

Molten mark classification using multi-scale directional cross-scale attention fusion

Taehi Kim

Korea University

Jonghwa Shim

Korea University

Eenjun Hwang

ehwang04@korea.ac.kr

Korea University

Article

Keywords:

Posted Date: August 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10442322/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on September 11th, 2026. See the published version at <https://doi.org/10.1038/s41598-026-70819-5>.

Molten mark classification using multi-scale directional cross-scale attention fusion

Taehi Kim¹, Jonghwa Shim¹ and Eenjun Hwang^{1,*}

¹Korea University, School of Electrical Engineering, Seoul, 02841, Korea

*corresponding.ehwang04@korea.ac.kr

ABSTRACT

In electrical fire investigation, it is critical to determine whether molten marks on wires were caused by a short circuit or by external heat. Existing analytical methods require substantial manual work and expensive equipment for preprocessing and classification. Recent convolutional neural network (CNN)-based approaches have demonstrated relatively good performance; however, their accuracy is limited by the insufficient utilization of various features of molten marks. To address this, we propose a molten mark classification method based on multi-scale directional cross-scale attention fusion. We first extract feature maps at four resolutions (scales) using a Swin transformer. These features are then fused using a bi-directional feature pyramid network neck integrated with directional cross-scale attention modules for channel and spatial attention, conditioned on features from different scales. The fused features are passed to a classification head for final classification. For performance evaluation, we conducted extensive experiments with five comparison models. The proposed method achieved an F1 score of 0.9613 and a Matthews correlation coefficient of 0.9257, improvements of 2.26 and 4.02 percentage points over a state-of-the-art model. Additionally, the proposed method exhibits only a 2.88% accuracy decrease on degraded images, demonstrating superior robustness to CNN-based models, which decrease by 15.75% on average.

1. Introduction

Investigating and identifying the causes of electrical fires is essential for attributing responsibility in accident scenarios and preventing future fires. Most molten marks observed at the scenes of electrical fires are either short-circuit beads caused by electric arcs or heat marks caused by external flames. Modern fire investigation techniques emphasize the identification of short circuits in electrical wiring; therefore, the accurate classification of molten marks is essential for identifying the causes of fires¹⁻³. Conventional methods for identifying molten marks rely on the metallographic analysis of wire cross-sections. However, this approach requires expensive equipment and substantial manual labor for preprocessing tasks, such as cutting and polishing, and also requires a significant amount of time for analysis⁴.

Recently, there have been attempts to train deep learning models that can classify molten marks automatically. For example, Batool and Bang⁵ applied transfer learning to pre-trained convolutional neural network (CNN) models and achieved 92.7% accuracy for molten mark classification, demonstrating the feasibility of automated classification. CNN models are specialized for extracting local features using fixed-size convolutional kernels and local receptive fields⁶. However, molten marks typically possess distinct local and global features. For example, short-circuit beads and heat marks may have similar shapes, but their surface gloss characteristics can differ significantly; conversely, their local textures may be similar, while the overall boundary sharpness may vary⁷⁻⁹. Therefore, to enhance classification performance, it is necessary to consider both local and global features. Furthermore, molten marks collected from real fire scenes often exhibit unclear local features owing to various factors. Ultimately, CNN-based models that utilize fixed receptive fields struggle to classify molten marks for these reasons^{6,10}.

To address these limitations, we propose a novel neural network architecture for molten mark classification that can effectively capture both local and global features through multi-scale directional cross-scale attention fusion. We first extract hierarchical visual features at four distinct resolution levels using a Swin transformer¹¹ as a backbone. We then fuse these features by revising the bi-directional feature pyramid network (BiFPN)¹², a well-established module for cross-scale feature aggregation. Because typical BiFPNs apply a uniform scalar weight across all pixels and channels of the features at each scale during fusion, they may fail to capture the subtle characteristics of molten marks, such as variations in boundary sharpness and surface glossiness. Therefore, we introduce a directional cross-scale attention (DCSA) module into each fusion node of the BiFPN. DCSA calculates attention weights representing

the importance of different channels and spatial regions for features at different resolutions. Finally, the binary classification of molten marks is performed by passing the fused features to a classification head consisting of a global pooling layer and two fully connected layers. By using DCSA, the proposed method can selectively emphasize classification-relevant features across multiple scales, thereby amplifying critical details that may be prominent at one scale but subtle or absent at another scale.

Extensive experiments on a molten mark dataset consisting of both laboratory and real fire scene data demonstrate that the proposed method outperforms a wide range of baseline models on various classification metrics by effectively learning the features of short-circuit beads and heat marks. Additionally, ablation studies reveal that each component of the proposed method makes a meaningful contribution to overall performance.

The main contributions of this paper are summarized as follows:

- We propose a novel molten mark classification method utilizing a Swin transformer and BiFPN-based multi-scale feature extraction and fusion. To the best of our knowledge, this is the first study to apply multi-scale feature fusion to molten mark classification.
- We propose the selective emphasis of channels and spatial regions for effective classification by replacing the weighting function of the fusion nodes in a BiFPN with the DCSA module.
- We enhance the practicality and robustness of model training by augmenting an existing training dataset of laboratory-generated samples with molten mark samples collected from actual electrical fire scenes.
- We verify the effectiveness of the proposed method through extensive comparative experiments and ablation studies.

The remainder of this paper is organized as follows. Section 2 introduces related works on molten mark classification, including the Swin transformer. Section 3 presents the overall structure and key components of the proposed method. Section 4 describes our experimental setup, and Section 5 presents experimental results. Finally, Section 6 concludes this paper and outlines promising directions for future research.

2. Related Work

This section summarizes existing research on molten mark classification and multi-scale feature learning in the field of computer vision.

2.1 Deep-Learning-Based Classification

Early deep-learning-based approaches for molten mark classification utilized metal structure images of wire cross-sections as model inputs. For example, Shi et al.¹³ segmented the molten regions in metallographic images of electric wires using TransUNet¹⁴, constructed feature vectors using conventional feature extraction techniques, and then performed classification using the XGBoost algorithm¹⁵. Although they achieved a promising F1 score of 82.9%, their method relies on manually engineered features, and its classification performance is sensitive to segmentation quality and feature extraction accuracy. Furthermore, this method requires costly preprocessing operations, such as wire cutting and polishing, limiting its applicability in the field.

Several attempts have been made to perform classification by learning the external features of molten marks. For example, Batool and Bang⁵ collected 1,734 images of short-circuit beads produced in a laboratory and expanded this dataset by a factor of 20 through data augmentation. They applied transfer learning to several pre-trained CNN models to classify primary and secondary short-circuit beads. Among the tested CNN models, VGG16¹⁶ achieved the best performance, with a validation accuracy of 92.7%, demonstrating that external images alone are sufficient for classifying detailed subtypes of short-circuit beads. However, because CNNs learn features only within fixed local receptive fields, they struggle to capture global context. This issue poses a limitation for molten mark classification because accurate classification requires the integrated analysis of global and local features^{6,10}.

2.2 Multi-Scale Feature Extraction

To analyze both the local and global features of molten marks effectively, features at various resolutions should be considered simultaneously. The Swin transformer proposed by Liu et al.¹¹ generates multi-scale feature maps at four

resolution levels by applying shifted-window-based self-attention within a hierarchical architecture. This hierarchical design produces feature maps at progressively lower resolutions, enabling it to capture fine features in small regions in the initial stage and global features in increasingly larger regions in subsequent stages. This multi-scale representation has proven effective across diverse vision tasks, including classification, detection, and segmentation^{17,18}. However, typical classification approaches relying solely on final-stage features discard the fine-grained features captured at higher resolutions, thereby limiting the ability to leverage the complementary multi-scale features required for molten mark classification^{12,19,20}.

Lin et al.²⁰ proposed the FPN, a seminal feature pyramid architecture that fuses multi-scale backbone features through a top-down pathway, enabling high-resolution feature maps to be enriched with features from lower-resolution stages. Tan et al.¹² proposed the BiFPN, which extends the original FPN architecture by adding a bottom-up pathway to enable bidirectional information flow and by incorporating learnable scalar weights into each fusion node. However, they assigned a single scalar weight to all feature maps at each scale, treating all channels and spatial locations equally during fusion, which prevents the network from selectively emphasizing the channels or spatial regions most informative for classification.

To address this issue, several studies have integrated standard channel and spatial attention modules into BiFPNs across various domains. For example, for medical image segmentation and object detection, convolutional block attention modules (CBAMs)²¹ have been incorporated into BiFPNs to jointly refine fused features along the channel and spatial dimensions^{22,23}. However, CBAMs recalibrate each fused output independently at a single resolution level and cannot model cross-scale dependencies at fusion nodes; therefore, they cannot determine how the features from one resolution level should selectively inform the features at another level during the fusion process.

To the best of our knowledge, no previous study has applied a Swin transformer and BiFPN-based multi-scale feature fusion to molten mark classification, and existing BiFPN models do not address the cross-scale interactions inherent to multi-scale fusion nodes. To address these issues, we propose the DCSA module, which introduces direction-aware channel and spatial attention into each fusion node of a BiFPN.

3. Proposed Method

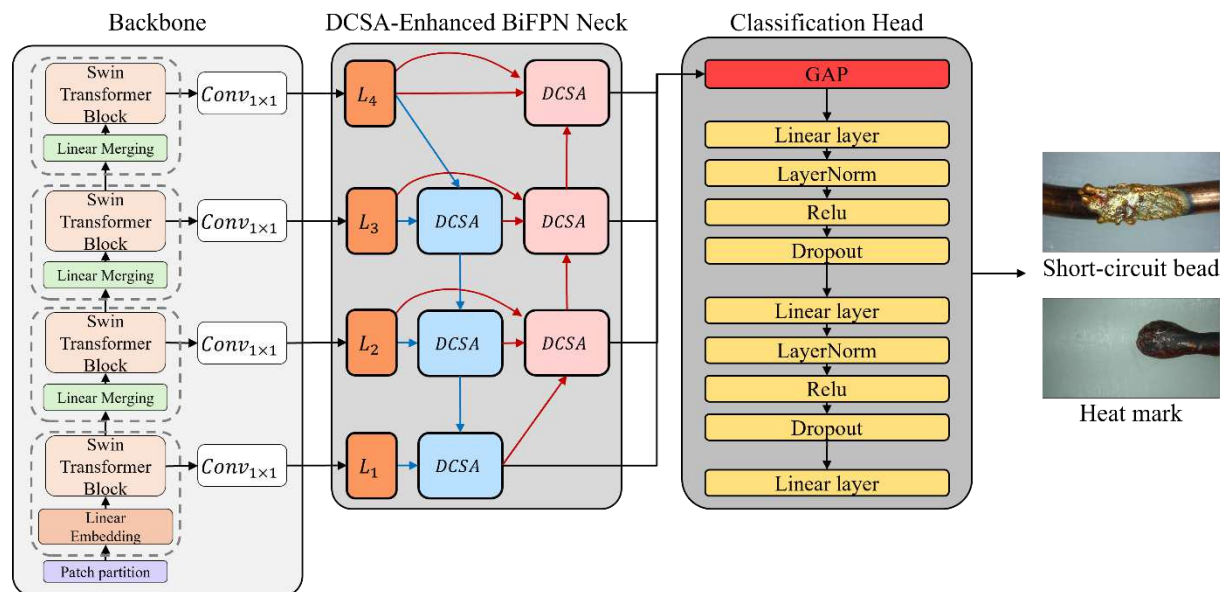


Figure 1 Overall architecture of the proposed method. The backbone extracts multi-scale feature maps, the neck fuses these maps through bidirectional pathways, and the classification head produces final predictions.

3.1 Overall architecture

Figure 1 illustrates the overall architecture of the proposed molten mark classification system. The system consists of

three main components: a backbone network, a BiFPN neck, and a classification head. First, the backbone network generates four feature maps for an input image. These feature maps are then fed into the DCSA-enhanced BiFPN neck, where a top-down pathway propagates global context from low-resolution to high-resolution levels, and a bottom-up pathway subsequently propagates local details from high-resolution to low-resolution levels. The DCSA module in each fusion node adaptively emphasizes discriminative features along both the channel and spatial dimensions, which are aggregated by the classification head to perform binary classification.

3.2 Backbone

The proposed system utilizes a Swin transformer as its backbone for hierarchical multi-scale feature extraction from input images. Given an input image $x \in \mathbb{R}^{H \times W \times 3}$, the four stages of the Swin transformer produce feature maps C_1, C_2, C_3 , and C_4 with progressively decreasing spatial resolution. C_1 possesses the highest spatial resolution, allowing it to encode rich local textural features, whereas C_4 captures global features at the lowest resolution, including the overall molten mark morphology. Before being fed into the BiFPN neck, these hierarchical feature maps are projected to a common channel dimensionality $d = 256$ via 1×1 convolutions as follows:

$$L_i = \text{Conv}_{1 \times 1}(C_i), i \in 1, 2, 3, 4. \quad (1)$$

This projection unifies the channel dimensions, thereby enabling cross-scale feature fusion in the BiFPN.

3.3 DCSA-Enhanced BiFPN Neck

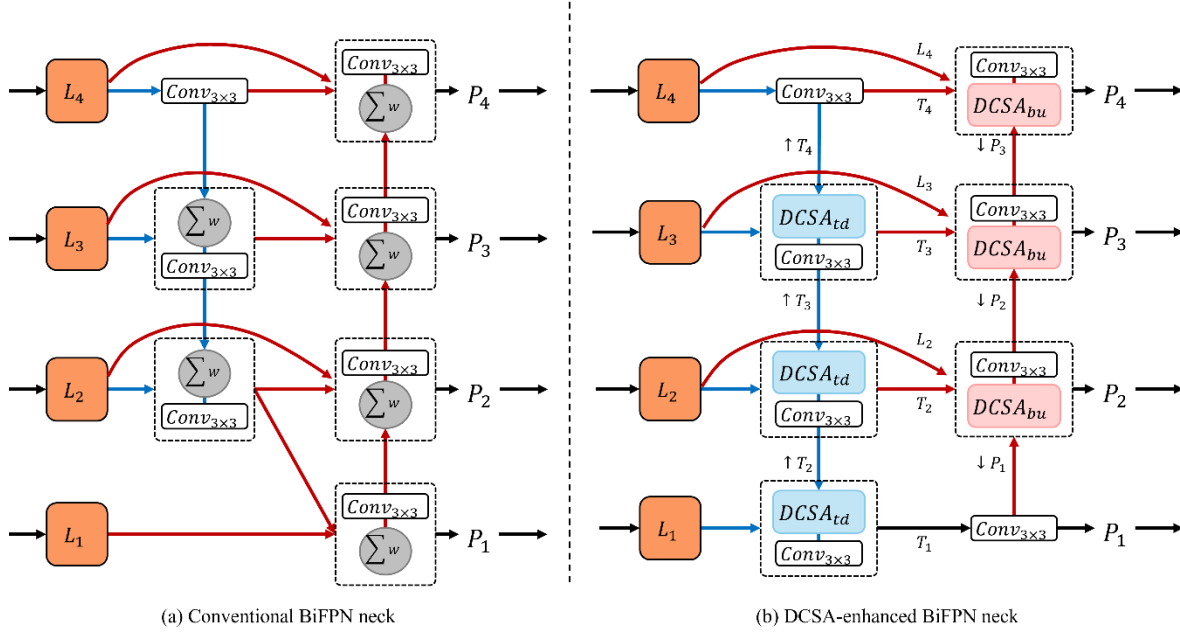


Figure 2 Comparison of the (a) conventional BiFPN with scalar weighted summation and (b) DCSA-enhanced BiFPN neck.

A conventional BiFPN fuses multi-scale features through top-down and bottom-up bidirectional pathways; however, a single scalar weight is assigned to all feature maps in each fusion node, and the maps are aggregated via summation. Although this approach can adjust the extent to which each scale contributes to fusion, it cannot selectively emphasize classification-relevant information along the channel and spatial dimensions of each feature. For example, if the surface glossiness of a molten mark is encoded in specific channels, and the boundary between molten and non-molten regions is encoded in specific spatial locations, then the inability to selectively highlight such discriminative cues at the channel and spatial levels may limit classification performance. In contrast, the proposed DCSA module can selectively highlight classification-relevant channels and spatial regions by leveraging the relationships between features at various scales and applying channel attention and spatial attention sequentially. Figure 2 compares the structure of the conventional BiFPN neck with the proposed DCSA-enhanced BiFPN neck. DCSA employs distinct

fusion strategies depending on the pathway direction. In the following description of our fusion strategy, $\text{Conv}_{3 \times 3}$ denotes a block consisting of a 3×3 convolution, batch normalization, and rectified linear unit (ReLU) activation, whereas $\psi(\cdot)$ denotes a block consisting of two 1×1 convolutions ($C \rightarrow C/r \rightarrow 1$, where r is the reduction ratio), batch normalization, and ReLU, which produces a single-channel spatial map from an input feature map.

3.3.1 Top-Down Pathway

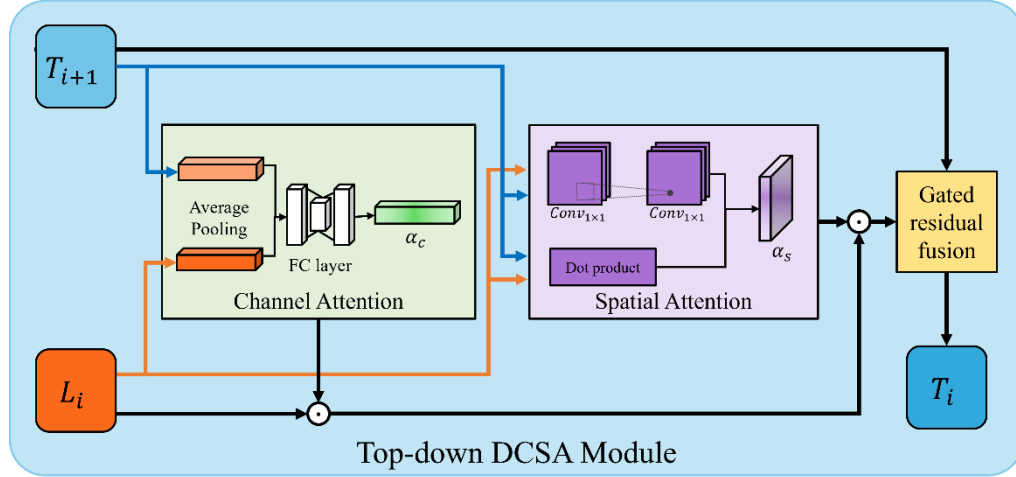


Figure 3 Detailed architecture of the binary fusion DCSA module for the top-down pathway: overall structure of top-down DCSA fusion, channel attention, and spatial attention.

The top-down pathway begins at L_4 and sequentially propagates global context toward finer scales. The starting point is defined as

$$T_4 = \text{Conv}_{3 \times 3}(L_4). \quad (2)$$

At each scale level $i \in \{3, 2, 1\}$, the DCSA_{td} module receives two inputs, namely a two-times upsampled coarser-scale feature $\uparrow T_{i+1}$ and the current-scale backbone feature L_i , and performs fusion as follows:

$$T_i = \text{Conv}_{3 \times 3}(\text{DCSA}_{td}(\uparrow T_{i+1}, L_i)). \quad (3)$$

Consequently, the final output T_1 of the top-down pathway contains both the global context propagated from coarser-scale levels and the local details captured at finer levels.

Figure 3 illustrates the architecture of the DCSA_{td} module, which consists of three stages for channel attention, spatial attention, and gated residual fusion. Because the global context propagated through the top-down pathway may not always be beneficial at the current scale, the final stage employs sigmoid-based gated residual fusion to control the propagation intensity of global context adaptively.

Channel attention is implemented by applying global average pooling (GAP) to the two inputs $\uparrow T_{i+1}$ and L_i , and the resulting vectors are concatenated and fed into a shared fully connected (FC) network, which generates a channel attention weight α_c for L_i as follows:

$$\alpha_c = \sigma(\text{FC}([\text{GAP}(\uparrow T_{i+1}); \text{GAP}(L_i)])), \quad (4)$$

$$L_i^c = L_i \odot \alpha_c. \quad (5)$$

Here, $[\cdot; \cdot]$ denotes concatenation along the channel dimension, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The weight α_c is intended to emphasize channels that are useful for molten mark classification selectively by recalibrating each channel of the current-scale L_i based on the global context from coarser scales. For example, the global morphological information encoded in L_4 may emphasize channels in L_3 related to the sharpness of the boundary between molten and non-molten regions.

The cross-scale similarity at each spatial location is calculated through the scaled element-wise dot product of $\uparrow T_{i+1}$ and L_i , and then combined with the spatial map produced by applying ψ to $\uparrow T_{i+1}$ to generate a spatial attention weight α_s as follows:

$$\alpha_s = \sigma \left(\langle \uparrow T_{i+1}, L_i \rangle \cdot C^{-1/2} + \psi(\uparrow T_{i+1}) \right), \quad (6)$$

$$L_i^{cs} = L_i^c \odot \alpha_s. \quad (7)$$

Here, $\langle \cdot, \cdot \rangle$ denotes the element-wise dot product along the channel axis, and $C^{-1/2}$ is a scaling factor. The dot-product term captures the spatial correspondence between the two scales. The ψ term identifies spatial locations in which classification cues are concentrated in $\uparrow T_{i+1}$ to compensate for locations that the dot-product alone may miss as a result of weak activation at the current scale. Through the combination of these two terms, the proposed module can selectively emphasize the most informative spatial locations.

In the final stage, the channel- and spatial-attended feature L_i^{cs} and the coarser-scale feature $\uparrow T_{i+1}$ are combined adaptively through a gating mechanism as follows:

$$\lambda = \sigma \left(\text{mean}_c(\text{GAP}(L_i^{cs})) \right), \quad (8)$$

$$T_i = \text{Conv}_{3 \times 3}(\lambda \cdot L_i^{cs} + (1 - \lambda) \cdot \uparrow T_{i+1}), \quad (9)$$

where mean_c denotes the mean over the channel dimension, which converts the GAP output vector into a single scalar. The gate value λ is determined by the overall activation magnitude of L_i^{cs} . When the current-scale feature following channel and spatial attention exhibits strong activation, λ increases dynamically to incorporate this information. Otherwise, the contribution of $\uparrow T_{i+1}$ dominates, preserving the global context from coarser scales.

3.3.2 Bottom-Up Pathway

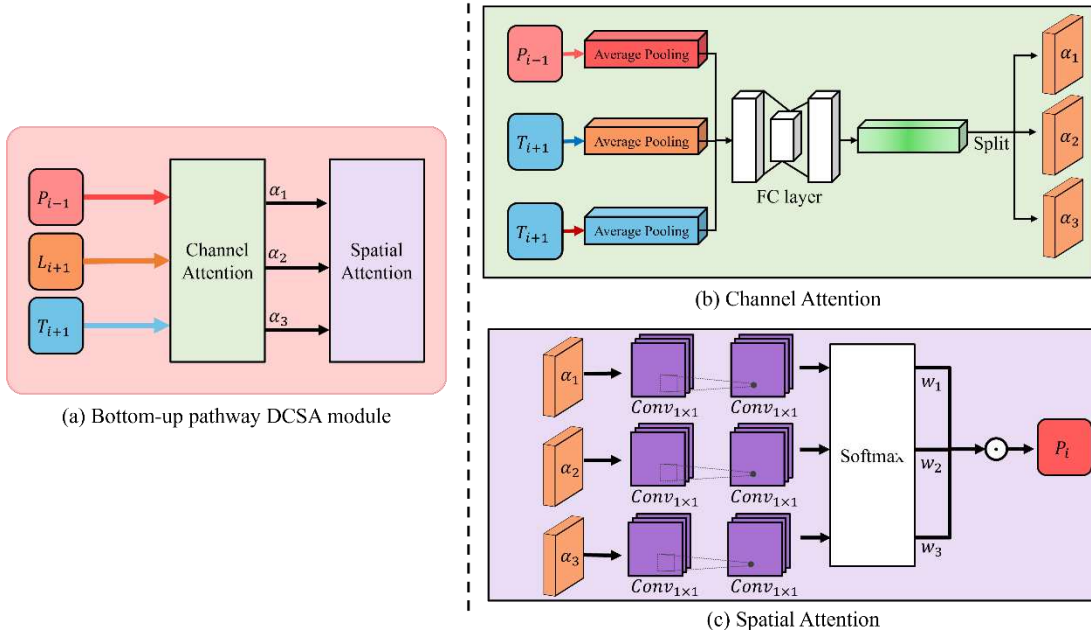


Figure 4 Detailed architecture of the ternary fusion DCSA module for the bottom-up pathway. This mechanism sequentially applies channel attention and spatial attention to perform fusion among P_{i-1} , L_i , and T_i features.

The bottom-up pathway begins at T_l and propagates local details at a fine scale toward coarser scales. The starting point is defined as

$$P_1 = \text{Conv}_{3 \times 3}(T_1). \quad (10)$$

At each scale level $i \in \{2, 3, 4\}$, the DCSA_{bu} module performs the fusion defined in Eq. (11) using three inputs as

$$P_i = \text{Conv}_{3 \times 3}(\text{DCSA}_{bu}(\downarrow P_{i-1}, L_i, T_i)), \quad (11)$$

where $\downarrow P_{i-1}$ is a downsampled bottom-up output from a finer scale, L_i is a current-scale backbone feature, and T_i is an intermediate result from the top-down pathway. These three inputs represent local details, the original feature at the current scale, and the feature augmented with global context, respectively.

Figure 4 illustrates the architecture of the DCSA_{bu} module. In the top-down pathway, the mixing of two inputs is controlled by a sigmoid gate, whereas in the bottom-up pathway, the three inputs provide different information

corresponding to the same spatial location. If independent sigmoid gates are applied to each of the three inputs, then the sum of their contributions may exceed one, leading to redundant feature accumulation. Therefore, SoftMax normalization is applied to select the most informative input at each spatial location.

Similar to the top-down pathway, GAP is applied to each of the three inputs, the resulting vectors are concatenated, and an independent channel attention weight is generated for each input by a shared FC network as follows:

$$\alpha_1, \alpha_2, \alpha_3 = \text{Split} \left(\sigma \left(FC \left([GAP(\downarrow P_{i-1}); GAP(L_i); GAP(T_i)] \right) \right) \right), \quad (12)$$

$$\downarrow P_{i-1}^c = \alpha_1 \odot \downarrow P_{i-1}, \quad L_i^c = \alpha_2 \odot L_i, \quad T_i^c = \alpha_3 \odot T_i, \quad (13)$$

where α_1 , α_2 , and α_3 selectively emphasize channels that are useful for classification at finer scales, the current scale, and coarser scales, respectively, augmented with global context. *Split* denotes an operation that separates the FC output into the channel attention weights corresponding to each of the three inputs. Because the FC network jointly considers the channel statistics of all three inputs, the generated weights reflect inter-input channel dependencies.

A shared ψ is applied to each channel-attended feature map to produce single-channel spatial maps, which are then normalized via SoftMax to determine the relative contributions of the three inputs at each spatial location as follows:

$$w_1, w_2, w_3 = \text{Softmax}(\psi(\downarrow T_{i-1}'), \psi(L_i'), \psi(T_i')), \quad (14)$$

$$P_i = w_1 \odot \downarrow P_{i-1}^c + w_2 \odot L_i^c + w_3 \odot T_i^c. \quad (15)$$

The ψ applied to the three inputs shares the same weights, and because SoftMax normalization enforces $w_1 + w_2 + w_3 = 1$ at each spatial location, the most discriminative input at each location is selected preferentially.

3.4 Classification Head

To derive a feature vector for each scale, GAP is applied to the four feature maps $\{P_1, P_2, P_3, P_4\}$ obtained from the bidirectional pathways of the BiFPN neck. The resulting features are then concatenated to form a unified representation vector as follows:

$$z = [GAP(P_1); GAP(P_2); GAP(P_3); GAP(P_4)]. \quad (16)$$

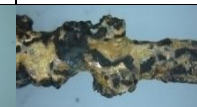
The concatenated vector $z \in \mathbb{R}^{1024}$ is passed through a fusion network consisting of two FC layers ($1024 \rightarrow 512 \rightarrow 256$), each of which is followed by layer normalization and ReLU activation, with dropout ($p = 0.3$) being applied after the first FC layer. The resulting 256-dimensional representation is fed into a linear classifier to perform the binary classification of short-circuit beads and heat marks.

4. Experimental Settings

This section describes the experimental setup, including the molten mark dataset, evaluation metrics, basic models, and implementation details, used to evaluate the performance of the proposed method.

4.1 Molten Mark Dataset

Table 1 Breakdown of the molten mark dataset.

Dataset type	Training				Testing	
Class	Short-Circuit bead		Heat mark		Short-Circuit bead	Heat mark
Image						
Source	Laboratory	Fire scene	Laboratory	Fire scene	Fire scene	Fire scene
No. of Images	6,016	2,071	7,033	3,630	2,549	2,356
Total	18,750				4,905	

In our experiments, we used a molten mark dataset provided by the Korea Electrical Safety Corporation (KESCO). The molten marks in this dataset were generated in a laboratory under controlled conditions or collected at actual fire

scenes and can be categorized into two classes: short-circuit bead and heat mark. Table 1 details the composition of the dataset.

For model training, the laboratory and field data were combined and then split into training and validation subsets. The testing set consisted exclusively of field data that were not used in the training dataset. The testing set was reserved for evaluating actual performance under realistic fire investigation conditions.

The dataset includes images of solid wires with cross-sections of 1.5 and 2.5 mm², cord wires, and stranded wires with cross-sections of 0.75, 1.0, 2.0, and 4.0 mm². The images were captured from diverse angles using digital and portable USB microscopes. When splitting the combined data into training and validation subsets, splitting was performed at the sample level to ensure that multiple images of the same wire were assigned exclusively to either the training or validation subset, thereby preventing data leakage from near-duplicate images across subsets.

Laboratory-generated short-circuit beads were produced through simulated short-circuit experiments intended to replicate the fault conditions that occur in typical residential settings. The experimental system consisted of a distribution panel, short-circuit current-limiting device, and short-circuit tester. The short-circuit current-limiting device restricted the fault current to 500 A through resistive components. When a short circuit occurs, the temperature at the fault location rises instantaneously from ambient temperature to several thousand degrees before cooling. During this period of intense heating, the characteristic globular morphology of the molten mark is formed.

Laboratory-generated heat marks were produced by applying external heat to electrical wires using a torch. Heating was terminated based on visual assessment, with the stopping criterion being the point at which the copper conductor exhibited deformation and solidification. Excessive heat application causes droplet formation as molten copper begins to drip, destroying the molten mark morphology. Therefore, the upper threshold was set at the point where melting and deformation occur without droplet detachment. The solidified samples obtained after cooling were used to capture heat mark images.

Field samples were collected from real electrical fire scenes by forensic specialists affiliated with local fire stations and the Accident Investigation Center of KESCO.

The wire samples collected from the fire scenes were heavily contaminated with foreign substances, such as melted insulation and soot, which may obstruct molten mark morphology analysis. To remove such contaminants, preprocessing was performed using an ultrasonic cleaner and compressed air prior to imaging, ensuring that only the bare copper conductor surface was exposed prior to image acquisition.

The class labels for the field samples were determined via consensus among fire investigators affiliated with the National Fire Agency and electrical fire forensic experts from KESCO to ensure the reliability of the labels.

4.2 Evaluation Metrics

To evaluate the performance of the proposed model comprehensively, we consider six evaluation metrics: accuracy, precision, recall, F1-score, MCC, and area under the curve (AUC).

Accuracy represents the proportion of correctly classified samples among the entire dataset, indicating the overall classification capabilities of a model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

Precision quantifies the fraction of samples predicted as short-circuit beads that are true positives, reflecting the model's robustness against false positives. In fire cause determination, false positives may lead to the erroneous identification of ignition sources, making this metric particularly critical.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (18)$$

Recall measures the proportion of actual short-circuit beads identified correctly, representing the model's sensitivity to false negatives. Because false negatives imply that the true fire cause is missed, this metric holds critical importance in fire investigations.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (19)$$

F1-score is the harmonic mean of precision and recall, which provides a balanced assessment of both metrics. Given the potential class imbalance in the molten mark dataset, F1-score provides a more faithful reflection of a model's practical discriminative capability compared with accuracy alone.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (20)$$

MCC is a correlation coefficient that incorporates all four entries of the confusion matrix in a balanced manner²⁴.

MCC is robust to class imbalance and ranges from -1 to $+1$, where $+1$ indicates perfect classification, 0 corresponds to random prediction, and -1 denotes complete inverse classification.

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (21)$$

AUC is a threshold-independent metric that aggregates a model’s discriminative ability across all possible classification thresholds. Because AUC evaluates class separability independently from threshold selection, this metric is well-suited to fair cross-model comparisons.

Beyond these quantitative metrics, we present confusion matrices and precision–recall (PR) curves to complement our numerical analysis with visual verifications of model performance.

4.3 Model Setup

To validate the effectiveness of the proposed model, we compared it with two CNN-based models (VGG16¹⁶ and EfficientNet²⁵) and three transformer-based models (ViT²⁶, DeiT²⁷, and the Swin transformer¹¹).

All models were initialized with weights pre-trained on ImageNet-1K²⁸, and the input images were rescaled to the resolution required by each architecture. Training was conducted using the AdamW optimizer²⁹ with a differential learning rate strategy. The pre-trained backbone was assigned a learning rate of 3×10^{-5} with a weight decay of 5×10^{-5} , whereas the newly added modules received a learning rate of 1×10^{-4} with a weight decay of 0.01 . The learning rate schedule consisted of a linear warmup over five epochs, followed by cosine annealing³⁰ with a minimum learning rate of 3×10^{-6} . The batch size was set to 32 , and training proceeded for a maximum of 45 epochs, with early stopping triggered when the validation loss exhibited no improvement for 20 consecutive epochs. Cross-entropy loss was adopted as the objective function.

The training data were augmented through horizontal and vertical flipping, as well as AugMix³¹, ColorJitter, Gaussian blur, sharpness adjustment, and random erasing. These augmentations simulate the natural variability in image acquisition conditions, including illumination, focus, and viewing angle, to improve generalization. During inference, only resizing and normalization were applied.

All experiments were conducted under stratified five-fold cross-validation, and the results are reported as the mean value over 10 independent runs. The proposed model was implemented using Python 3.9 and PyTorch 2.1, and all experiments were executed on an NVIDIA GeForce RTX 3090 GPU.

5. Experimental Results

5.1 Quantitative Evaluation

5.1.1 Performance comparisons with baseline models

To verify whether multi-scale feature learning provides a significant benefit for molten mark classification, we considered CNN-based models and transformer-based models as baselines for comparison. Table 2 summarizes the classification performance of the proposed model and baseline models on the standard testing dataset.

Table 2 Performance comparisons with baseline models.

Model	Accuracy	Precision	Recall	F1-Score	AUC	MCC
VGG16	0.9344	0.9866	0.8752	0.9276	0.9898	0.8731
EfficientNet	0.9248	0.8992	0.9499	0.9238	0.9802	0.8509
ViT	0.9023	0.9480	0.8430	0.8924	0.9748	0.8082
DeiT	0.9309	0.9684	0.8850	0.9248	0.9882	0.8641
Swin	0.9425	0.9613	0.9172	0.9387	0.9877	0.8855
Proposed model	0.9629	0.9633	0.9593	0.9613	0.9933	0.9257

One can see that the proposed model provides the highest scores across all metrics other than precision. Among the baseline models, VGG16 yields the highest precision, but exhibits the largest gap between precision and recall. This pattern indicates that a fixed receptive field struggles to capture global morphological features, causing many short-circuit bead instances to be missed. In contrast, the proposed model maintains balanced precision (0.9633) and recall (0.9593) values, resulting in the highest F1-score. Among the transformer-based models, ViT, which operates at a single resolution, exhibits the worst overall performance with an F1-score of 0.8924 and MCC of 0.8082 , whereas the Swin transformer, which adopts a hierarchical structure, ranks second with an F1-score of 0.9387 and MCC of 0.8855 .

This gap confirms the efficiency of multi-scale feature extraction for molten mark classification. Compared with the Swin transformer, the proposed model exhibits improvements of 2.26 percentage points in terms of F1-Score and 4.02 percentage points in terms of MCC.

5.1.2 Comparison of attention mechanisms

We compared the effects of different attention mechanisms in the fusion nodes, with all variants sharing the same backbone (Swin transformer) and neck (BiFPN). The compared models are the standard BiFPN, +CBAM²¹, and +Coordinate Attention³², with the latter two replacing scalar weighted summation with a channel-space attention module. Table 3 presents the comparison results.

Table 3 Performance comparisons of alternative attention mechanisms.

Model	Accuracy	Precision	Recall	F1-Score	AUC	MCC
BiFPN	0.9535	0.9478	0.9559	0.9518	0.9903	0.9069
+CBAM	0.9535	0.9586	0.9439	0.9512	0.9900	0.9069
+Coordinate Attention	0.9602	0.9651	0.9516	0.9583	0.9937	0.9204
Proposed model	0.9629	0.9633	0.9593	0.9613	0.9933	0.9257

One can see that the BiFPN yields an accuracy of 0.9535 and F1-score of 0.9518, surpassing the standalone Swin transformer (Table 2, accuracy of 0.9425) and confirming the fundamental benefit of bidirectional feature fusion. The proposed model provides the highest accuracy of 0.9629 and MCC of 0.9257, improving upon the MCC value of the BiFPN by 1.88 percentage points. Notably, the proposed model maintains a high F1-score (0.9613), demonstrating that cross-scale conditioning in DCSA selectively emphasizes discriminative features without discarding useful information.

These results confirm that the observed performance gains of the proposed model originate not from the mere inclusion of attention but from the cross-scale interaction of DCSA, which conditions attention on features from different resolution levels, rather than refining each fused output independently.

5.1.3 Confusion matrix and PR curve comparisons with baseline models

Figure 5 presents the confusion matrix of the proposed model and the PR curves of all baseline models.

The confusion matrix reveals that the proposed model correctly classified 2,463 out of 2,549 short-circuit beads (96.6%) and 2,260 out of 2,356 heat marks (95.9%), demonstrating balanced performance across both classes. The misclassifications consist of 86 false negatives and 96 false positives, indicating that the proposed model can minimize the error rate for both classes without bias.

According to the PR curves, the proposed model achieved the highest average precision (AP) of 0.9919, followed by the Swin transformer (0.9874) and DeiT (0.9872). Additionally, the curve of the proposed model remains closest to the upper-right corner, even in the high-recall region (above 0.9), without any abrupt decline. In contrast, the CNN-based models (VGG16 and EfficientNet) exhibit steep precision drops in the high-recall region, while ViT (AP 0.9739) consistently traces a lower curve across the entire range, confirming the discriminative limitations stemming from the absence of multi-scale features.

5.2 Quantitative Evaluation on the Degraded Testing Dataset

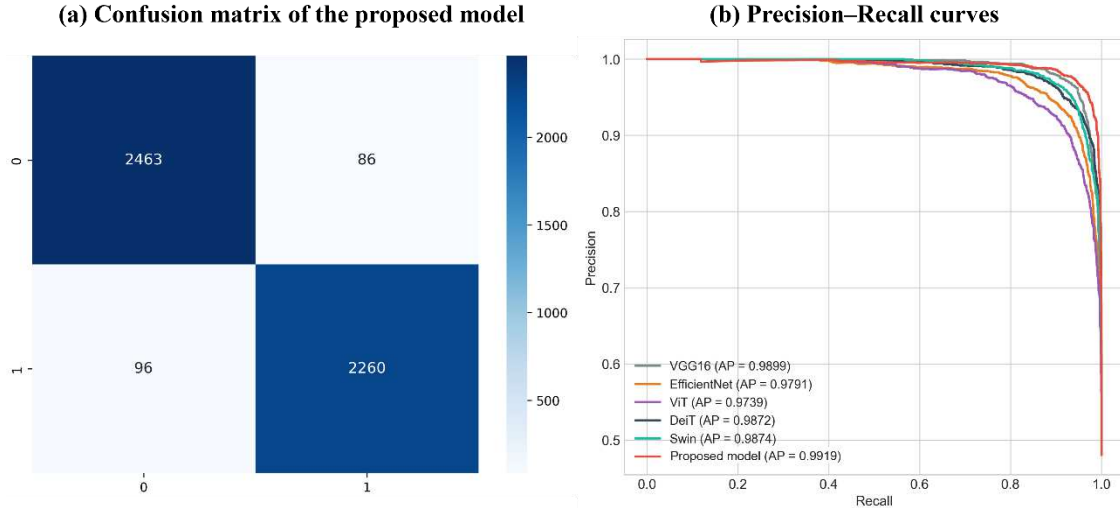


Figure 5 (a) Confusion matrix of the proposed model and (b) PR curve comparisons.

In real fire scenes, acquiring high-quality images is often difficult as a result of diverse external factors, including out-of-focus captures, uneven lighting, and surface contamination. This subsection quantifies the robustness of the proposed model and the five baseline models by evaluating their classification performance on such degraded fire images. The degraded testing dataset was constructed by randomly applying severe Gaussian blurring, additive noise, brightness and contrast distortions, and JPEG compression artifacts to the standard testing dataset.

These corruption types and their corresponding severity ranges were selected to approximate the imaging artifacts commonly encountered in fire scene evidence collection, including out-of-focus capture, uneven lighting, and lossy image transmission. To ensure reproducibility, the random seed for corruption assignment was fixed across all model evaluations.

5.2.1 Performance comparisons with baseline models

Table 4 Performance comparisons with baseline models on the degraded testing dataset.

Model	Accuracy	Precision	Recall	F1-Score	AUC	MCC
VGG16	0.7908	0.7065	0.9635	0.8152	0.9251	0.6238
EfficientNet	0.7535	0.6655	0.9788	0.7923	0.9113	0.5750
ViT	0.8887	0.9436	0.8171	0.8758	0.9670	0.7825
DeiT	0.8966	0.8720	0.9198	0.8953	0.9610	0.7945
Swin	0.9107	0.9057	0.9087	0.9072	0.9746	0.8212
Proposed model	0.9341	0.9298	0.9334	0.9316	0.9826	0.8681

Table 4 presents the classification performance of the proposed model and baseline models on the degraded testing dataset. The proposed model yields the best overall performance. Compared with the normal testing dataset, accuracy drops by only 2.88 percentage points, and the F1-Score by 2.97 percentage points. In contrast, VGG16 and EfficientNet exhibit decreases of 14.36 and 17.13 percentage points, respectively, in terms of accuracy. Their precision collapses from 0.9866 to 0.7065 and from 0.8992 to 0.6655, respectively, while their recall increases to 0.9635 and 0.9788, respectively. These two models classified most inputs in the degraded set as short-circuit beads because blurring and noise remove the fine surface textures relied on by these models, which is consistent with the texture bias of ImageNet-trained CNNs.

The transformer-based models demonstrate relatively high robustness. The Swin transformer yields an accuracy of 0.9107, ranking second overall. The MCC gap grows from 4.02 percentage points on the normal testing dataset to 4.69 percentage points on the degraded testing dataset. MCC drops sharply when errors pile up on one class, and the widening gap clearly indicates that the proposed model keeps its errors balanced as image quality deteriorates, whereas the baselines lean toward one class.

5.2.2 Performance comparisons of attention mechanisms

Table 5 Performance comparisons of alternative attention mechanisms on the degraded testing dataset.

Model	Accuracy	Precision	Recall	F1-Score	AUC	MCC
BiFPN	0.9257	0.8851	0.9715	0.9234	0.9832	0.8532
+CBAM	0.9253	0.8948	0.9571	0.9249	0.9818	0.8528
+Coordinate Attention	0.9288	0.9656	0.8832	0.9226	0.9809	0.8599
Proposed model	0.9341	0.9298	0.9334	0.9316	0.9826	0.8681

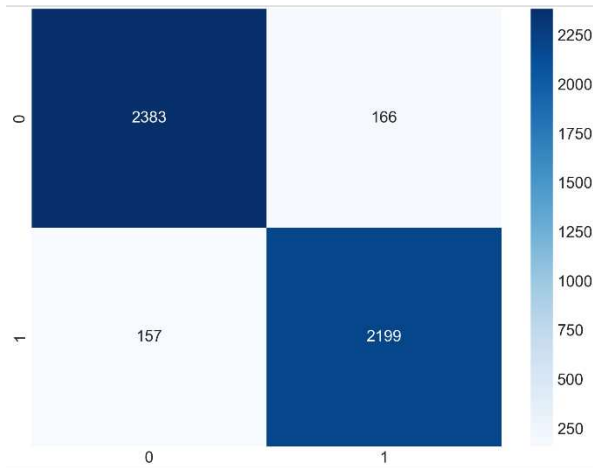
Table 5 presents the classification performance with different attention mechanisms in the fusion nodes, where all variants share the same backbone (Swin transformer) and neck (BiFPN).

The BiFPN exhibits a high recall of 0.9715 and low precision of 0.8851, suggesting the generation of many false positives. +CBAM achieves a precision of 0.8948; however, its MCC is approximately the same as that of the BiFPN. +Coordinate Attention yields a slightly improved MCC (0.8599), but its recall drops to 0.8832.

The proposed model raises precision by 4.47 percentage points compared with the BiFPN and achieves the highest accuracy (0.9341), F1-score (0.9316), and MCC (0.8681) among all variants. The MCC gap of 1.49 percentage points over the BiFPN demonstrates that the proposed model keeps its errors balanced under image corruption, whereas the other attention modules struggle with error bias. This pattern confirms that the cross-scale design of DCSA is effective at maintaining stable classification performance, even under degraded image quality conditions.

5.2.3 Confusion matrix and PR curve comparisons with baseline models

(a) Confusion matrix of the proposed model



(b) Precision–Recall curves

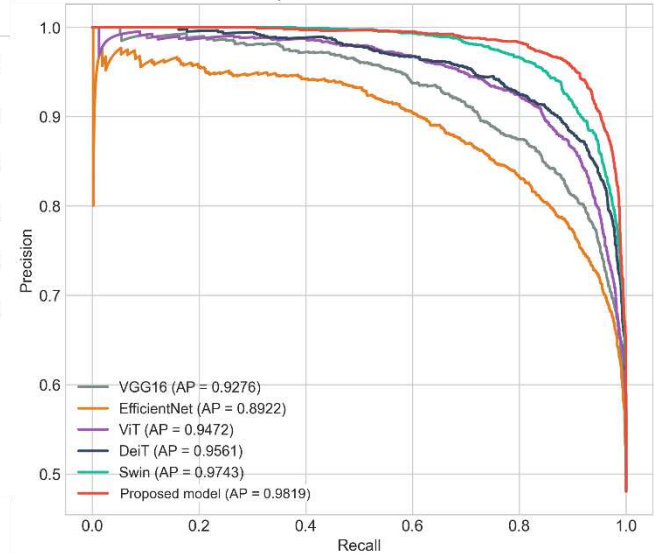


Figure 6 (a) Confusion matrix and (b) PR curve comparisons on the degraded testing dataset.

Figure 6 presents the confusion matrix of the proposed model and the PR curves of all baseline models on the degraded testing dataset. The confusion matrix reveals that the proposed model correctly classified 2,383 out of 2,549 short-circuit beads (93.5%) and 2,199 out of 2,356 heat marks (93.3%). Compared with the normal testing dataset, false negatives increase from 86 to 166, and false positives increase from 96 to 157. Although both error types increase with image quality degradation, the error distribution remains balanced, resulting in an overall accuracy of 93.41%. The observed moderate increase in false negatives suggests that severe blurring can occasionally mask the surface glossiness of short-circuit beads, causing them to mimic the smooth texture of heat marks.

The PR curves highlight the robustness disparity among the models. The proposed model records the highest AP of 0.9819, followed by the Swin transformer (0.9743) and DeiT (0.9561). The CNN-based models exhibit sharp AP declines to the range of 0.89–0.93, with precision falling below 0.7 in the high-recall region, substantially limiting

their practical applicability. In this experiment on a degraded dataset, the proposed model exhibited significantly less AP degradation compared with that observed in all other models, confirming that DCSA-based multi-scale fusion becomes increasingly advantageous as image quality deteriorates.

5.3 Ablation Study

Table 6 Ablation study of the proposed model on the normal testing dataset.

Model				Accuracy	F1-Score	AUC	MCC
Top-down		Bottom-up					
Channel	Spatial	Channel	Spatial				
✓	✓			0.9556	0.9527	0.9934	0.9115
		✓	✓	0.9573	0.9548	0.9908	0.9151
✓		✓		0.9553	0.9530	0.9899	0.9106
	✓		✓	0.9545	0.9517	0.9904	0.9094
✓	✓	✓	✓	0.9629	0.9613	0.9933	0.9257

In this analysis, we remove DCSA components one at a time and evaluate the resulting variants on both the normal and degraded testing datasets. All variants share the same Swin transformer backbone and BiFPN neck.

Table 6 presents the ablation results on the normal testing dataset. In the first row, DCSA is applied only to the top-down pathway, whereas in the second row, it is applied only to the bottom-up pathway. The third row applies only channel attention to both pathways, the fourth row applies only spatial attention to both pathways, and the final row represents the full proposed model combining all components.

Applying DCSA to a single pathway yields accuracy values of 0.9556 (top-down) and 0.9573 (bottom-up), and using a single attention type yields accuracy values of 0.9553 (channel) and 0.9545 (spatial). The proposed model, which combines all components, achieves an accuracy of 0.9629, F1-Score of 0.9613, and MCC of 0.9257. The MCC gap is the widest at 1.06 to 1.63 percentage points. Because MCC reflects all four entries of the confusion matrix, each partial variant causes errors to tilt toward one class, and only the full combination keeps errors balanced.

Table 7 Ablation study of the proposed model on the degraded testing dataset.

Model				Accuracy	F1-Score	AUC	MCC
Top-down		Bottom-up					
Channel	Spatial	Channel	Spatial				
✓	✓			0.9221	0.9213	0.9826	0.8457
		✓	✓	0.9286	0.9284	0.9835	0.8598
✓		✓		0.9137	0.9145	0.9802	0.8318
	✓		✓	0.9213	0.9215	0.9807	0.8458
✓	✓	✓	✓	0.9341	0.9316	0.9826	0.8681

Table 7 presents the results of the same ablation described above, but conducted on the degraded test dataset. Single-pathway variants yield accuracy values of 0.9221 (top-down) and 0.9286 (bottom-up), while single-attention variants reach values of 0.9137 (channel) and 0.9213 (spatial). The proposed model, which combines all components, achieves an accuracy of 0.9341, F1 score of 0.9316, and MCC of 0.8681, demonstrating superior performance compared with all other variants. The largest difference of 3.63 percentage points can be observed in MCC when comparing the full model with the variant applying only channel attention. This pattern demonstrates that the performance gap between individual elements widens in environments with degraded image quality, and demonstrates that the combination of all components is essential for ensuring stability.

5.4 Qualitative Evaluation

5.4.1 GradCAM visualization

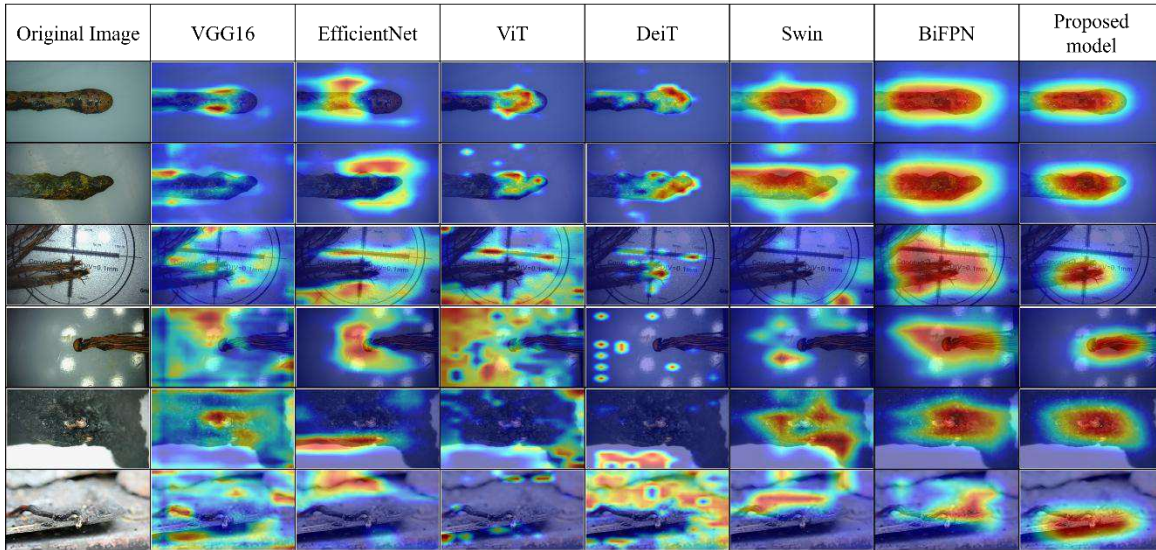


Figure 7 Qualitative comparison of Grad-CAM attention maps.

To examine which image regions are assigned the greatest attention by the proposed method during classification, we conducted GradCAM- based³³ attention region visualization for each model. Figure7 presents the GradCAM results for six molten mark images.

In rows 1 and 2, which represent a typical case with a clean background, one can see that most models, including the proposed model, focused their attention on the molten mark region. However, VGG16 and EfficientNet exhibit activations that are either biased toward a partial area of the molten mark or diffused beyond its boundary. In rows 3 and 4, the inter-model differences become more pronounced in the presence of background interference, such as gridlines and specular reflections. Most of the baseline models respond strongly to the gridlines, reflections, or regions unrelated to the molten mark. The BiFPN forms activations over the molten mark region, but with broad diffusion. In contrast, the proposed model effectively suppresses background interference and concentrates its activation exclusively on the molten mark. The advantages of the proposed model are clear for the small molten mark represented in row 5. The baseline models disperse their activations across surrounding soot or background regions, whereas both the BiFPN and proposed model precisely localize the small molten mark. This shared capability is attributable to the bidirectional multi-scale feature learning common to both models.

Row 6 represents an image captured at a real fire scene with a complex background and small molten mark, which is the most challenging type of scenario. Most of the baseline models either disperse their attention to the background or concentrate activations on regions unrelated to the molten mark, whereas the proposed model maintains focused activation in the molten mark region. These results qualitatively confirm that the proposed model possesses the discriminative ability required to assign attention to molten mark regions selectively, regardless of background interference and scale variation.

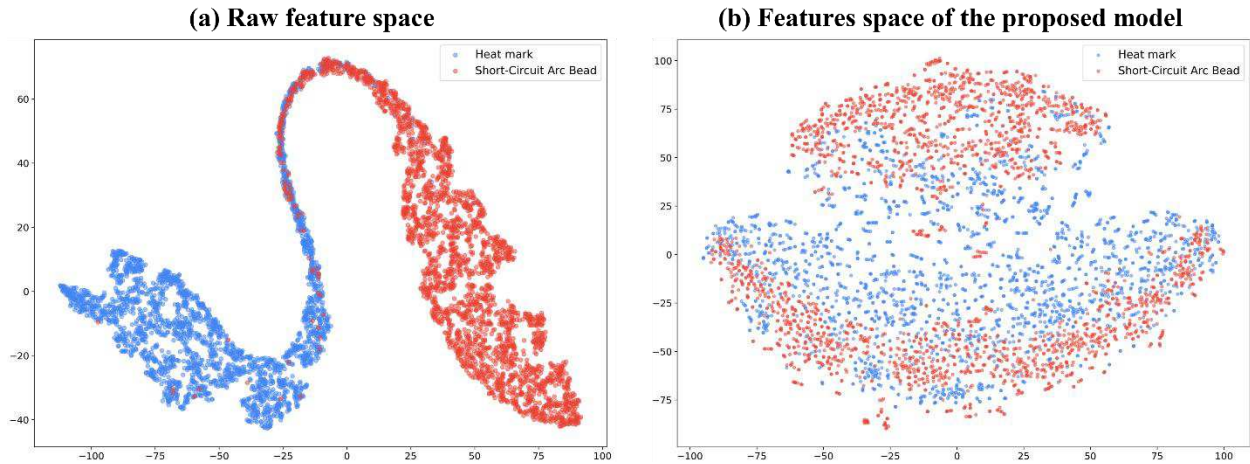


Figure 8 t-SNE visualization of molten mark feature representations.

To assess the discriminative quality of the learned feature representations intuitively, we conducted t-SNE³⁴ visualization on the testing dataset. Figure 8 presents projections of (a) the raw pixel features of the original images and (b) the features learned by the proposed model into a two-dimensional space. Blue points represent heat marks, while red points represent short-circuit beads.

In the raw feature space, the two classes are largely intermixed, confirming that the high visual similarity between short-circuit beads and heat marks renders pixel-level discrimination inherently challenging. In contrast, our model achieves clear class separation, with the heat marks and short-circuit beads forming distinct clusters. Notably, the proposed model produces multiple sub-clusters within each class, rather than mapping each class onto a single continuous manifold. This pattern suggests that the DCSA module learns fine-grained feature representations that differentiate diverse sub-patterns of molten marks, including variations in wire type, specimen size, and surface morphology. This enriched intra-class structure is consistent with the robustness observed on the degraded dataset in Section 5.2, as the capacity to discriminate diverse sub-patterns translates into enhanced adaptability to image quality variations.

6. Conclusion

In this paper, we proposed a molten mark classification method based on multi-scale directional cross-scale attention fusion. The proposed model extracts multi-scale features at four resolution levels using the Swin transformer as a hierarchical backbone and fuses the extracted features through a DCSA-enhanced BiFPN neck. In each fusion node, the proposed DCSA module assigns channel and spatial attention conditioned on features from different scales to emphasize classification-relevant features. To evaluate the effectiveness of the proposed method, we conducted extensive experiments with CNN-based and transformer-based comparative models on normal and degraded datasets. In our experiments on the normal dataset, the proposed model achieved an F1 score of 0.9613 and MCC of 0.9257, outperforming the baseline models in terms of overall classification performance. On the degraded dataset, the proposed model demonstrated the smallest performance degradation, verifying its robustness. Additionally, we conducted ablation studies and qualitative analyses. The former confirmed that both the top-down and bottom-up pathways contribute meaningfully and that channel and spatial attention are both required to achieve the full observed gain. GradCAM and t-SNE analyses verified that the proposed model effectively focuses on molten mark regions and learns discriminative feature representations.

References

1. Liu, K.-H., Shih, Y.-H., Chen, G.-J. & Chou, J.-M. Microstructural study on molten marks of fire-causing copper wires. *Materials* **8**, 3776–3790 (2015).
2. Deng, J. *et al.* Metallurgical analysis of the 'cause' arc beads pattern characteristics under different short-circuit currents. *J. Loss Prev. Process. Ind.* **68**, 104339 (2020).

3. Yu, Z., Chen, S., Deng, J., Xu, X. & Wang, W. Microstructural characteristics of arc beads with overcurrent fault in the fire scene. *Materials* **13**, 4521 (2020).
4. Babrauskas, V. Arc beads from fires: Can 'cause' beads be distinguished from 'victim' beads by physical or chemical testing? *J. Fire Prot. Eng.* **14**, 125–147 (2004).
5. Batool, S. & Bang, S. Classification of short circuit marks in electric fire case with transfer learning and fine-tuning the convolutional neural network models. *J. Electr. Eng. Technol.* **18**, 4329–4339 (2023).
6. Geirhos, R. *et al.* Imagenet-trained CNNs are biased towards textures; increasing shape bias improves accuracy and robustness. In *Proc. International Conference on Learning Representations (ICLR)* (2019).
7. National Fire Protection Association. *NFPA 921: Guide for Fire and Explosion Investigations* (NFPA, Quincy, MA, 2021).
8. Vyas, V. & Williams, R. Physical and metallographic analysis of arc beads and fire-melted copper. *Fire Mater.* **37**, 1–15 (2013).
9. Howitt, D. The metallographic analysis of arcing damage in electrical wires. *Fire Mater.* **33**, 385–395 (2009).
10. Luo, W., Li, Y., Urtasun, R. & Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 4898–4906 (2016).
11. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022 (2021).
12. Tan, M., Pang, R. & Le, Q. EfficientDet: Scalable and efficient object detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10781–10790 (2020).
13. Shi, W. *et al.* Wire melted mark metallographic image recognition and classification based on semantic segmentation. *Expert. Syst. Appl.* **238**, 122146 (2024).
14. Chen, J. *et al.* TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).
15. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 785–794 (2016).
16. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations (ICLR)* (2015).
17. Xu, W., Zhang, C., Wang, Q. & Dai, P. FEA-Swin: Foreground enhancement attention swin transformer network for accurate UAV-based dense object detection. *Sensors* **22**, 6993 (2022).
18. Xu, X. *et al.* An improved Swin Transformer-based model for remote sensing object detection and instance segmentation. *Remote. Sens.* **13**, 4779 (2021).
19. Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path aggregation network for instance segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8759–8768 (2018).
20. Lin, T. *et al.* Feature pyramid networks for object detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117–2125 (2017).
21. Woo, S., Park, J., Lee, J. & Kweon, I. CBAM: Convolutional block attention module. In *Proc. European Conference on Computer Vision (ECCV)*, 3–19 (2018).
22. Zhao, Y., Cui, Y. & Wang, Z. Improved YOLOv5 based on CBAM and BiFPN for rice pest and disease detection. In *Proc. International Conference on Computer, Internet of Things and Smart City (CIoTSC)*, 29–36 (2023).
23. Yu, H., Wang, J., Han, Y., Fan, B. & Zhang, C. Research on an intelligent identification method for wind turbine blade damage based on CBAM-BiFPN-YOLOV8. *Processes* **12**, 205 (2024).
24. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophys. Acta (BBA)-Protein Struct.* **405**, 442–451 (1975).
25. Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. International Conference on Machine Learning (ICML)*, 6105–6114 (2019).
26. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations (ICLR)* (2021).

27. Touvron, H. *et al.* Training data-efficient image transformers and distillation through attention. In *Proc. International Conference on Machine Learning (ICML)*, 10347–10357 (2021).
28. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255 (2009).
29. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations (ICLR)* (2019).
30. Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *Proc. International Conference on Learning Representations (ICLR)* (2017).
31. Hendrycks, D. *et al.* AugMix: A simple data processing method to improve robustness and uncertainty. In *Proc. International Conference on Learning Representations (ICLR)* (2020).
32. Hou, Q., Zhou, D. & Feng, J. Coordinate attention for efficient mobile network design. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13713–13722 (2021).
33. Selvaraju, R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 618–626 (2017).
34. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605 (2008).

Funding

This research was funded by the Fire Agency and the Korea Evaluation Institute of Industrial Technology (KEIT), grant number RS-2025-03372990.

Data availability

The datasets analysed during the current study are not publicly available because they were provided by the Korea Electrical Safety Corporation (KESCO) under restricted-use conditions, but are available from the corresponding author on reasonable request and with permission of KESCO.

Author contributions statement

T.K. conceived the study, developed the methodology, implemented the software, conducted the experiments, performed the validation and visualization, and wrote the original draft. J.S. provided critical feedback on the manuscript structure and reviewed and edited the manuscript. E.H. reviewed and edited the manuscript and supervised the study. All authors have read and agreed to the published version of the manuscript.

Additional information

Competing interests: The authors declare no competing interests.