http://dx.doi.org/10.62762/TSCC.2026.978086

RESEARCH ARTICLE

FocusNet: Feature Oriented Contextual Understanding via Bidirectional Supervision for Surface Defect Detection

Muhammad Hamza Ali[1], **Farhan Ali**[2] **and Zaid Muhammad**[3,*]

[1] Department of Computer Science, Virtual University of Pakistan, Islamabad 44000, Pakistan
[2] Department of Computer Science, Graz University of Technology, Graz 8010, Austria
[3] Global Degree College, Peshawar 25120, Pakistan

Abstract

Surface defect segmentation in metallic materials presents significant challenges due to irregular defect shapes, extreme scale variations ranging from small blowholes to large uneven regions, and low contrast with complex background textures. While Convolutional Neural Networks excel at local feature extraction, their limited receptive fields hinder effective global context modeling. Conversely, Vision Transformers capture long-range dependencies but struggle with fine-grained boundary details critical for accurate defect localization. To address these limitations, we propose FocusNet, a novel architecture integrating multi-scale feature refinement, hybrid attention mechanisms, and progressive decoding with bidirectional supervision for robust metallic surface defect segmentation. FocusNet introduces a Feature Refinement Module combining Residual Feature Enhancement Blocks and Atrous Spatial Pyramid Pooling to capture multi-scale contextual information while preserving fine-grained details. A Hybrid Attention Mechanism synergistically fuses channel, spatial, and coordinate attention to enhance discriminative feature representations across both global and local scales. The Bidirectional Feature Pyramid Network enables efficient weighted multi-scale feature fusion, facilitating robust detection across diverse defect characteristics. Our progressive decoder employs deep supervision at multiple hierarchical levels, enforcing intermediate representations to learn defect patterns and improving boundary localization accuracy. Extensive experiments on the MT-Defect benchmark demonstrate state-of-the-art performance with 81.6% mIoU, surpassing LGGFormer (80.0%) and SegFormer (66.1%). Notably, FocusNet achieves particularly strong gains on challenging defect categories, including Crack (70.7% vs. 64.9%) and Break (74.3% vs. 72.4%) compared with the best-performing prior method LGGFormer, demonstrating robust segmentation across defects with diverse scales, orientations, and boundary complexities.

Keywords: surface defect segmentation, metallic tile inspection, hybrid attention mechanism, bidirectional

Submitted: 04 March 2026
Revised: 22 June 2026
Accepted: 23 July 2026
Published: 17 September 2026

Vol. 3, **No.** 3, 2026.
10.62762/TSCC.2026.978086

***Corresponding author:**
✉ Zaid Muhammad
zaidmuhammadsaid14@gmail.com

Citation
Ali, M. H., Ali, F., & Muhammad, Z. (2026). FocusNet: Feature Oriented Contextual Understanding via Bidirectional Supervision for Surface Defect Detection. *ICCK Transactions on Sensing, Communication, and Control*, 3(3), 160–175.

feature pyramid network, deep supervision, industrial quality control.

1 Introduction

Surface defect detection plays a crucial role in industrial quality control, directly impacting product reliability and manufacturing efficiency [1]. Traditional inspection methods, including manual examination and non-destructive testing techniques, face significant limitations. Manual inspection is time-consuming, labor-intensive, and prone to human error. Such limitations have driven the adoption of automated vision-based inspection systems powered by deep learning.

Deep learning approaches for surface defect detection can be categorized into three paradigms: image-level classification, region-level detection, and pixel-level segmentation. Image-level methods classify defects but cannot localize their positions [2]. Region-level detection using bounding boxes provides approximate localization but fails to capture precise defect boundaries and often includes irrelevant background information [3, 4]. In contrast, pixel-level segmentation methods offer the most comprehensive solution by accurately delineating defect regions while simultaneously providing location, type, and precise boundary information [5–7]. Consequently, semantic segmentation has become the preferred approach for industrial defect inspection.

However, metallic surface defect segmentation presents unique challenges that distinguish it from conventional segmentation tasks. Defects in magnetic tiles exhibit irregular defect shapes, extreme scale variations ranging from small blowholes to large uneven regions, and low contrast with complex background textures [8, 9]. Additionally, industrial imaging conditions introduce illumination variations, shadows, and reflective artifacts that further obscure defect boundaries, resulting in weak textural features that are difficult to distinguish from normal surfaces [10]. These factors impose stringent requirements on segmentation architectures to simultaneously capture fine-grained local details and broad contextual information.

Convolutional Neural Networks (CNNs) have dominated defect detection due to their strong local feature extraction capabilities via hierarchical convolutions [11–13]. However, CNNs are inherently limited by their restricted receptive fields, which hinder their ability to model long-range dependencies necessary for understanding global defect patterns. While techniques such as large kernel convolutions [14] and attention mechanisms [15] partially alleviate this limitation, the fundamental constraint of locality-focused convolution operations persists. This results in suboptimal performance when segmenting large-area defects or capturing holistic defect structures across the image.

Vision Transformers have emerged as a compelling alternative, leveraging self-attention mechanisms to model global dependencies by treating images as sequences of patches [16–20]. While Transformers excel at capturing long-range contextual relationships and overall defect shapes, their global focus comes at the expense of local detail preservation. Self-attention's emphasis on token-to-token interactions across the entire image can dilute fine-grained textural information critical for detecting microcracks, subtle scratches, and precise defect boundaries. This trade-off between global context and local detail presents a fundamental challenge for defect segmentation architectures.

Recent research has explored hybrid CNN-Transformer architectures to leverage complementary strengths [21–23]. These approaches can be categorized into serial or parallel combinations of CNN and Transformer blocks, or integration of convolutional locality into self-attention mechanisms [24, 25]. Despite these efforts, performance gains remain limited due to inherent inconsistencies between CNN and Transformer computational paradigms [26]. Standard convolutions are input-agnostic with fixed kernels, whereas self-attention dynamically computes attention weights based on input content. This fundamental difference in representational capacity creates integration challenges, and naive fusion strategies such as element-wise addition or concatenation often fail to fully exploit their complementary advantages [27]. Additionally, existing hybrid methods struggle with sample imbalance prevalent in industrial settings, where defective samples are significantly scarcer than normal samples [12]. Collectively, these limitations of restricted receptive fields in CNNs, loss of fine-grained local detail in Transformers, and suboptimal integration in hybrid architectures underscore the need for a unified framework that jointly addresses multi-scale contextual learning, discriminative feature recalibration, and robust cross-scale fusion for metallic surface defect segmentation.

To address these challenges, we propose FocusNet, a novel architecture specifically designed for metallic tile defect detection that systematically integrates multi-scale feature refinement, hybrid attention mechanisms, and progressive decoding with deep supervision. Our approach makes the following key contributions, forming a synergistic pipeline where FRM prepares multi-scale contextual features, HAM recalibrates them for discriminative representation, BiFPN fuses information across hierarchical scales, and the progressive decoder refines predictions with boundary-aware deep supervision:

- We propose a Feature Refinement Module (FRM) combining Residual Feature Enhancement Blocks (RFEB) and Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale contextual information while preserving fine-grained details critical for detecting defects with diverse scales.
- We introduce a Hybrid Attention Mechanism (HAM) that synergistically combines channel, spatial, and coordinate attention to enhance discriminative feature representations, effectively capturing both inter-channel dependencies and spatial relationships essential for distinguishing subtle defects from complex backgrounds.
- We design a Bidirectional Feature Pyramid Network (BiFPN) with weighted multi-scale feature fusion that enables efficient information flow across hierarchical levels, facilitating robust detection of both small-scale and large-area defects.
- We develop a progressive decoder with deep supervision strategy that applies auxiliary losses at multiple scales, enforcing intermediate representations to learn hierarchical defect patterns and improving boundary localization accuracy.
- Comprehensive experiments on the MT-Defect benchmark demonstrate that FocusNet achieves state-of-the-art performance with 81.6% mIoU, surpassing recent methods while maintaining computational efficiency suitable for industrial deployment.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 details the proposed FocusNet architecture, covering the FRM, HAM, BiFPN, and progressive decoder with deep supervision, Section 4 presents experimental results and analysis, and Section 5 concludes the paper.

2 Related Work

2.1 CNN-Based Defect Segmentation

Convolutional Neural Networks have long dominated surface defect segmentation due to their strong local feature extraction capabilities through hierarchical convolutions [28, 29]. Early work such as FFCNN [68] demonstrated the effectiveness of deep convolutional networks for surface defect detection of magnetic tiles, establishing a foundation for subsequent CNN-based defect inspection research. Liu and He [30] proposed a triple attention semantic segmentation network that captures defect contextual information through a context attention module for small defect segmentation. Zhang et al. [31] introduced a dual-branch real-time network incorporating boundary detail encoding and semantic context through auxiliary tasks, employing Global Context Upsampling to address local similarity issues. Yeung and Lam [32] developed a transformer-augmented model integrating spatial and channel attention for merging boundary and context information. Zuo et al. [33] proposed a multi-scale segmentation model combining feature pyramids and attention mechanisms for detecting defects of varying sizes. Yin et al. [34] presented a dual-branch network enriching feature representations through semantic-boundary integration with a global-local fusion module. Cao et al. [35] proposed a pixel-level segmentation network based on deep feature fusion, integrating multi-scale convolutional features to improve detection accuracy across defect categories of varying sizes and shapes. Qiu et al. [36] introduced a region and edge-aware network utilizing Feature Pyramid Edge modules and Cross-Level Fusion for railway defect detection.

To enhance spatial information capture, several methods incorporate attention mechanisms. Wang et al. [37] proposed the lightweight BSRA module for capturing long-range dependencies in high-resolution feature maps. Yang et al. [38] integrated spatial attention into residual units for adaptive defect area focus, while Yang et al. [39] proposed GSRM to determine global semantic relevance for improved feature representation. Despite these advances, CNNs remain limited by restricted receptive fields, hindering effective global dependency modeling and resulting in suboptimal performance on large-area or scattered defects. Furthermore, industrial defect datasets inherently suffer from severe class imbalance and sample scarcity, as defective samples are significantly scarcer than defect-free samples [12]; this has motivated few-shot approaches that learn robust

representations from limited labeled examples [40], posing additional challenges for CNN-based pixel-level segmentation.

2.2 Transformer-Based Defect Segmentation

Vision Transformers have been successfully adapted to a wide range of computer vision tasks. Dosovitskiy et al. [17] proposed Vision Transformer (ViT) by introducing self-attention mechanisms through image patching, linear embedding, and positional encoding. Wang et al. [41] developed Pyramid Vision Transformer (PVT) with pyramid structures for multi-scale feature handling. Zheng et al. [42] applied pure Transformer architectures to semantic segmentation, leveraging global modeling capabilities. Liu et al. [43] introduced Swin Transformer with hierarchical structures and window-based self-attention for improved efficiency. Xie et al. [44] designed SegFormer with hierarchical Transformer encoders and lightweight MLP decoders for efficient segmentation.

To address the quadratic complexity of self-attention, two primary approaches have emerged. The first employs local attention by restricting attention to windows [45, 46]. Swin [43] introduced window-based self-attention to enhance feature locality while reducing computational complexity, which CSWin [45] further improved with cross-shaped windows. Zhu et al. [47] proposed an efficient LSwin window shift computation. However, these approaches reduce receptive fields and long-range modeling capability. The second approach generates coarse attention matrices through spatial downsampling while maintaining global receptive fields [41, 48, 49].

While Transformers excel at global feature extraction and mitigating background interference[18], they struggle with local detail capture critical for defect segmentation, particularly for small-scale defects with weak appearances. Shang et al. [50] proposed DAT-Net, a defect-aware Transformer that incorporates graph position encoding and a defect-aware module into the self-attention mechanism to capture defect geometry and suppress background interference. Zhou et al. [51] introduced ETDNet, an efficient Transformer-based detection network integrating multi-head attention with lightweight classification and perception heads for accurate segmentation of surface defects with varying shapes and scales. Among the most recent methods targeting the MT-Defect benchmark, Zhang et al. [52] proposed LGGFormer, a dual-branch network combining CNN and Transformer through a Local-Guided Global Attention Self-Attention (LGGSA) mechanism, wherein local sliding-window features actively guide global attention computation. A Cross-Branch Feature Interaction (CBFI) module enables correlation-weighted CNN-Transformer fusion at each encoder stage, and an Edge-Guided Decoder (EGD) leverages CNN-extracted boundary information to refine high-level features. Despite these promising results, existing methods generally rely on complex multi-branch designs or dedicated boundary supervision modules, which increase architectural overhead and do not explicitly address the joint challenges of multi-scale feature refinement, discriminative attention recalibration, and cross-scale fusion required for robust segmentation across diverse defect scales and orientations.

2.3 Hybrid CNN-Transformer Architectures

The complementary strengths and limitations of CNNs and Transformers have motivated the development of hybrid architectures that leverage local detail extraction from CNNs and global context modeling from Transformers for robust defect segmentation [21–23]. These approaches can be categorized into serial or parallel combinations of CNN and Transformer blocks, or integration of convolutional locality into self-attention mechanisms [24, 25]. For instance, Jiang et al. [53] proposed CINFormer, which injects multi-level CNN features into successive stages of a Transformer encoder to preserve local detail capture while suppressing background noise, achieving competitive performance on magnetic tile defect datasets. Prior CNN-based methods such as MCNet [54] have demonstrated the value of multi-scale context aggregation for industrial surface defect segmentation under complex backgrounds. However, such CNN-based methods remain constrained by the limited global modeling capacity of purely convolutional architectures, underscoring the need for hybrid CNN-Transformer designs. Despite these efforts, performance gains remain limited due to inherent inconsistencies between CNN and Transformer computational paradigms [26]. Standard convolutions are input-agnostic with fixed kernels, whereas self-attention dynamically computes attention weights based on input content. This fundamental difference in representational capacity creates integration challenges, and naive fusion strategies such as element-wise addition

or concatenation often fail to fully exploit their complementary advantages [27]. Additionally, existing hybrid methods struggle with sample imbalance prevalent in industrial settings, where defective samples are significantly scarcer than normal samples [12]. These limitations motivate FocusNet's unified design, which addresses multi-scale contextual learning, discriminative feature recalibration, and robust cross-scale fusion within a single coherent framework.

3 Proposed Methodology

3.1 Overview of FocusNet Architecture

The proposed FocusNet addresses the challenges of metallic tile defect detection through an encoder-decoder architecture integrating five key components (as shown in Figure 1): (1) EfficientNet-B4 backbone for multi-scale feature extraction, (2) Feature Refinement Modules (FRMs) with Atrous Spatial Pyramid Pooling for contextual aggregation, (3) Hybrid Attention Mechanisms (HAMs) combining channel, spatial, and coordinate attention for adaptive feature recalibration, (4) Bidirectional Feature Pyramid Network (BiFPN) enabling efficient cross-scale fusion through learnable weighted connections, and (5) Progressive Decoder with deep supervision for precise boundary-aware surface defect localization. This hierarchical design captures both fine-grained details and global context while ensuring effective gradient flow through multi-level supervision for robust magnetic tile inspection. Unless otherwise specified, all convolutional operations throughout FocusNet use a padding of 1 and stride of 1 to preserve spatial dimensions; the sole exception is the Down$(\cdot)$ max-pooling operation in BiFPN, which uses stride 2 for spatial downsampling.

3.2 Backbone Network

FocusNet employs EfficientNet-B4 [55] as the backbone due to its optimal balance between accuracy and computational efficiency. EfficientNet-B4 utilizes compound scaling to uniformly scale network width, depth, and resolution, achieving superior feature extraction capability with reduced parameters compared to conventional architectures. The backbone extracts hierarchical features from five stages, producing multi-scale feature maps $\{C_1, C_2, C_3, C_4, C_5\}$ with downsampling factors $\{2, 4, 8, 16, 32\}$ relative to the input. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the backbone generates:

$$\mathcal{F}_{\text{backbone}} = \{C_1, C_2, C_3, C_4, C_5\} \tag{1}$$

where $C_i \in \mathbb{R}^{H_i\times W_i\times D_i}$ represents features at level i with spatial dimensions $H_i = H/2^i$, $W_i = W/2^i$, and channel dimensions $\{D_1, D_2, D_3, D_4, D_5\} = \{24, 32, 56, 160, 448\}$ for EfficientNet-B4. Pretrained ImageNet [56] weights provide robust initialization for metallic surface inspection, enabling effective transfer learning despite limited labeled defect data. This initialization preserves generalizable low-level features (edges, textures, corners) essential for industrial defect detection.

3.3 Feature Refinement Module

To enhance backbone features with multi-scale contextual information, we introduce a Feature Refinement Module (FRM) consisting of two sequential components: a Residual Feature Enhancement Block (RFEB) and an Atrous Spatial Pyramid Pooling (ASPP) module.

3.3.1 *Residual Feature Enhancement Block*

The RFEB employs residual learning to facilitate gradient flow and enable incremental feature refinement when adapting pretrained features to metallic surface defect detection. For input feature C_i, the RFEB operation is:

$$\begin{aligned}\text{RFEB}(C_i) = \text{ReLU}\Big(&\text{BN}\big(\text{Conv}_2\big(\text{ReLU}\big(\text{BN}\big(\text{Conv}_1(C_i)\big)\big)\big)\big)\\ &+ \mathcal{S}(C_i)\Big)\end{aligned} \tag{2}$$

where Conv_1 and Conv_2 are 3×3 convolutions, BN denotes batch normalization, and $\mathcal{S}(\cdot)$ is the shortcut connection:

$$\mathcal{S}(C_i) = \begin{cases} C_i & \text{if } D_i = D_{\text{fpn}} \\ \text{BN}\left(\text{Conv}_{1\times 1}(C_i)\right) & \text{otherwise}\end{cases} \tag{3}$$

where $D_{\text{fpn}} = 256$ is the unified channel dimension. The 1×1 convolution ensures compatibility when channel dimensions differ, preventing information loss during transformation.

3.3.2 *Atrous Spatial Pyramid Pooling*

ASPP [57] captures multi-scale context through parallel atrous convolutions with varying dilation rates, enabling detection of diverse defect sizes without increasing parameters or reducing resolution. The module processes features through five parallel

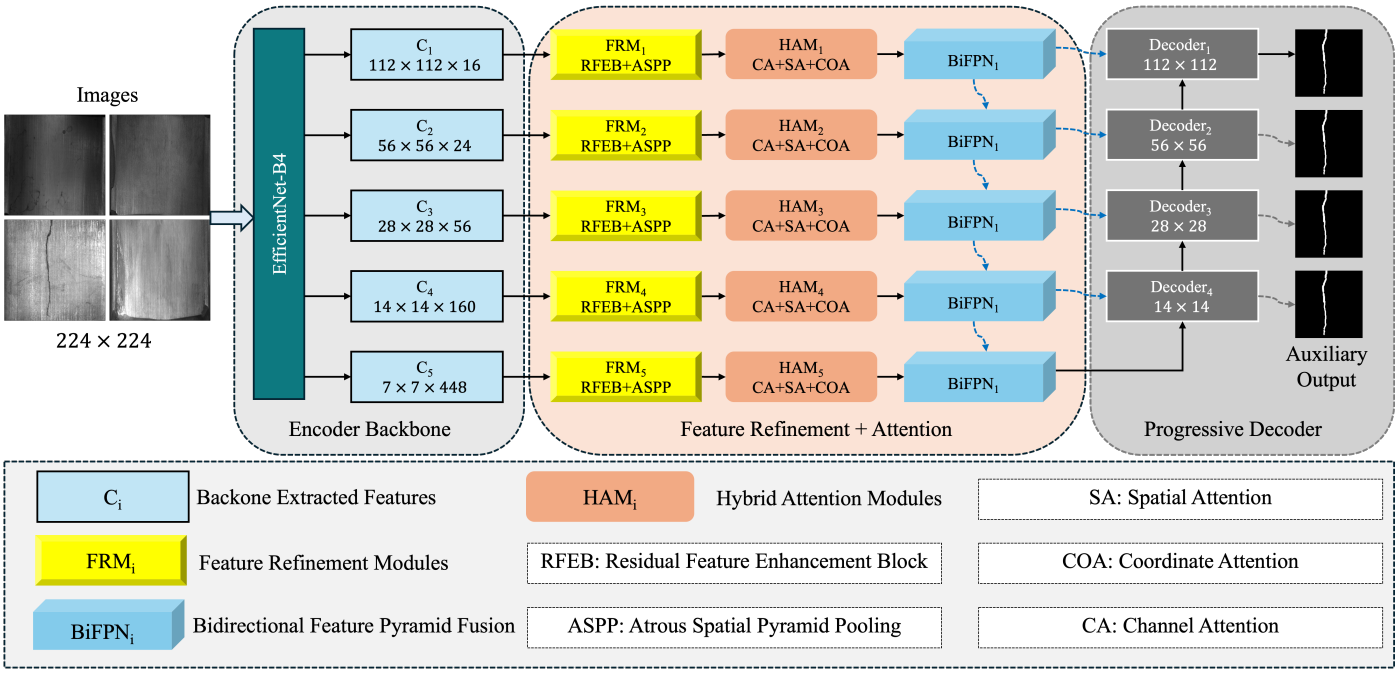

Figure 1. Overall architecture of the proposed FocusNet showing the encoder-decoder structure with multi-scale feature extraction, ASPP-based refinement, hybrid attention, BiFPN cross-scale fusion, and progressive decoding with deep supervision for surface defect detection.

branches:

$$\text{ASPP}(x) = \text{Conv}_{1\times1}\Big(\text{Concat}\big[\text{Conv}_{1\times1}(x), \text{Conv}^{d=6}_{3\times3}(x), \text{Conv}^{d=12}_{3\times3}(x), \text{Conv}^{d=18}_{3\times3}(x), \text{GAP}(x)\big]\Big) \tag{4}$$

where $\text{Conv}^{d}_{k\times k}$ denotes $k\times k$ convolution with dilation rate d, Concat$[\cdot]$ is channel-wise concatenation, and GAP represents global average pooling:

$$\text{GAP}(x) = \text{Upsample}\left(\text{Conv}_{1\times1}\left(\text{AdaptiveAvgPool}(x)\right)\right) \tag{5}$$

The five branches capture features at receptive fields of 1×1 (local), 13×13 ($d=6$), 25×25 ($d=12$), 37×37 ($d=18$), and global scale, enabling simultaneous detection of small scratches and large contamination regions in magnetic tiles. The concatenated features are fused through 1×1 convolution with batch normalization and ReLU activation. The complete FRM transforms each backbone feature as:

$$F_{\text{refined},i} = \text{ASPP}\left(\text{RFEB}(C_i)\right), \quad i \in \{1,2,3,4,5\} \tag{6}$$

producing refined features $\mathcal{F}_{\text{refined}} = \{F_{\text{refined},1}, \ldots, F_{\text{refined},5}\}$ with unified dimension $D_{\text{fpn}} = 256$.

3.4 Hybrid Attention Mechanism

FocusNet integrates a hybrid attention mechanism combining three complementary attention types to adaptively recalibrate features based on *what* (channel attention), *where* (spatial attention), and *directional* information (coordinate attention) [58–60].

Channel attention recalibrates features by modeling channel interdependencies, emphasizing informative channels while suppressing less useful ones. Given input feature $F \in \mathbb{R}^{H\times W\times C}$, channel attention aggregates spatial information through both pooling operations:

$$\text{CA}(F) = \sigma\left(\text{MLP}\left(\text{AvgPool}(F)\right) + \text{MLP}\left(\text{MaxPool}(F)\right)\right) \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function and MLP is a two-layer bottleneck: $\text{MLP}(z) = W_2 \cdot \text{ReLU}(W_1 \cdot z)$ with $W_1 \in \mathbb{R}^{(C/r)\times C}$, $W_2 \in \mathbb{R}^{C\times(C/r)}$, and reduction ratio $r = 16$. The channel weights $\mathcal{M}_c = \text{CA}(F) \in \mathbb{R}^{1\times1\times C}$ are applied via element-wise multiplication: $F' = F \odot \mathcal{M}_c$. Spatial attention highlights informative spatial locations for defect localization by aggregating channel information:

$$\text{SA}(F) = \sigma\Big(\text{Conv}_{7\times7}\big(\text{Concat}[\text{AvgPool}_c(F), \text{MaxPool}_c(F)]\big)\Big) \tag{8}$$

where $\text{AvgPool}_c(\cdot)$ and $\text{MaxPool}_c(\cdot)$ denote pooling along the channel dimension. The spatial weights $\mathcal{M}_s = \text{SA}(F) \in \mathbb{R}^{H\times W\times1}$ are applied as: $F'' = F' \odot \mathcal{M}_s$. The 7×7 kernel captures spatial relationships over a wider receptive field. Unlike traditional

spatial attention that compresses channel information, coordinate attention preserves directional positional information crucial for detecting oriented defects in metallic surfaces. It performs directional pooling along horizontal and vertical axes:

$$z_h^c(i) = \frac{1}{W}\sum_{j=0}^{W-1} x^c(i,j), \quad z_w^c(j) = \frac{1}{H}\sum_{i=0}^{H-1} x^c(i,j) \tag{9}$$

The pooled features $z = \text{Concat}[z_h, z_w] \in \mathbb{R}^{(H+W)\times C}$ are transformed via $f = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(z)))$ with channel reduction to C/r. After splitting along the spatial dimension, separate attention weights are generated:

$$g_h = \sigma(\text{Conv}_{1\times1}(f_h)) \in \mathbb{R}^{H\times1\times C} \tag{10}$$

$$g_w = \sigma(\text{Conv}_{1\times1}(f_w)) \in \mathbb{R}^{1\times W\times C} \tag{11}$$

The coordinate attention output $\text{CA}_{\text{coord}}(F) = F \odot g_h \odot g_w$ enables detection of linear defects such as scratches and cracks with strong directional characteristics.

Integration. The three attention mechanisms are applied sequentially for comprehensive feature refinement:

$$\begin{aligned} F_1 &= F_{\text{refined}} \odot \text{CA}(F_{\text{refined}}) \\ F_2 &= F_1 \odot \text{SA}(F_1) \\ F_{\text{att}} &= \text{CA}_{\text{coord}}(F_2) \end{aligned} \tag{12}$$

This hybrid approach leverages complementary strengths: channel attention identifies *which* features are important, spatial attention determines *where* to focus, and coordinate attention preserves *directional* relationships, resulting in both discriminative and spatially precise features for magnetic tile defect detection.

3.5 Bidirectional Feature Pyramid Network

Multi-scale feature fusion is essential for detecting defects of varying sizes. FocusNet employs BiFPN with learnable weights for adaptive feature fusion across scales. Given refined features $\mathcal{F}_{\text{refined}} = \{P_1, P_2, P_3, P_4, P_5\}$ where $P_i \in \mathbb{R}^{H_i\times W_i\times D_{\text{fpn}}}$, BiFPN performs bidirectional fusion through two pathways. The top-down pathway propagates high-level semantic information from coarse to fine resolutions:

$$P_i^{\text{td}} = \text{Conv}\left(\frac{w_1\cdot P_i + w_2\cdot \text{Up}(P_{i+1}^{\text{td}})}{w_1+w_2+\epsilon}\right), \; i\in\{4,3,2\} \tag{13}$$

$$P_1^{\text{out}} = \text{Conv}\left(\frac{w_7\cdot P_1 + w_8\cdot \text{Up}(P_2^{\text{td}})}{w_7+w_8+\epsilon}\right) \tag{14}$$

where $\text{Up}(\cdot)$ denotes upsampling via nearest neighbor interpolation, Conv represents 3×3 convolution with batch normalization and ReLU, $w_i = \text{ReLU}(\tilde{w}_i) \geq 0$ are learnable scalar weights, and $\epsilon = 10^{-4}$ ensures numerical stability. Specifically, the ReLU constraint ensures all learnable weights satisfy $w_i \geq 0$, guaranteeing that the denominator $\sum_i w_i + \epsilon$ is strictly positive at all times. The small constant $\epsilon = 10^{-4}$ prevents the denominator from approaching zero, avoiding division by zero during gradient computation while remaining sufficiently small to have negligible influence on the fused feature representations. This formulation allows the network to learn adaptive, data-driven fusion weights across scales without numerical instability during training. The bottom-up pathway enhances low-level features with refined semantic information:

$$P_i^{\text{out}} = \text{Conv}\left(\frac{w_1\cdot P_i + w_2\cdot P_i^{\text{td}} + w_3\cdot \text{Down}(P_{i-1}^{\text{out}})}{w_1+w_2+w_3+\epsilon}\right), \quad i\in\{2,3,4\} \tag{15}$$

$$P_5^{\text{out}} = \text{Conv}\left(\frac{w_{18}\cdot P_5 + w_{19}\cdot \text{Down}(P_4^{\text{out}})}{w_{18}+w_{19}+\epsilon}\right) \tag{16}$$

where $\text{Down}(\cdot)$ denotes max pooling with stride 2. The bottom-up pathway fuses three feature sources: original encoder features (P_i), top-down refined features (P_i^{td}), and adjacent finer-scale features (P_{i-1}^{out}). This weighted fusion enables adaptive emphasis on fine-scale features for small defects or coarse-scale features for large contamination regions in metallic tiles.

Multi-layer Refinement. FocusNet employs two stacked BiFPN blocks for iterative feature refinement:

$$\mathcal{F}_{\text{BiFPN}}^{(2)} = \text{BiFPN}_2(\text{BiFPN}_1(\mathcal{F}_{\text{refined}})) \tag{17}$$

where $\mathcal{F}_{\text{BiFPN}}^{(2)} = \{P_1^{\text{final}}, \ldots, P_5^{\text{final}}\}$ represents the final multi-scale features. This iterative process allows information to flow multiple times between scales, enhancing detection of challenging cases with multiple defect types or ambiguous boundaries.

3.6 Progressive Decoder with Deep Supervision

The decoder progressively combines multi-scale BiFPN features to generate high-resolution defect segmentation maps through hierarchical refinement at each scale. Deep supervision with auxiliary prediction heads provides multi-level gradient signals to facilitate training and improve feature learning.

Decoder Block. Each decoder block performs upsampling, skip connection fusion, and feature

refinement:

$$\begin{aligned}\tilde{D}_{i+1} &= \text{Upsample}(D_{i+1}) \\ \hat{D}_i &= \text{Concat}[\tilde{D}_{i+1}, P_i^{\text{final}}] \\ D_i &= \text{Conv}_2\left(\text{Conv}_1(\hat{D}_i)\right)\end{aligned} \tag{18}$$

where D_{i+1} is the decoder output from the coarser level, Upsample$(\cdot)$ denotes bilinear upsampling $(\times 2)$, P_i^{final} is the corresponding BiFPN feature, and Conv_1, Conv_2 are consecutive 3×3 convolutions with batch normalization and ReLU activation for feature refinement.

Hierarchical Decoding. The decoder processes features from coarse to fine resolutions through four stages with progressive channel reduction $\{1024 \to 512 \to 256 \to 128\}$:

$$D_4 = \text{DecoderBlock}(P_5^{\text{final}}, P_4^{\text{final}}) \in \mathbb{R}^{H/16 \times W/16 \times 1024} \tag{19}$$

$$D_3 = \text{DecoderBlock}(D_4, P_3^{\text{final}}) \in \mathbb{R}^{H/8 \times W/8 \times 512} \tag{20}$$

$$D_2 = \text{DecoderBlock}(D_3, P_2^{\text{final}}) \in \mathbb{R}^{H/4 \times W/4 \times 256} \tag{21}$$

$$D_1 = \text{DecoderBlock}(D_2, P_1^{\text{final}}) \in \mathbb{R}^{H/2 \times W/2 \times 128} \tag{22}$$

This strategy gradually recovers spatial resolution while maintaining semantic information from deeper layers and spatial details from shallower layers through skip connections.

Deep Supervision. Three auxiliary heads attached to intermediate decoder outputs $\{D_4, D_3, D_2\}$ generate multi-scale supervision:

$$\begin{aligned}\text{Aux}_i = \text{Conv}_{1\times1}\left(\text{Conv}_{3\times3}(D_{5-i})\right) \in \mathbb{R}^{H\times W\times 1}, \\ i \in \{1,2,3\}\end{aligned} \tag{23}$$

The final prediction head processes the finest decoder output:

$$\begin{aligned}\text{Final} = \text{Conv}_{1\times1}\Big(\text{ReLU}\big(\text{BN}\big(\text{Conv}_{3\times3}(D_1)\big)\big)\Big) \\ \in \mathbb{R}^{H\times W\times 1}\end{aligned} \tag{24}$$

All predictions are upsampled to input resolution via bilinear interpolation. The total training loss combines all supervision levels with decreasing weights for coarser predictions:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{final}} + \lambda_1\mathcal{L}_{\text{aux1}} + \lambda_2\mathcal{L}_{\text{aux2}} + \lambda_3\mathcal{L}_{\text{aux3}} \tag{25}$$

where $\mathcal{L}$ represents the segmentation loss (Section 3.7), and $\{\lambda_1, \lambda_2, \lambda_3\} = \{0.4, 0.3, 0.2\}$ emphasize the final prediction while providing intermediate gradient signals. This multi-scale supervision mitigates gradient vanishing, accelerates convergence, and improves feature learning for metallic surface defect detection.

3.7 Loss Function

The network is optimized using a hybrid loss combining Binary Cross-Entropy (BCE) and Dice loss to address both pixel-level accuracy and global region overlap. BCE loss measures pixel-wise classification error:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{26}$$

where $N = H \times W$ is the total number of pixels, $y_i \in \{0,1\}$ is the ground truth label, and $\hat{y}_i \in (0,1)$ is the predicted probability. Dice loss optimizes region overlap between predictions and ground truth:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i\hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N}\hat{y}_i + \epsilon} \tag{27}$$

where $\epsilon = 1$ is a smoothing constant. The final loss combines both components with equal weights:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{BCE}} + \beta \cdot \mathcal{L}_{\text{Dice}} \tag{28}$$

where $\alpha = \beta = 0.5$. This hybrid approach leverages BCE for pixel-level precision and Dice loss for handling class imbalance, effectively capturing both fine-grained boundaries and global defect regions in metallic surface defect segmentation.

4 EXPERIMENT AND RESULTS

4.1 Dataset

We evaluate FocusNet on the MT-Defect dataset [61], a challenging benchmark for metallic tile surface defect segmentation. The dataset comprises 1,344 images (392 defect images and 952 non-defect images) with five defect categories: blowhole, break, crack, fray, and uneven. Image resolutions range from 105×283 to 388×516 pixels. Following prior work on this benchmark [52, 62], the dataset is partitioned into training and test sets using an 80/20 stratified random split. The dataset presents significant challenges including complex background textures, diverse defect scales, and random illumination variations. Representative samples from the dataset are shown in Figure 2.

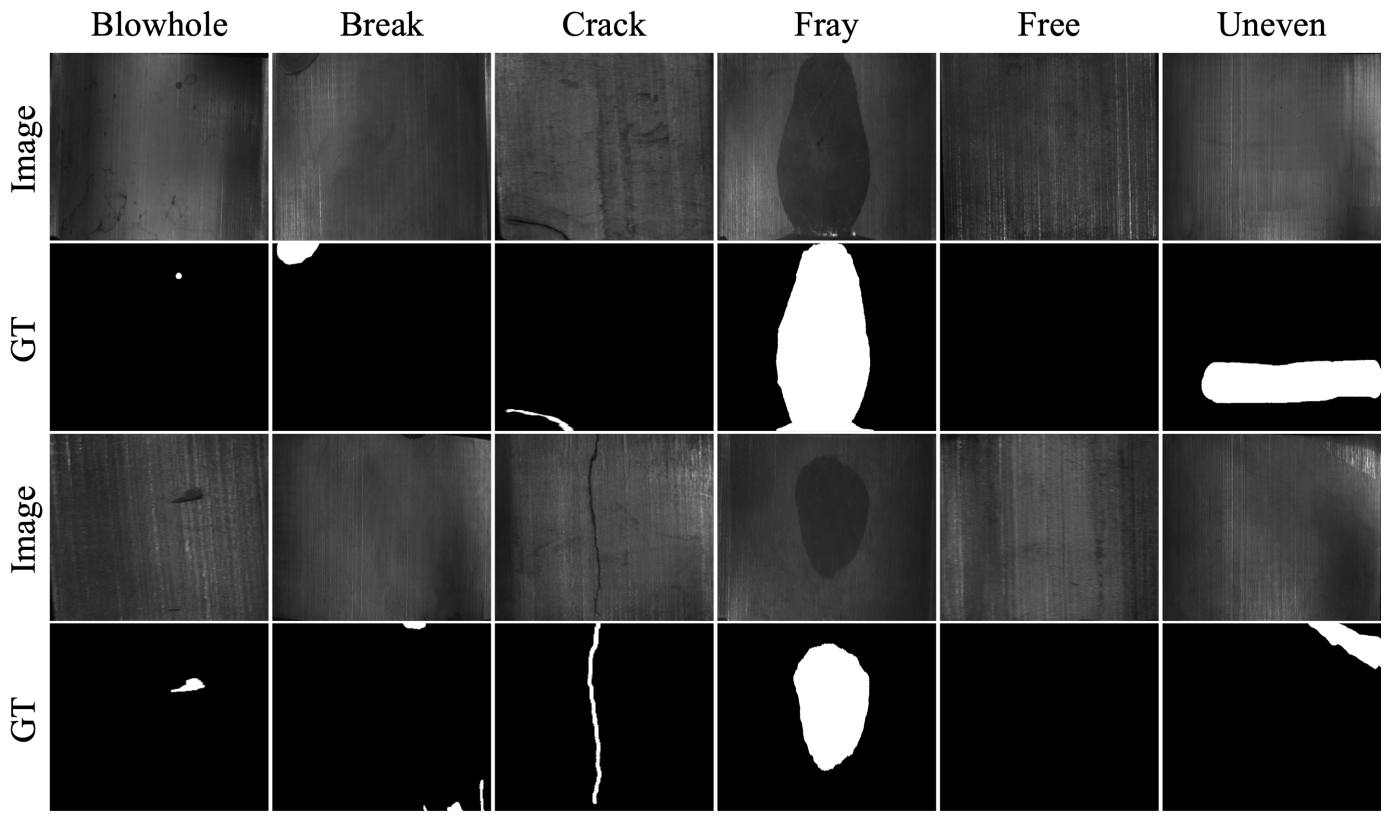

Figure 2. Representative samples from MT-Defect dataset showing defect categories. For each category, we display original metallic tile images (top row) and corresponding ground truth segmentation masks (middle row). The dataset exhibits diverse defect characteristics: small-scale Blowhole defects, large irregular Uneven regions, and thin elongated Crack structures.

4.2 Evaluation Metric

We adopt mean Intersection over Union (mIoU) as the primary evaluation metric, which measures the ratio of intersection to union between predicted segmentation and ground truth:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{29}$$

where N is the number of classes (including background), TP_i denotes true positive pixels, FP_i false positive pixels, and FN_i false negative pixels for class i. The mIoU provides a comprehensive assessment of segmentation accuracy by considering both precision and recall, making it particularly suitable for imbalanced defect detection tasks.

4.3 Implementation Details

FocusNet is implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU. We employ the AdamW optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-4}, using a cosine annealing schedule. Training uses batch size 16 with input images resized to 224×224 pixels following standard protocols [52, 63, 64]. Data augmentation includes random horizontal and vertical flips ($p = 0.5$), random rotation within $\pm 30°$, color jittering (brightness = 0.2, contrast = 0.2, saturation = 0.2, hue = 0.1), and Gaussian blur with kernel size randomly chosen from $\{3, 5\}$ and $\sigma \in [0.1, 2.0]$. The network is trained for 200 epochs. The EfficientNet-B4 backbone is initialized with ImageNet pretrained weights at a reduced learning rate of 1×10^{-5}. Deep supervision weights are $\lambda_1 = 0.4$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$, and the hybrid BCE-Dice loss uses $\alpha = \beta = 0.5$.

4.4 Ablation Study

We conduct extensive ablation experiments to validate the effectiveness of each component in FocusNet. All experiments are performed on the MT-Defect dataset using identical training configurations.

4.4.1 Component-wise Analysis

Table 1 presents the ablation results for key components of FocusNet. The baseline model consists of EfficientNet-B4 backbone with a simple decoder, achieving 75.3% mIoU. The Feature Refinement

Table 1. Component-wise ablation study on MT-Defect dataset.

FRM	HAM	BiFPN	DS	IoU (%)						mIoU (%)
				Bg	Bl	Br	Cr	Fr	Un	
✗	✗	✗	✗	95.8	64.2	66.5	62.1	88.3	74.7	75.3
✓	✗	✗	✗	96.9	67.1	69.2	65.3	90.1	77.4	77.7
✓	✓	✗	✗	97.5	69.4	71.8	67.9	91.6	79.2	79.6
✓	✓	✓	✗	98.1	70.3	73.1	69.5	92.4	80.1	80.6
✓	✓	✓	✓	**98.4**	**71.4**	**74.3**	**70.7**	**93.2**	**81.3**	**81.6**

Table 2. Comparison of different attention mechanisms on MT-Defect dataset.

Attention Mechanism	mIoU (%)
w/o Attention	77.7
Channel Attention	78.4
Spatial Attention	78.9
Coordinate Attention	79.3
Channel + Spatial	79.8
Channel + Coordinate	79.6
Spatial + Coordinate	80.1
Hybrid (Channel + Spatial + Coordinate)	**80.5**

Table 3. Impact of deep supervision weight configuration on MT-Defect dataset.

λ_1	λ_2	λ_3	mIoU (%)
w/o Deep Supervision			80.6
0.33	0.33	0.33	81.1
0.5	0.3	0.2	81.3
0.4	**0.3**	**0.2**	**81.6**
0.3	0.4	0.3	81.4
0.2	0.3	0.5	81.2
0.6	0.25	0.15	80.9

Module (FRM) with RFEB and ASPP provides multi-scale contextual information, improving mIoU by 2.4% to 77.7%. The Hybrid Attention Mechanism (HAM) effectively captures inter-channel relationships and spatial dependencies, contributing a further 1.9% gain to reach 79.6%. The Bidirectional Feature Pyramid Network (BiFPN) enables efficient multi-scale feature fusion, adding 1.0% to reach 80.6%. Finally, the progressive decoder with deep supervision further refines segmentation boundaries, yielding an additional 1.0% improvement to achieve the final 81.6% mIoU, demonstrating that each component contributes meaningfully to the overall performance.

4.4.2 Attention Mechanism Comparison

To validate our hybrid attention design, we compare different attention mechanisms as shown in Table 2. Channel attention alone achieves 78.4% mIoU by modeling inter-channel relationships. Spatial attention achieves 78.9% by capturing spatial dependencies. Coordinate attention provides 79.3% by encoding positional information. Our hybrid attention mechanism, which integrates all three attention types, achieves the highest performance at 80.5% mIoU, demonstrating the complementary nature of different attention mechanisms for metallic surface defect detection. Note that this comparison is conducted with FRM enabled but without BiFPN or deep supervision, isolating the contribution of the attention module alone; the corresponding result in Table 1 (79.6%) reflects the same setting within the full ablation protocol.

4.4.3 Deep Supervision Analysis

We analyze the impact of deep supervision and its weight configuration in Table 3. Without deep supervision (DS), the model achieves 80.6% mIoU. Adding DS with equal weights improves performance to 81.1%. We empirically find that assigning higher weights to deeper layers yields better results, as they produce more refined features. Our configuration with $\lambda_1 = 0.4$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$ achieves the best performance at 81.6% mIoU. Assigning excessive weight to shallow layers ($\lambda_1 = 0.6$) degrades performance to 80.9%, confirming that deeper features are more discriminative for defect segmentation.

4.5 Comparison with State-of-the-Art Methods

To demonstrate the effectiveness of FocusNet, we conduct comprehensive comparisons with 16 state-of-the-art segmentation methods on the MT-Defect dataset. The compared methods span various architectural paradigms including pure CNN-based approaches (UNet-2022 [24]), hybrid CNN-Transformer architectures (SegNeXt [65],

Table 4. Quantitative comparison of different detection methods on the MT-Defect dataset.

Method	Venue	Year	Defect Categories (IoU)						mIoU (%)
			Background	Blowhole	Break	Crack	Fray	Uneven	
SegNeXt	NeurIPS	2022	96.1	39.5	57.6	16.7	38.2	67.8	52.7
DDRNet	TITP	2022	98.2	59.4	74.6	54.9	93.3	81.1	76.9
UniFormer	ICLR	2022	97.4	6.1	53.7	0.0	80.7	74.8	52.1
IFormer	NeurIPS	2022	96.3	31.4	71.3	4.5	86.8	56.7	57.9
Swin-UNet	ECCV	2022	93.6	25.4	26.1	30.8	55.7	30.5	43.7
MISSFormer	TMI	2023	91.1	35.6	25.0	30.3	31.5	9.0	37.1
PIDNet-M	CVPR	2023	98.2	64.3	76.9	56.1	92.1	79.9	77.9
UNet-2022	MICAD	2023	97.9	58.2	71.8	57.6	89.9	78.3	75.6
CFDANet	JIM	2024	98.1	62.4	76.6	55.2	93.0	79.6	77.5
Segformer	NeurIPS	2021	96.7	54.2	60.2	41.4	80.6	63.5	66.1
TopFormer	CVPR	2022	98.1	53.3	65.9	44.1	88.6	81.4	71.9
SeaFormer	ICLR	2023	98.0	41.6	65.9	24.1	90.3	80.2	66.7
AFFormer	AAAI	2023	97.8	50.7	64.6	42.3	88.5	77.6	70.2
PIDNet-S	CVPR	2023	98.0	58.8	66.9	51.1	88.9	79.7	73.9
LEFormer	ICASSP	2024	96.9	54.2	66.3	50.5	88.6	65.2	70.3
LGGFormer	AEI	2025	98.2	70.6	72.4	64.9	92.9	80.6	80.0
FocusNet (Ours)	TSCC	2026	**98.4**	**71.4**	**74.3**	**70.7**	**93.2**	**81.3**	**81.6**

DDRNet [66], PIDNet [67]), pure Transformer-based models (SegFormer [44], Swin-UNet [51], MISSFormer [50]), and efficient or general-purpose segmentation methods (LEFormer [68], AFFormer [69], CFDANet [70], UniFormer [71], IFormer [72], SeaFormer [73], TopFormer [74], and LGGFormer [52]).

4.5.1 Quantitative Results

Table 4 presents the quantitative comparison on the MT-Defect dataset. FocusNet achieves 81.6% mIoU, outperforming all compared methods including recent state-of-the-art LGGFormer (80.0%) and SegFormer (66.1%). Examining per-category results, FocusNet achieves the highest IoU across five of the six defect classes: Background (98.4%), Blowhole (71.4%), Break (74.3%), Crack (70.7%), and Uneven (81.3%). The most significant gains are observed on the Crack category, where FocusNet achieves 70.7% compared to LGGFormer's 64.9% (+5.8%) and PIDNet-M's 56.1% (+14.6%), demonstrating that our progressive decoder with deep supervision captures thin elongated crack structures that other methods consistently miss. For the Blowhole category, FocusNet achieves 71.4% versus LGGFormer's 70.6% (+0.8%) and CFDANet's 62.4% (+9.0%), confirming the effectiveness of the FRM in detecting small-scale anomalies. For Break defects, FocusNet achieves 74.3% versus LGGFormer's 72.4% (+1.9%), and remains within 0.3% of DDRNet (74.6%), a general-purpose segmentation network not specifically designed for defect inspection, demonstrating strong competitiveness on this category. For Fray and Uneven, FocusNet achieves 93.2% and 81.3% respectively, matching or surpassing nearly all compared methods. The sole exception is DDRNet on the Fray category (93.3%, -0.1%), a negligible margin that does not diminish the overall competitiveness of our approach. FocusNet nonetheless surpasses CFDANet (93.0% on Fray; 79.6% on Uneven) and all other methods on Uneven (81.3% vs. DDRNet's 81.1%). While Transformer-based methods excel at capturing long-range dependencies, they often struggle with fine-grained boundary details, as evidenced by SegFormer's 66.1% and SeaFormer's 66.7% mIoU on this dataset. Against efficient architectures like TopFormer (71.9%) and SeaFormer (66.7%), FocusNet demonstrates that multi-scale feature refinement and bidirectional feature pyramid fusion effectively enhance segmentation quality across all defect categories, achieving superior or highly competitive results in every class.

4.5.2 Qualitative Analysis

Figure 3 presents qualitative results of FocusNet on representative samples from three challenging defect categories: Blowhole, Uneven, and Crack. For Blowhole defects, FocusNet successfully detects small-scale anomalies ranging from tiny pinhole-sized

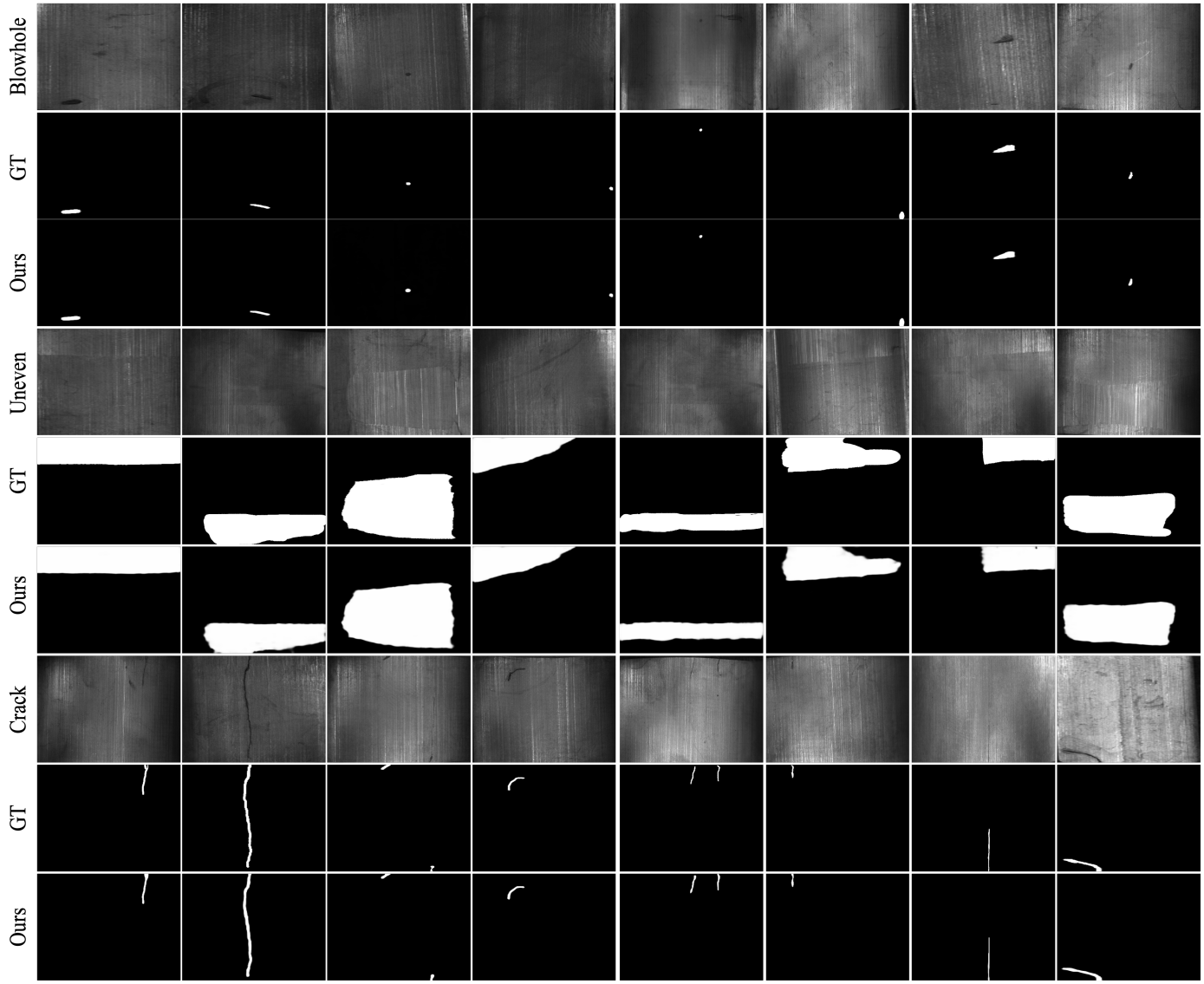

Figure 3. Visual comparison of ground truth and predicted segmentation masks on MT-Defect dataset. Top to bottom: input images, ground truth (GT), and FocusNet predictions for Blowhole, Uneven, and Crack defect categories. The results demonstrate robust multi-scale defect detection and accurate boundary localization across diverse defect characteristics.

voids to larger circular defects, with predictions closely matching ground truth masks even in complex scenes with multiple blowholes. Uneven defects, characterized by large spatial extent and irregular boundaries, are accurately segmented with smooth and continuous boundaries, demonstrating the effectiveness of our hybrid attention mechanism in distinguishing defective regions from normal surface textures. Crack defects, the most challenging category due to thin elongated structures with varying orientations, are accurately traced including vertical, diagonal, and curved patterns without fragmentation. The model preserves fine crack structures and maintains structural connectivity even for low-contrast cracks barely visible in the original images. Overall, the qualitative results confirm that FocusNet generalizes well across diverse defect types, scales, and appearances, consistently producing high-quality segmentation masks with accurate boundary localization suitable for industrial inspection applications.

5 Conclusion

This paper presented FocusNet, a novel architecture for metallic surface defect segmentation that integrates multi-scale feature refinement, hybrid attention mechanisms, and bidirectional supervision to address challenges of irregular shapes, extreme scale variations, and low-contrast backgrounds. By combining a Feature Refinement Module, a

Hybrid Attention Mechanism fusing channel, spatial, and coordinate attention, Bidirectional Feature Pyramid Network, and progressive decoder with deep supervision, FocusNet achieves robust segmentation across diverse defect characteristics. Experiments on the MT-Defect benchmark demonstrate state-of-the-art performance with 81.6% mIoU, outperforming recent methods. Ablation studies validate each component's contribution, and qualitative results confirm accurate segmentation with precise boundary delineation. Future work will explore extensions to other industrial domains, lightweight variants for resource-constrained environments, and semi-supervised learning for limited labeled data scenarios.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that Claude (Anthropic) was used for language editing and grammar refinement of the manuscript. The authors have carefully reviewed, revised, and verified the AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Tabernik, D., Šela, S., Skvarč, J., & Skočaj, D. (2020). Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing, 31*(3), 759-776. [CrossRef]

[2] Aydin, I., Akin, E., & Karakose, M. (2021). Defect classification based on deep features for railway tracks in sustainable transportation. *Applied Soft Computing, 111*, 107706. [CrossRef]

[3] Gao, Y., Lin, J., Xie, J., & Ning, Z. (2020). A real-time defect detection method for digital signal processing of industrial inspection applications. *IEEE Transactions on Industrial Informatics, 17*(5), 3450-3459. [CrossRef]

[4] Zeng, N., Wu, P., Wang, Z., Li, H., Liu, W., & Liu, X. (2022). A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-14. [CrossRef]

[5] Ye, B., Lai, J., & Xie, X. (2024). Teacher-student collaboration: Effective semi-supervised model for defect instance segmentation. *IEEE Transactions on Automation Science and Engineering, 22*, 6932-6943. [CrossRef]

[6] Ma, Y., Liu, M., Jiang, S., Wang, X., Bian, Y., & Wang, Y. (2025). Multi-context aggregation network with foreground correction for automated few-shot defect segmentation. *IEEE Transactions on Automation Science and Engineering, 22*, 13777-13787. [CrossRef]

[7] Huang, Y., Jing, J., & Wang, Z. (2021). Fabric defect segmentation method based on deep learning. *IEEE Transactions on Instrumentation and Measurement, 70*, 1-15. [CrossRef]

[8] Liu, T., He, Z., Lin, Z., Cao, G. Z., Su, W., & Xie, S. (2022). An adaptive image segmentation network for surface defect detection. *IEEE Transactions on Neural Networks and Learning Systems, 35*(6), 8510-8523. [CrossRef]

[9] Bao, Y., Song, K., Liu, J., Wang, Y., Yan, Y., Yu, H., & Li, X. (2021). Triplet-graph reasoning network for few-shot metal generic surface defect segmentation. *IEEE Transactions on Instrumentation and Measurement, 70*(5), 1-11. [CrossRef]

[10] Yu, X., Lyu, W., Zhou, D., Wang, C., & Xu, W. (2022). ES-Net: Efficient scale-aware network for tiny defect detection. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-14. [CrossRef]

[11] Pan, Y., & Zhang, L. (2022). Dual attention deep learning network for automatic steel surface defect segmentation. *Computer-Aided Civil and Infrastructure Engineering, 37*(11), 1468-1487. [CrossRef]

[12] Song, K., Feng, H., Cao, T., Cui, W., & Yan, Y. (2024). MFANet: Multifeature aggregation network for cross-granularity few-shot seamless steel tubes surface defect segmentation. *IEEE Transactions on Industrial Informatics, 20*(7), 9725-9735. [CrossRef]

[13] Liang, W., Sun, Y., Zhang, S., Bai, L., & Yang, J. (2024). SmallNet: A small defects detection network for magnetic chips based on context-weighted aggregation and feature multiscale loop fusion. *IEEE Transactions on Automation Science and Engineering, 22*, 10095-10106. [CrossRef]

[14] Ding, X., Zhang, X., Han, J., & Ding, G. (2022, June). Scaling up your kernels to 31× 31: Revisiting large kernel design in CNNs. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11953-11965). IEEE. [CrossRef]

[15] Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019, October). Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 1971-1980). IEEE. [CrossRef]

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. [CrossRef]

[18] Yao, H., Luo, W., Yu, W., Zhang, X., Qiang, Z., Luo, D., & Shi, H. (2023). Dual-attention transformer and discriminative flow for industrial visual anomaly detection. *IEEE Transactions on Automation Science and Engineering, 21*(4), 6126-6140. [CrossRef]

[19] Strudel, R., Garcia, R., Laptev, I., & Schmid, C. (2021, October). Segmenter: Transformer for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 7242-7252). IEEE. [CrossRef]

[20] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021, October). Cvt: Introducing convolutions to vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 22-31). IEEE. [CrossRef]

[21] Xu, Z., Wu, D., Yu, C., Chu, X., Sang, N., & Gao, C. (2024, March). Sctnet: Single-branch cnn with transformer semantic information for real-time segmentation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 38, No. 6, pp. 6378-6386). [CrossRef]

[22] Xia, C., Wang, X., Lv, F., Hao, X., & Shi, Y. (2024, June). Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5493-5502). IEEE. [CrossRef]

[23] Shi, D. (2024, June). Transnext: Robust foveal visual perception for vision transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 17773-17783). IEEE. [CrossRef]

[24] Guo, J., Zhou, H. Y., Wang, L., & Yu, Y. (2022, June). UNet-2022: Exploring dynamics in non-isomorphic architecture. In *International conference on medical imaging and computer-aided diagnosis* (pp. 465-476). Singapore: Springer Nature Singapore. [CrossRef]

[25] Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., & Ye, Q. (2021). Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 367-376).

[26] Lou, M., Zhang, S., Zhou, H. Y., Yang, S., Wu, C., & Yu, Y. (2025). TransXNet: learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *IEEE Transactions on Neural Networks and Learning Systems, 36*(6), 11534-11547. [CrossRef]

[27] Xu, G., Jia, W., Wu, T., Chen, L., & Gao, G. (2024). HAFormer: Unleashing the power of hierarchy-aware features for lightweight semantic segmentation. *IEEE Transactions on Image Processing, 33*, 4202-4214. [CrossRef]

[28] Shen, X., Liu, J., Zhang, H., Jiang, L., Zhao, H., & Yang, H. (2024). A novel incremental defect detection method via elastic heterogeneous distillation network. *IEEE Transactions on Automation Science and Engineering, 22*, 10149-10161. [CrossRef]

[29] Zhang, Y., Wu, J., Li, Q., Zhao, X., & Tan, M. (2021). Beyond crack: Fine-grained pavement defect segmentation using three-stream neural networks. *IEEE Transactions on Intelligent Transportation Systems, 23*(9), 14820-14832. [CrossRef]

[30] Liu, T., & He, Z. (2022). TAS 2-Net: Triple-attention semantic segmentation network for small surface defect detection. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-12. [CrossRef]

[31] Zhang, J., Ding, R., Ban, M., & Guo, T. (2022, May). FDSNeT: An accurate real-time surface defect segmentation network. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3803-3807). IEEE. [CrossRef]

[32] Yeung, C. C., & Lam, K. M. (2023). Attentive boundary-aware fusion for defect semantic segmentation using transformer. *IEEE Transactions on Instrumentation and Measurement, 72*, 1-13. [CrossRef]

[33] Zuo, L., Xiao, H., Wen, L., & Gao, L. (2023). A pixel-level segmentation convolutional neural network based on global and local feature fusion for surface defect detection. *IEEE Transactions on Instrumentation and Measurement, 72*, 1-10. [CrossRef]

[34] Yin, Z., Qin, L., Han, G., Shi, X., Zhang, F., Xu, G., & Bi, Y. (2024). Ddsnet: Deep dual-branch networks for surface defect segmentation. *IEEE Transactions on Instrumentation and Measurement, 73*, 1-16. [CrossRef]

[35] Cao, J., Yang, G., & Yang, X. (2020). A pixel-level segmentation convolutional neural network based on deep feature fusion for surface defect detection. *IEEE Transactions on Instrumentation and Measurement, 70*, 1-12. [CrossRef]

[36] Qiu, Y., Liu, H., Liu, J., Shi, B., & Li, Y. (2024). Region and edge-aware network for rail surface defect segmentation. *IEEE Transactions on Instrumentation and Measurement, 73*, 1-13. [CrossRef]

[37] Wang, C., Chen, H., & Zhao, S. (2023). RERN: Rich edge features refinement detection network for polycrystalline solar cell defect segmentation. *IEEE Transactions on Industrial Informatics, 20*(2), 1408-1419. [CrossRef]

[38] Yang, L., Xu, S., Fan, J., Li, E., & Liu, Y. (2023). A pixel-level deep segmentation network for automatic defect detection. *Expert Systems with Applications, 215*, 119388. [CrossRef]

[39] Yang, H., Hu, J., Yin, Z., & Wang, Z. (2022). A semantic information decomposition network for accurate

segmentation of texture defects. *IEEE Transactions on Industrial Informatics, 19*(7), 8319-8327. [CrossRef]

[40] Yu, R., Guo, B., & Yang, K. (2022). Selective prototype network for few-shot metal surface defect segmentation. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-10. [CrossRef]

[41] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., ... & Shao, L. (2021, October). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF international conference on computer vision (ICCV)* (pp. 548-558). IEEE. [CrossRef]

[42] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., ... & Zhang, L. (2021, June). 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). IEEE. [CrossRef]

[43] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021, October). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)* (pp. 9992-10002). Ieee. [CrossRef]

[44] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems, 34*, 12077-12090.

[45] Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., ... & Guo, B. (2022, June). Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12114-12124). IEEE. [CrossRef]

[46] Pan, X., Ye, T., Xia, Z., Song, S., & Huang, G. (2023, June). Slide-transformer: Hierarchical vision transformer with local self-attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2082-2091). IEEE. [CrossRef]

[47] Zhu, W., Zhang, H., Zhang, C., Zhu, X., Guan, Z., & Jia, J. (2023). Surface defect detection and classification of steel using an efficient Swin Transformer. *Advanced Engineering Informatics, 57*, 102061. [CrossRef]

[48] Wu, Y. H., Liu, Y., Zhan, X., & Cheng, M. M. (2022). P2T: Pyramid pooling transformer for scene understanding. *IEEE transactions on pattern analysis and machine intelligence, 45*(11), 12760-12771. [CrossRef]

[49] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., ... & Shao, L. (2022). Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media, 8*(3), 415-424. [CrossRef]

[50] Shang, H., Sun, C., Liu, J., Chen, X., & Yan, R. (2023). Defect-aware transformer network for intelligent visual surface defect detection. *Advanced Engineering Informatics, 55*, 101882. [CrossRef]

[51] Zhou, H., Yang, R., Hu, R., Shu, C., Tang, X., & Li, X. (2023). ETDNet: Efficient transformer-based detection network for surface defect detection. *IEEE transactions on instrumentation and measurement, 72*, 1-14. [CrossRef]

[52] Zhang, G., Lu, Y., Jiang, X., Jin, S., Li, S., & Xu, M. (2025). LGGFormer: A dual-branch local-guided global self-attention network for surface defect segmentation. *Advanced Engineering Informatics, 64*, 103099. [CrossRef]

[53] Jiang, X., Guo, K., Lu, Y., Yan, F., Liu, H., Cao, J., ... & Tao, D. (2023). CINFormer: Transformer network with multi-stage CNN feature injection for surface defect segmentation. *arXiv preprint arXiv:2309.12639*. [CrossRef]

[54] Zhang, D., Song, K., Xu, J., He, Y., Niu, M., & Yan, Y. (2020). MCnet: Multiple context information segmentation network of no-service rail surface defects. *IEEE Transactions on Instrumentation and Measurement, 70*, 1-9. [CrossRef]

[55] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PmLR.

[56] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). IEEE. [CrossRef]

[57] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence, 40*(4), 834-848. [CrossRef]

[58] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141). [CrossRef]

[59] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September). Cbam: Convolutional block attention module. In *European conference on computer vision* (pp. 3-19). Cham: Springer International Publishing. [CrossRef]

[60] Hou, Q., Zhou, D., & Feng, J. (2021, June). Coordinate attention for efficient mobile network design. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 13708-13717). IEEE. [CrossRef]

[61] Huang, Y., Qiu, C., & Yuan, K. (2020). Surface defect saliency of magnetic. *The Visual Computer, 36*(1), 85-96. [CrossRef]

[62] Zhang, S., Jiang, L., Ge, L., Wu, C., Wang, Y., Lu, K., & Zhang, C. (2025). SGTP-Net: semantic guidance and texture priors-based dual-branch segmentation network for surface defect detection. *IEEE Transactions*

on Automation Science and Engineering, 23, 417-432. [CrossRef]

[63] Dong, H., Song, K., He, Y., Xu, J., Yan, Y., & Meng, Q. (2019). PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection. *IEEE Transactions on Industrial Informatics, 16*(12), 7448-7458. [CrossRef]

[64] Zhou, X., Fang, H., Liu, Z., Zheng, B., Sun, Y., Zhang, J., & Yan, C. (2021). Dense attention-guided cascaded network for salient object detection of strip steel surface defects. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-14. [CrossRef]

[65] Guo, M. H., Lu, C. Z., Hou, Q., Liu, Z., Cheng, M. M., & Hu, S. M. (2022). Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in neural information processing systems, 35*, 1140-1156.

[66] Hong, Y., Pan, H., Sun, W., & Jia, Y. (2021). Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv preprint arXiv:2101.06085*. [CrossRef]

[67] Xu, J., Xiong, Z., & Bhattacharyya, S. P. (2023, June). PIDNet: A real-time semantic segmentation network inspired by PID controllers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 19529-19539). IEEE. [CrossRef]

[68] Xie, L., Xiang, X., Xu, H., Wang, L., Lin, L., & Yin, G. (2020). FFCNN: A deep neural network for surface defect detection of magnetic tile. *IEEE Transactions on Industrial Electronics, 68*(4), 3506-3516. [CrossRef]

[69] Dong, B., Wang, P., & Wang, F. (2023, June). Head-free lightweight semantic segmentation with linear transformer. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, No. 1, pp. 516-524). [CrossRef]

[70] Ma, S., Song, K., Niu, M., Tian, H., & Yan, Y. (2024). Cross-scale fusion and domain adversarial network for generalizable rail surface defect segmentation on unseen datasets. *Journal of Intelligent Manufacturing, 35*(1), 367-386. [CrossRef]

[71] Li, K., Wang, Y., Gao, P., Song, G., Liu, Y., Li, H., & Qiao, Y. (2022). Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*. [CrossRef]

[72] Si, C., Yu, W., Zhou, P., Zhou, Y., Wang, X., & Yan, S. (2022). Inception transformer. *Advances in neural information processing systems, 35*, 23495-23509.

[73] Wan, Q., Huang, Z., Lu, J., Yu, G., & Zhang, L. (2025). SeaFormer++: Squeeze-enhanced axial transformer for mobile visual recognition. *International journal of computer vision, 133*(6), 3645-3666. [CrossRef]

[74] Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., ... & Shen, C. (2022, June). Topformer: Token pyramid transformer for mobile semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12073-12083). IEEE. [CrossRef]

Muhammad Hamza Ali is a Master's candidate in Computer Science at the Virtual University of Pakistan. His academic interests lie in computer vision, machine learning, and artificial intelligence. He has been involved in multiple research-oriented activities, with a strong focus on dataset preparation, data acquisition, and annotation, contributing to the development of datasets for vision-based and AI-driven applications. (Email: hamzaali441809@gmail.com)

Farhan Ali is a graduate student currently pursuing a master's in computer science with a specialization in Data Science at Technische Universität Graz, Austria. He has a strong academic background and hands-on experience in data science, machine learning, and computer vision. He worked as an Associate Software Engineer at OpusAI, where he was involved in building user-centric web applications. In addition, he completed data science internships with Oasis Infobyte and Info AidTech, respectively, gaining valuable experience. (Email: farhan.ali@student.tugraz.at)

Zaid Muhammad is a Computer Science student with a strong foundation in Python programming and computational problem-solving, complemented by experience in graphic and visual system design. His work focuses on developing intelligent, efficient, and visually robust digital solutions. His research interests include artificial intelligence, machine learning, and technology-driven system design. (Email: zaidmuhammadsaid14@gmail.com)