoU for our algorithm is 78.14%, showcasing a significant improvement compared to DeepLabv3+ with ResNet101. This underscores that careful consideration of fine-grained and contextual information enhances segmentation performance in detecting tunnel lining defects.

Due to the high cost associated with data annotation, we encounter significant challenges stemming from the limited coverage of our dataset. In future work, our goal is to extend the tunnel lining defect detection network into a multi-category defect segmentation network. This expansion holds the potential to become a robust architecture for ensuring more accurate defect predictions. To effectively handle this difficulty, it is crucial to annotate a large number of instances of defects accurately. Additionally, incorporating few-shot learning techniques, especially in the context of imbalanced samples, may contribute to constructing a more robust architecture.

## CRediT authorship contribution statement

**Bo Guo:** Writing – review & editing, Methodology. **Zhihai Huang:** Writing – original draft. **Haitao Luo:** Data curation. **Perpetual Hope Akwensi:** Visualization. **Ruisheng Wang:** Project administration. **Bo Huang:** Writing – review & editing. **Tsz Nam Chan:** Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. U21A20139).

## Data availability

The authors do not have permission to share data.

## References

Afshani, A., Kawakami, K., Konishi, S., Akagi, H., 2019. Study of infrared thermal application for detecting defects within tunnel lining. Tunn. Undergr. Space Technol. 86, 186–197. http://dx.doi.org/10.1016/j.tust.2019.01.013.

Agrawal, A., Kundu, S., Ahmad, T., Bhatt, M., 2024. ReLAP-Net: Residual learning and attention based parallel network for hyperspectral and multispectral image fusion. Photogramm. Eng. Remote Sens. 90 (7), 395–403. http://dx.doi.org/10.14358/PERS.24-00003R2.

Alidoost, F., Hahn, M., Austen, G., 2023. Development of a machine vision system for damage and object detection in tunnels using convolutional neural networks. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 10, 1–8. http://dx.doi.org/10.5194/isprs-annals-X-1-W1-2023-1-2023.

Attard, L., Debono, C.J., Valentino, G., Di Castro, M., 2018. Tunnel inspection using photogrammetric techniques and image processing: A review. ISPRS J. Photogramm. Remote Sens. 144, 180–188. http://dx.doi.org/10.1016/j.isprsjprs.2018.07.010.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495. http://dx.doi.org/10.48550/arXiv.1511.00561.

Baheti, B., Innani, S., Gajre, S., Talbar, S., 2020. Semantic scene segmentation in unstructured environment with modified DeepLabV3+. Pattern Recognit. Lett. 138, 223–229. http://dx.doi.org/10.1016/j.patrec.2020.07.029.

Bhowmick, S., Nagarajaiah, S., Veeraraghavan, A., 2020. Vision and deep learning-based algorithms to detect and quantify cracks on concrete surfaces from UAV videos. Sensors 20 (21), 6299. http://dx.doi.org/10.3390/s20216299.

Bi Yang, Z.J., 2016. Review of convolution neural network. J. Univ. South China (Sci. Technol.) 30 (3), 7.

Cai, H., Gan, C., Han, S., 2022. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. http://dx.doi.org/10.48550/arXiv.2205.14756, arXiv preprint arXiv:2205.14756.

Cao, Z., Chen, D., Peethambaran, J., Zhang, Z., Xia, S., Zhang, L., 2021. Tunnel reconstruction with block level precision by combining data-driven segmentation and model-driven assembly. IEEE Trans. Geosci. Remote Sens. 59 (10), 8853–8872. http://dx.doi.org/10.1109/TGRS.2020.3046624.

Cao, Z., Chen, D., Shi, Y., Zhang, Z., Jin, F., Yun, T., Xu, S., Kang, Z., Zhang, L., 2019. A flexible architecture for extracting metro tunnel cross sections from terrestrial laser scanning point clouds. Remote Sens. 11 (3), 297. http://dx.doi.org/10.3390/rs11030297.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Springer, pp. 205–218. http://dx.doi.org/10.1007/978-3-031-25066-8_9.

Cha, Y.-J., Choi, W., Büyüköztürk, O., 2017. Deep learning-based crack damage detection using convolutional neural networks. Comput.-Aided Civ. Infrastruct. Eng. 32 (5), 361–378. http://dx.doi.org/10.1111/mice.12263.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. http://dx.doi.org/10.48550/arXiv.2102.04306, arXiv preprint arXiv:2102.04306.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848. http://dx.doi.org/10.1109/TPAMI.2017.2699184.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation. http://dx.doi.org/10.48550/arXiv.1706.05587, arXiv e-prints, arXiv–1706.

Chen, C., Qi, F., 2019. Review on development of convolutional neural network and its application in computer vision. Comput. Sci. 46 (3), 63–73.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818. http://dx.doi.org/10.48550/arXiv.1802.02611.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1251–1258. http://dx.doi.org/10.48550/arXiv.1610.02357.

Deng, P., 2015. Research on Detection of Railway Tunnel Seepage with Train-Mounted GPG (Ph.D. thesis). Southwest Jiaotong University.

Dian, S., Huang, J., Wu, K., et al., 2022. TCS-Net: A tiny crack segmentation network for nuclear containment vessel. Adv. Eng. Sci. 54 (5), 249–256.

Diao, Z., Chunjiang, Z., Gang, W., et al., 2010. Application research of mathematical morphology in image processing crop disease. J. Image Graph. 15 (2), 194–199.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. http://dx.doi.org/10.48550/arXiv.2010.11929, arXiv preprint arXiv:2010.11929.

Dung, C.V., Anh, L.D., 2019. Autonomous concrete crack detection using deep fully convolutional neural network. Autom. Constr. 99, 52–58. http://dx.doi.org/10.1016/j.autcon.2018.11.028.

Erdem, F., Ocer, N.E., Matci, D.K., Kaplan, G., Avdan, U., 2023. Apricot tree detection from UAV-images using mask R-CNN and U-Net. Photogramm. Eng. Remote Sens. 89 (2), 89–96. http://dx.doi.org/10.14358/PERS.22-00086R2.

Feng, Y., Feng, S.-J., Zhang, X.-L., Zhang, D.-M., Zhao, Y., 2024. A two-step deep learning-based framework for metro tunnel lining defect recognition. Tunn. Undergr. Space Technol. 150, 105832. http://dx.doi.org/10.1016/j.tust.2024.105832.

Feng, Y., Zhang, X., Feng, S., Chen, H., Zhao, Y., Chen, Y., 2023. Improved SOLOv2 detection method for shield tunnel lining water leakages. J. Intell. Constr. 1 (1), 9180004. http://dx.doi.org/10.26599/JIC.2023.9180004.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3146–3154. http://dx.doi.org/10.48550/arXiv.1809.02983.

He, B., 2020. Research on Real-time Semantic Segmentation Based on Lightweight Encoder-decoder (Ph.D. thesis). XiDian University.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. http://dx.doi.org/10.48550/arXiv.1512.03385.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. http://dx.doi.org/10.48550/arXiv.1704.04861, arXiv preprint arXiv:1704.04861.

Huang, Z., Zhangg, C.-l., Fu, H.-l., Ma, S.-k., Fan, X.-d., 2021. Machine inspection equipment for tunnels: a review. J. Highw. Transp. Res. Dev. (Engl. Ed.) 15 (2), 40–53. http://dx.doi.org/10.1061/JHTRCQ.0000774.

Huang, H., Zhao, S., Zhang, D., Chen, J., 2022. Deep learning-based instance segmentation of cracks from shield tunnel lining images. Struct. Infrastruct. Eng. 18 (2), 183–196. http://dx.doi.org/10.1080/15732479.2020.1838559.

Islam, M.A., Jia, S., Bruce, N.D., 2019. How much position information do convolutional neural networks encode? In: International Conference on Learning Representations. http://dx.doi.org/10.48550/arXiv.2001.08248.

Ji, A., Xue, X., Wang, Y., Luo, X., Xue, W., 2020. An integrated approach to automatic pixel-level crack detection and quantification of asphalt pavement. Autom. Constr. 114 (6), 1–15. http://dx.doi.org/10.1016/j.autcon.2020.103176.

Kim, I., Lee, C., 2018. Development of video shooting system and technique enabling detection of micro cracks in the tunnel lining while driving. J. Korean Soc. Hazard Mitig. 18 (5), 217–229. http://dx.doi.org/10.9798/KOSHAM.2018.18.5.217.

Lei, Y., Zuo, M.J., 2009. Gear crack level identification based on weighted K nearest neighbor classification algorithm. Mech. Syst. Signal Process. 23 (5), 1535–1547. http://dx.doi.org/10.1016/j.ymssp.2009.01.009.

Lin, G., Milan, A., Shen, C., Reid, I., 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1925–1934. http://dx.doi.org/10.48550/arXiv.1611.06612.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al., 2022b. Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 12009–12019. http://dx.doi.org/10.48550/arXiv.2111.09883.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 10012–10022. http://dx.doi.org/10.48550/arXiv.2103.14030.

Liu, S., Sun, H., Zhang, Z., Li, Y., Zhong, R., Li, J., Chen, S., 2022a. A multiscale deep feature for the instance segmentation of water leakages in tunnel using MLS point cloud intensity images. IEEE Trans. Geosci. Remote Sens. 60, 1–16. http://dx.doi.org/10.1109/TGRS.2022.3158660.

Liu, H., Zhang, H.K., Huang, B., Yan, L., Tran, K.K., Qiu, Y., Zhang, X., Roy, D.P., 2024. Reconstruction of seamless harmonized Landsat Sentinel-2 (HLS) time series via self-supervised learning. Remote Sens. Environ. 308, 114191. http://dx.doi.org/10.1016/j.rse.2024.114191.

Liujie Sun, Z.Y., Wenju, W., 2022. Lightweight semantic segmentation network for RGB-D image based on attention mechanism. Packag. Eng. 43 (3), 10.

Long, J., Shelhamer, E., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (4), 640–651. http://dx.doi.org/10.1109/TPAMI.2016.2572683.

Menendez, E., Victores, J.G., Montero, R., Martínez, S., Balaguer, C., 2018. Tunnel structural inspection and assessment using an autonomous robotic system. Autom. Constr. 87, 117–126. http://dx.doi.org/10.1016/j.autcon.2017.12.001.

Peng, M., 2016. Discussion on improving the maintenance and management efficiency of electromechanical system equipment in urban subways. Archit. Eng. Technol. Des. (023), 1588.

Qu, Z., Feng, H., Zeng, Z., Zhuge, J., Jin, S., 2010. A SVM-based pipeline leakage detection and pre-warning system. Measurement 43 (4), 513–519. http://dx.doi.org/10.1016/j.measurement.2009.12.022.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Rui Gong, D.S., Chaohua, Z., 2020. Lightweight and multi-pose face recognition method based on deep learning. J. Comput. Appl. 40 (3), 6.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015). Computational and Biological Learning Society, http://dx.doi.org/10.48550/arXiv.1409.1556.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. http://dx.doi.org/10.48550/arXiv.1706.03762, arXiv preprint arXiv:1706.03762.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 568–578. http://dx.doi.org/10.48550/arXiv.2102.12122.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2022. Pvtv2: Improved baselines with pyramid vision transformer. Comput. Vis. Media 8 (3), 1–10. http://dx.doi.org/10.1007/s41095-022-0274-8.

Weibo Xu, Z.H., 2017. Research progress in image segmentation based on region growing. Beijing Biomed. Eng. 36 (3), 6.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 16133–16142. http://dx.doi.org/10.48550/arXiv.2301.00808.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19. http://dx.doi.org/10.48550/arXiv.1807.06521.

Xia, C., Wang, X., Lv, F., Hao, X., Shi, Y., 2024. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5493–5502. http://dx.doi.org/10.48550/arXiv.2403.07392.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1492–1500. http://dx.doi.org/10.48550/arXiv.1611.05431.

Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., Luo, P., 2021a. Segmenting transparent object in the wild with transformer. http://dx.doi.org/10.48550/arXiv.2101.08461, arXiv e-prints, arXiv–2101.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021b. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090. http://dx.doi.org/10.48550/arXiv.2105.15203.

Xu, X., Yang, H., 2019. Intelligent crack extraction and analysis for tunnel structures with terrestrial laser scanning measurement. Adv. Mech. Eng. 11 (9), 1687814019872650. http://dx.doi.org/10.1177/1687814019872650.

Xue, Y., Li, Y., 2018. A fast detection method via region-based fully convolutional neural networks for shield tunnel lining defects. Comput.-Aided Civ. Infrastruct. Eng. 33 (8), 638–654. http://dx.doi.org/10.1111/mice.12367.

Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K., 2018. Denseaspp for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3684–3692. http://dx.doi.org/10.1109/CVPR.2018.00388.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. http://dx.doi.org/10.48550/arXiv.1511.07122, arXiv preprint arXiv:1511.07122.

Zhang, F., Liu, B., Liu, L., Wang, J., Lin, C., Yang, L., Li, Y., Zhang, Q., Yang, W., 2019. Application of ground penetrating radar to detect tunnel lining defects based on improved full waveform inversion and reverse time migration. Near Surf. Geophys. 17 (2), 127–139. http://dx.doi.org/10.1002/nsg.12032.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2881–2890. http://dx.doi.org/10.48550/arXiv.1612.01105.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., Zhang, L., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 6877–6886. http://dx.doi.org/10.1109/CVPR46437.2021.00681.

Zhou, Q., Qu, Z., Li, Y.-X., Ju, F.-R., 2022. Tunnel crack detection with linear seam based on mixed attention and multiscale feature fusion. IEEE Trans. Instrum. Meas. 71, 1–11. http://dx.doi.org/10.1109/TIM.2022.3184351.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer International Publishing, Springer, pp. 3–11. http://dx.doi.org/10.48550/arXiv.1807.10165.

Zhou, Z., Zhang, J., Gong, C., 2023. Hybrid semantic segmentation for tunnel lining cracks based on Swin Transformer and convolutional neural network. Comput.-Aided Civ. Infrastruct. Eng. 38 (17), 2491–2510. http://dx.doi.org/10.1111/mice.13003.

Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.W., 2023. BiFormer: Vision transformer with Bi-level routing attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 10323–10333. http://dx.doi.org/10.48550/arXiv.2303.08810.