

Received 4 September 2023, accepted 19 September 2023, date of publication 27 September 2023, date of current version 6 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3319670

RESEARCH ARTICLE

CECvT: Initial Diagnosis of Anomalies in Thermal Images

DONGHYUN KIM¹, HOSEONG HWANG¹, AND HOCHUL KIM²

¹Department of Medical Artificial Intelligent, Eulji University, Seongnam-si, Gyeonggi-do 13135, Republic of Korea

²Department of Radiological Science, Eulji University, Seongnam-si, Gyeonggi-do 13135, Republic of Korea

Corresponding author: Hochul Kim (tiger1005@eulji.ac.kr)


This work was supported in part by the Ministry of Trade, Industry and Energy (MOTIE), in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP), and in part by the National Information Society Agency (NIA) under Grant 2022202090010A.

ABSTRACT Given the global competitive landscape, it is imperative that businesses maintain and manage their facilities continuously to enhance efficiency and productivity for sustaining competitiveness. Hence, a new hybrid model called contrast enhancement convolutional vision transformer (CECvT) was developed in this study that enables fault diagnosis without physical contact with factory equipment to ensure accurate initial fault detection without the risk of machine damage or interference. This model leverages thermal imaging as an apt source for early anomaly detection in equipment. A new contrast enhancement module employing contrast enhancement techniques was integrated to address the edge information loss when utilizing thermal images. Moreover, the network performance was enhanced by fusing the advantages of convolutional neural network (CNN) and Transformer models. Notably, the model design allows deriving detailed feature information necessary for the initial diagnostics by harnessing multiscale information to extract and concatenate features. The proposed method's performance was evaluated using the thermal imaging dataset provided by AI Hub. When juxtaposed with CNN, Transformer, and hybrid CNN–Transformer models, the proposed model demonstrated a superior accuracy of 96.17%. Furthermore, it achieved the most accurate diagnosis at the inception of abnormalities than the other networks. The proposed model thus has potential and is preferable for various thermal-imaging-based fault diagnosis applications owing to its excellent performance and precision during initial diagnosis.

INDEX TERMS Deep learning, infrared imaging, anomaly detection, factory equipment, contrast enhancement, multi-scale feature fusion.

I. INTRODUCTION

In recent times, factory equipment has garnered increasing attention in various sectors, including mobile telecommunications, automotive, maritime, and railway industries as well as the academic community, underscoring their pivotal roles in manufacturing. Ensuring the sustained operations of these critical facilities necessitate a keen emphasis on the diagnostic assessment of equipment abnormalities. To achieve this imperative task, real-time sensor data pertaining to parameters, such as temperature, pressure, vibration, and flow rate,

The associate editor coordinating the review of this manuscript and approving it for publication was Ravibabu Mulaveesala .

are gathered meticulously [1]. These data are used to continuously monitor equipment conditions and identify potential malfunctions or anomalies. Furthermore, using microphones or alternative auditory sensors to procure acoustic data aids in detecting irregularities in the sound or noise levels, facilitating diagnosis of factory equipment problems [2].

Among the myriad sensor signals, vibration signals stand out as they can depict the dynamic state of factory equipment directly and clearly. Consequently, vibration-based diagnostic methods are being researched extensively [3]. Vibration signals that reflect the vibrational phenomena originating within the equipment assist in detecting internal defects, such as interactions between components, wear and tear of parts,

and gaps. This makes them highly efficient for sensing the internal conditions of equipment. However, practical industrial applications of conventional vibration-based diagnostics for factory equipment present challenges; these include the considerable time and efforts required to install vibration sensors on each equipment. Moreover, their designs must be robust against various noise and interference sources, such as ambient vibrations, electrical noise, or mechanical noise from adjacent equipment [4].

The second most frequently employed method to diagnose abnormalities in factory equipment is based on acoustics. Acoustic data obtained from sounds produced by the equipment can reflect a range of defects and anomalies, facilitating comprehensive assessment of the equipment's overall condition. Furthermore, acoustic data are invaluable for detecting various defects [5]. However, leveraging acoustic data for inspection poses challenges, especially within factories. Typically, factories entail multiple equipment and processes generating significant background noise, complicating the isolation and identification of specific acoustic signals associated with equipment anomalies. Additionally, given the nature of acoustic signals that attenuate upon propagation through air or other media, detecting and analyzing acoustic data from equipment that are either distant or obstructed by other structures becomes problematic. Such challenges compromise the accuracy and reliability of monitoring systems, and several such problems persist in the current applications [6].

Utilizing thermal imaging to diagnose abnormalities in factory equipment offers significant advantages over vibration and acoustic data. Thermal cameras can capture temperature data without physical contact with the equipment, thereby reducing the risk of mechanical damage or interference during inspection. The visual representation of temperature distribution allows easy identification of potential equipment problems, such as hotspots, thermal leakages, or anomalous heat patterns, thus ensuring a high level of inspection efficacy. Moreover, the ability to detect abnormal temperatures at the early stages facilitates preventative maintenance, potentially reducing costly operational downtimes [7]. However, there are inherent challenges in using thermal imaging for equipment diagnostics. Interpreting thermal imaging data demands specialized knowledge and experience. Owing to the data complexities and their subjective interpretability, rigorous training and expertise are essential for accurate anomaly detection. Establishing precise criteria to identify abnormalities in the imaging data is imperative, which means distinctly defining the temperature patterns and variations between normal and anomalous states. Given the equipment characteristics and operating conditions, setting appropriate benchmarks can be challenging. Lastly, the spatial resolution of a thermal camera is often limited, posing difficulties in accurately identifying minute defects or detailed anomalies. Distinguishing the precise appearances of parts can be challenging, particularly when diagnosing

small or intricate equipment components [8]. Nevertheless, the benefits of thermal-imaging-based equipment condition monitoring are substantial. Recent advancements in fusing artificial intelligence with thermal imaging diagnostics have exhibited promising results, with enhanced anomaly detection performances.

In recent times, machine learning approaches, such as the hidden Markov model (HMM), support vector machine (SVM), k-nearest neighbor (KNN), and artificial neural network (ANN), have been applied for anomaly detection using thermal imaging. Among these, for anomaly detection based on deep learning with thermal imagery, the convolutional neural network (CNN) that has been applied extensively across various computer vision domains is actively used [9]. However, there are continuous concerns regarding CNN's limitations in utilizing global information. To address this, recent advancements in computer vision have integrated Transformer technology actively. Moreover, compared to CNNs, Transformers demonstrate superior performance on noisy or augmented images; this is due to the self-attention mechanism of the Transformer, which allows using image information from the highest to lowest layers to provide enhanced performance over CNN. However, the Transformer technology that is rapidly gaining traction in the computer vision domain also has challenges, such as the vast amount of data required for training. CNNs can generalize better with smaller datasets, resulting in greater accuracy, whereas Transformers necessitate more extensive training data since the images are divided into smaller patches, introducing more diverse inter image relationships. Recognizing these challenges of the CNN and Transformer methods, there is emerging emphasis on the need for research into deep-learning networks that integrate both these models. Current research highlights the potential benefits of such integration; notably, recent studies have indicated that the ensemble with CNN and Transformer yields up to 10% higher accuracy on the ImageNet-C benchmark than various other networks [10], [11].

To address the challenges of CNN and Transformer technologies, research is underway to develop deep-learning networks that integrate the strengths of both techniques for applications in various fields. In the present study, the convolutional vision transformer (CvT) is used as the base model for further research [12].

In contemporary research utilizing thermal imaging to diagnose equipment malfunctions in factories, the primary focus has historically been on accurately diagnosing equipment failures once they have occurred, rather than early-stage detection by identifying minor anomalous temperature elevations. Consequently, extant methods are limited for diagnosing subtle temperature anomalies during the initial phases. Additionally, the spatial resolution of a thermal camera is inherently limited, presenting challenges in pinpointing small defects or diagnosing early-stage abnormalities. Research endeavors to overcome these constraints have been sparse.

In this study, to address the problem of edge information losses in thermal images—a critical feature in early equipment abnormality diagnosis—a new contrast enhancement module is introduced by leveraging contrast enhancement techniques. By harnessing the strengths of both CNN and Transformer, the integrated CvT network structure is presented. Further, by employing multiscale information, such as fully convolutional cross-scale flows (CS-Flow), multiscale image features were extracted and concatenated to obtain detailed feature information essential for preliminary diagnoses [13]. Therefore, the primary aim of this research is the development of the contrast enhancement convolutional vision transformer (CECvT) network that can achieve early abnormality diagnosis efficiently.

The main contributions of the model proposed in this study over those in literature are summarized below.

- 1) By introducing a new contrast enhancement module, the edge information loss problem of thermal images is addressed successfully.
- 2) Using multiscale information from the CS-Flow network, the proposed system detects and diagnoses anomalies even in small-sized regions via extracting and utilizing detailed features, leading to significant findings.
- 3) Through restructuring and optimizing a network that integrates only the advantages of both the CNN and Transformer methods, a new network structure that demonstrates superior performance to the conventional CNN or Transformer network is developed.

The remainder of this manuscript is organized as follows. Section II discusses the literature and background on thermal-imaging-based anomaly detection with deep-learning techniques. Section III outlines the details of the thermal image dataset composed of abnormal data from various cases. Section IV presents detailed descriptions of the contrast enhancement module and proposed network structure. Section V explains the proposed experimental approach and experimental evaluations of different techniques compared to the proposed technique. Section VI presents the conclusions of this work.

II. RELATED WORKS

A. THERMAL-IMAGING-BASED ANOMALY DETECTION

To date, studies on equipment anomaly detection using thermal imaging have predominantly relied on simple image processing. Over the past few decades, there has been a global surge in the installation of photovoltaic (PV) power plants. The output efficiencies of these stations degrade over time owing to several factors. Advancements in drone technology have allowed researchers to employ drones equipped with thermal cameras to monitor PV power plants. These drones are often fitted with both red-green-blue (RGB) and thermal cameras. The proposed system identifies defects from among hundreds or even thousands of PV modules in a power

plant, extracting details through contour detection algorithms like the canny edge detector [14]. Furthermore, studies have been conducted on stainless-steel plates with circular defects captured using thermal cameras, where filtering was used to possibly improve the signal-to-noise ratio, followed by automated defect detection based on thresholding of binary images [15]. In addition, studies have been conducted on BLDC motors whose conditions were captured using thermal cameras, and the defects were identified through a feature extraction method called the common part of arithmetic mean of thermographic images (CPoAMoTI) [16].

However, recent advances include active research deploying ANN techniques to diagnose anomalies in thermal images. Using ANNs, the unforeseen anomalies, such as gear wear in gearboxes, can be detected. Thermal analyses of thermal images were used as a novel noninvasive approach to diagnose and categorize uniform wear levels of gears via automated defect diagnosis using ANNs [17]. Additionally, detecting thermal bridges in building envelopes is a critical aspect that needs prioritized resolution to enhance the thermal performances of buildings. Recently, thermal imaging measurements have been adopted to detect thermal bridges. There is a proposal for an image-processing- and a machine-learning-based linear thermal bridge detection method using images captured by thermal cameras; this method involves clustering the thermal anomaly regions, feature extraction, and thermal bridge detection using ANNs [18].

To accurately identify leak defects in the pipelines of mine air compressors, wavelet noise reduction and Otsu-GrabCut image segmentation were employed, followed by defect diagnosis using SVM [19]. In another study, an automated diagnostic method was applied to inspect PV power plants and identify anomalies within the panels; this involved capturing thermal images using an unmanned aerial vehicle (UAV) and diagnosing defects with SVM [20]. To address the problem of high failure rates of electric heating devices (EORs) during rail transportation, thermal imaging was employed to detect EOR errors and malfunctions using SVM [21]. Thermal images of concrete structures were preprocessed to emphasize and detect cracks on the exterior walls of buildings; SVM was then used to compute and classify the visual characteristics of each region to achieve accurate crack detection [22].

When employing the generative adversarial network (GAN) to prevent power system collapse, it is essential to detect various overheating defects in the operational states of the power transformers that play crucial roles in the system. Here, GAN was utilized to recognize and diagnose the overheating locations during transformer operation [23].

A deep-learning approach was previously used for real-time detection of equipment components using CNN to predict and diagnose the component coordinates, orientations, and grade types [24]. Moreover, the efficacy of the CNN was validated for diagnosing anomalies in rotating machinery based on infrared thermographic imaging [25].

Infrared imagery has been employed in conveyor systems for binary classification of thermographic images to monitor the status of belt conveyor idlers using CNN [26]. Many solar power plants face challenges due to numerous defects that cause non-negligible power losses; to address these, drone-based thermographic imaging was utilized along with CNN-based anomaly detection [27].

The Transformer network has been used to detect small infrared targets. When the CNN method was used for this purpose, there was a problem with modeling the long-range dependencies of the images owing to the locality of the convolutional kernel. Thus, the Transformer was applied to detect small infrared targets in 640×512 field-of-view (FOV) images [28].

B. THERMAL-IMAGING-BASED ANOMALY DETECTION WITH COMBINED CNN AND TRANSFORMER

In recent deep-learning research, it has been consistently reported that the fusion of CNN and Transformer exhibits the best performance. Following this trend, studies based on thermal imagery have been conducted, including those focused on detecting small infrared ships from space. The aim here was to differentiate small ships from images captured by Earth-orbit satellites. Owing to the vast image-coverage area, the potential targets in such images appear much smaller and fainter than those observed using aerial and ground-based imaging devices. To extract multistage features from such images, Transformer and CNN were fused. The local feature maps were first extracted from several convolutional layers; using the Transformer module to derive long-range dependencies, a high-performance network was designed [29].

III. MATERIALS

The AI Hub thermal imaging dataset, which is a public dataset, was used to accurately compare various methods and studies from literature. Detailed information regarding this dataset that has been used for the first time is provided below [30].

A. DATASET

For experiments on equipment malfunction diagnosis using thermal imaging, the AI Hub thermal imaging dataset was employed. Each image in the dataset has dimensions of $256 \times 256 \times 3$. For the training dataset, a total of 500,512 normal and 171,708 malfunctioning thermal images were used. For the validation dataset, a total of 55,612 normal and 19,079 malfunctioning thermal images were employed. The evaluation dataset consisted of 69,521 normal and 23,854 malfunctioning thermal images to assess the proposed network performance. In total, about 625,645 normal and 214,641 malfunctioning thermal images were utilized. The overall quantity of thermal imaging data used by class is listed in Table 1.

TABLE 1. Dataset information.

	Train		Validation		Test	
	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Storage Tank	57193	18590	6355	2066	7944	2583
Transfer Valve	54850	18545	6094	2061	7618	2576
Transfer Piping	54643	18586	6071	2065	7590	2582
Switch Board	55267	19200	6141	2133	7677	2667
Outdoor Unit	56325	18591	6258	2066	7823	2583
Inside the Factory	55498	20578	6166	2286	7709	2859
Outside the Factory	56063	19016	6229	2113	7787	2642
Car	54579	20344	6064	2260	7581	2826
Ship	56095	18259	6233	2029	7792	2536
Total	500512	171708	55612	19079	69521	23854

IV. METHODS

The CvT model is one of the popular architectures in literature for image classification [12]. One of the primary reasons why the CvT is favored is that it blends the advantages of CNNs, which generalize well even with smaller datasets to achieve superior accuracy, with the strengths of Transformers. The latter, with its self-attention mechanism, is adept at utilizing whole-image information from the highest to lowest layers. Hence, the fused model performance surpasses those of the conventional CNN and Transformer models. Accordingly, this study adopts the CvT model as the base with further improvement to the network structure. The foundational CvT model is first discussed; then, the enhanced CECvT model obtained by incorporating contrast enhancement and coupling modules for utilizing multiscale information is detailed. Optimization of the CECvT network structure is then elaborated. Thus, the distinctions between the proposed CECvT and base CvT models are elucidated. Finally, the performance metrics used to gauge the proposed system are presented.

A. BASE MODEL: CvT

The CvT integrates the hierarchical architecture, a hallmark of the CNN, to enhance the locality lacking in the vision transformer (ViT). Through the hierarchical architecture, the CvT learns low-level features like the broad contours of objects in the low layer, while the high-level features such as the detailed characteristics of objects are learned in the high layer. To incorporate the hierarchical nature of CNNs into ViT, convolutional token embedding and convolutional projection for attention are applied to the network structure.

Convolutional token embedding applies overlapping convolution with stride operations to the 2D token map,

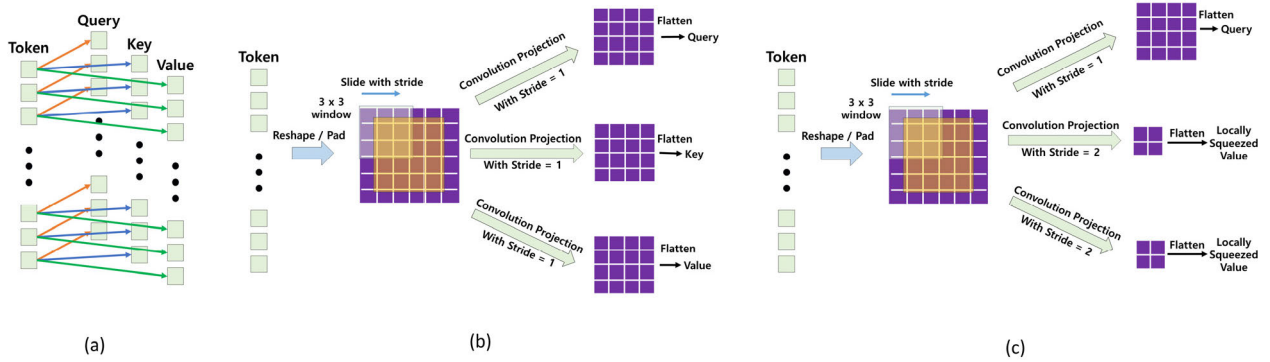


FIGURE 1. Convolutional projection.

harnessing both local information capture and spatial down-sampling concurrently. Convolutional token embedding is incorporated to augment the convolutional characteristics of the CvT and aims to model the local spatial context spanning from low- to high-level features.

$$H_{i+1} = \frac{H_i + 2p - s}{s - o} + 1, \quad W_{i+1} = \frac{W_i + 2p - s}{s - o} + 1 \quad (1)$$

Equation (1) describes the convolution operation applied when given a 2D input image or the output token map from the previous step to derive a new token map with the corresponding height and width. Here, H represents height, W denotes width, C signifies channel, s is the kernel size, $s-o$ stands for stride, and p indicates padding; i is the instance when the 2D input image or output token map is fed in, while $i+1$ is the moment for which the new output token map is generated. By leveraging convolutional token embedding, the token feature dimensions and number of tokens at each stage can be adjusted by modifying the convolutional operation parameters. Here, while the token feature dimensions increase at each stage, the length of the token sequence progressively decreases; this enables the tokens to represent increasingly complex visual patterns over larger spatial extents, akin to the layers of a CNN.

The convolutional projection for attention replaces the linear projections used in ViT with the depthwise convolution operation commonly employed in CNNs, thereby harnessing the structural characteristics of a CNN. The goal of the convolutional projection layer is to achieve additional modeling of the local spatial context and permit undersampling of the key and value matrices for efficiency. The original positionwise linear projections for multihead self-attention (MHSA) are replaced with depthwise convolutions to form the convolutional projections. Figure. 1(a) depicts the positionwise linear projections used in ViT, while Figure. 1(b) represents the proposed $s \times s$ convolutional projections, which when articulated in mathematical terms is given by (2):

$$x_i^{query, key, value} = F(CP(Reshape(x_i), s)) \quad (2)$$

Here, $x_i^{query, key, value}$ represents the token inputs for the corresponding query, key, and value matrices. The tokens are first restructured into a 2D token map; then, a convolutional projection is implemented using a convolution with a kernel size of $s \times s$. The *Reshape* function transforms the tokens into a 2D map, as seen in Figure. 1(b). The *CP* function denotes the depthwise separable convolution implemented in the order of depthwise Conv2d \rightarrow BatchNorm2d \rightarrow pointwise Conv2d. The function F signifies flattening the tokens to 1D for subsequent processing. Figure. 1(c) demonstrates the use of convolution with a stride larger than 1 to reduce the computational cost of the MHSA operation by decreasing the numbers of tokens for key and value by a factor of four. This results in a slight performance degradation but achieves a fourfold reduction in the computational cost. Since the adjacent pixels/patches in an image tend to have redundant shapes/semantics, the performance decline is minimal. Furthermore, local context modeling of the convolutional projections compensates for the information losses due to resolution reduction, resulting in only a marginal performance drop.

B. PROPOSED MODEL: CECvT

The contrast enhancement module is used to address the problem where all the image pixels in a camera are not properly exposed due to dynamic range limitations [31]. While increasing exposure can reveal some underexposed areas, it also risks overexposing other areas that were previously well-lit. The module not only resolves this problem but also strengthens edge information, ensuring that the target edges are distinctly visible. Initially, the contrast enhancement module employs the illumination estimation technique for designing the weight matrix for image fusion. Subsequently, the camera response model is used to composite multiple exposure images. Then, the optimal exposure ratio is determined to ensure that the underexposed regions in the input image are well-exposed in the composite image. Lastly, based on the weight matrix, the input and composite images are fused to produce the final result.

Weight matrix estimation is performed for the exposure fusion framework, where the estimation algorithm enhances the low contrast of underexposed areas while preserving the contrast of well-exposed regions. It assigns higher weights to the well-exposed pixels and lower weights to underexposed ones. The weight matrix is positively correlated with image illumination; brighter areas that are more likely to be well-exposed are assigned higher weights to maintain their contrast. The camera response model used in the framework incorporates beta-gamma correction. The exposure ratio determination for the framework seeks the optimal exposure ratio to ensure that the composite image is well-exposed in areas where the original image is underexposed. First, an image with overall underexposure is obtained, excluding the well-exposed pixels. The underexposed image is then rectified using the exposure ratio determination algorithm.

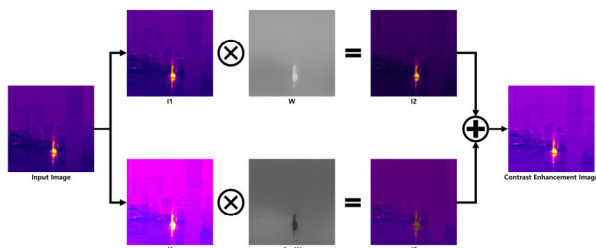


FIGURE 2. Contrast enhancement module process

Figure 2 illustrates the previously described contrast enhancement module. From the input image, both $I1$ and $J1$ images are generated, where $I1$ is identical to the input image. However, the $J1$ image identifies the underexposed areas of the input image and enhances the low contrast of these regions. W represents the weight matrix, where the well-exposed pixels with high weights can be observed. The $1-W$ image is the inverse of W , where the underexposed regions have higher weights. The $I2$ image is produced by pixelwise multiplication of values of the $I1$ and W images. Similarly, the $J2$ image is derived by pixelwise multiplication of the values of the $J1$ and $1-W$ images. The final contrast enhanced image is obtained by the addition of $I2$ and $J2$.

Figure 3 visually presents the contrast enhanced image resulting from enhancing the input image. The edges of areas with abnormally high temperatures are defined clearly. Even in small regions with high temperature anomalies, it is evident that the contrast is improved, making them visually distinguishable.

In the architecture of the CECvT network illustrated in Figure 4, the approach deviates from the previously developed CvT network structure. The input image or 2D token map is resized to three different dimensions using the Resize function and subsequently enhanced using the contrast enhancement module. Then, the contrast-enhanced output images with three different sizes employ a method

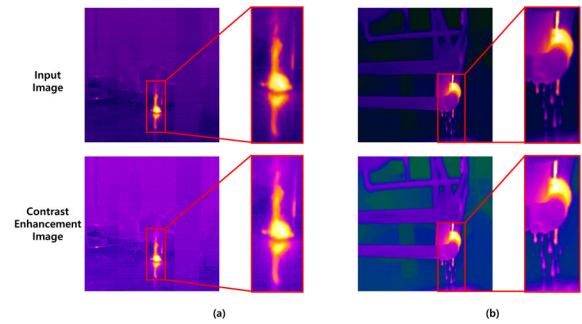


FIGURE 3. Contrast enhancement examples.

inspired by CS-Flow, which leverages multiscale information [13]. This method known as cross-scale convolution in the broader context is referred to as the coupling module in the proposed structure. As depicted in Figure 5, the cross-scale convolution inside the coupling module consists of two convolutional stages. The first stage is a typical 2D convolution, while the second is the cross convolution. In the cross convolution stage, the differently sized feature maps are adjusted to the same size before summation and is achieved by altering the stride or through upsampling. Following the two convolutional stages, the outputs are summed elementwise, and the three distinct-sized feature maps are concatenated. One key distinction here from the conventional CS-Flow method is that previously, only the second input feature map leveraged information from the first and third inputs. In contrast, the proposed coupling module design allows all three maps to be added elementwise, enabling richer utilization of information.

In Stage 1, features are extracted at three distinct scales by resizing the input image of dimensions 256×256 to 256×256 , 128×128 , and 64×64 . In Stage 2, by resizing these images to 56×56 , 28×28 , and 14×14 , the architecture extracts features at three different scales. Similarly, in Stage 3, features are yet again extracted from three distinct scales with image dimensions of 28×28 , 14×14 , and 7×7 . This design ensures that the network architecture can capture a diverse range of feature information.

C. IMPLEMENTATION DETAIL

The present study was conducted on a workstation equipped with an i9 processor (i9-13900k), a DDR5 PC5-44800 64GB RAM, a 64MB cache, CUDA version 11.0 or higher, CuDNN 8.8.1, a 24GB GPU (NVIDIA 3090), and a 64-bit operating system. The experiments were conducted using Python 3.7 and Pytorch 1.13.0.

D. EVALUATION MEASURES

In this study, six parameters were utilized as the evaluation measures. “NG Detection” denotes the quantity accurately identified as defects, while “OK Detection” indicates the quantity accurately determined as nondefective. “Overkill” refers to instances where a nondefective status is mistakenly

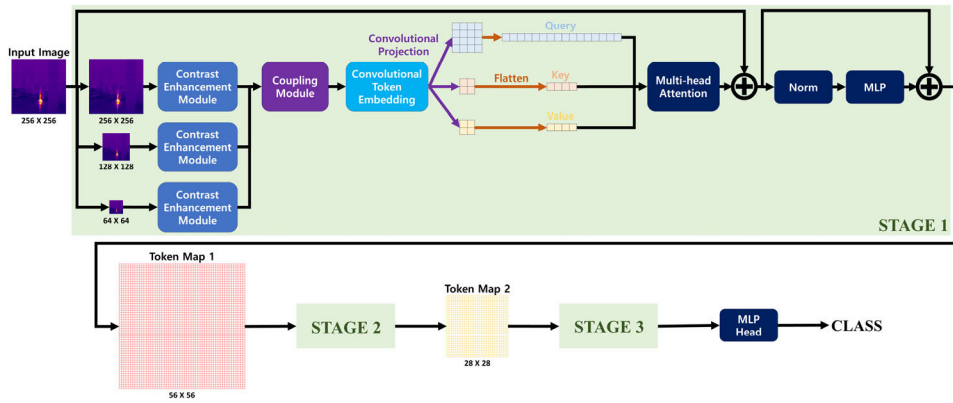


FIGURE 4. CECvT network structure.

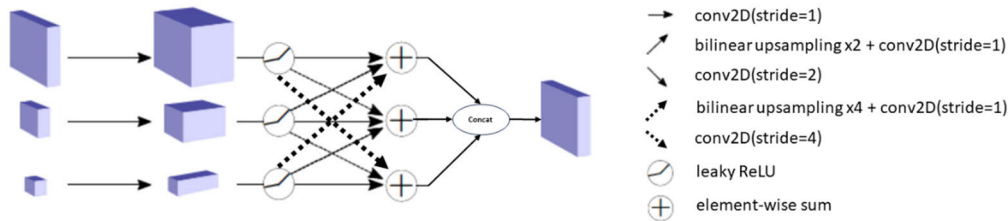


FIGURE 5. Coupling module.

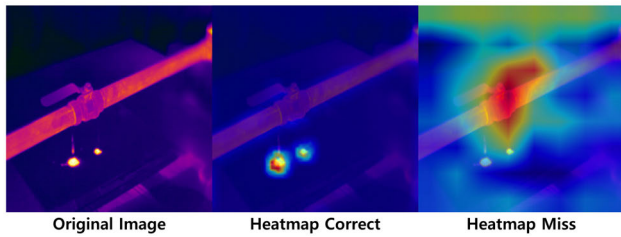


FIGURE 6. Heatmap correct and heatmap miss example.

classified as defective, whereas “Escape” denotes instances where a defective status is incorrectly classified as nondefective. “Heatmap Miss” is a scenario where a defect is identified as such, but upon inspection of the heatmap, it is not precisely activated at the defect location. A heatmap provides information about the parts of the input image that significantly influence the final determination. If the influential areas on the heatmap do not overlap with the defect locations, it cannot be said that a correct judgment was made. As shown in Figure. 6, one can distinguish between “Heatmap Correct,” which accurately pinpoints the defect location, and “Heatmap Miss,” which misidentifies the defect location. “Accuracy” is defined as per (3) and serves as an evaluation measure for the overall performance of the deep-learning network. Among the six parameters, the most critical metric is “Escape.” Incorrectly classifying a defect as nondefective can have catastrophic consequences for the equipment,

making it the most essential evaluation measure.

$$Accuracy = \frac{OK\ Detection + NG\ Detection + Heatmap\ Miss}{Total\ Input\ Image} \tag{3}$$

V. RESULTS AND DISCUSSION

To compare the superior performance of the proposed method, CECvT, the infrared camera image dataset from AI Hub was used. For performance evaluations, a total of 69,521 normal and 23,854 defective infrared images were used as the test dataset. Performance comparison experiments were conducted in two stages. The first experiment compared the detection capability based on the infrared images for anomaly detection. The second experiment was aimed at early anomaly diagnosis by dividing the anomalies from infrared images into three stages, namely initial, middle, and last, and comparing the detection capability for each stage. All experiments involved performance comparisons between CNN-based and Transformer-based models as well as hybrid networks combining CNN and Transformer features, against the proposed network CECvT.

A. DETECTION CAPABILITY COMPARISON WITH STATE-OF-THE-ART STUDIES

The first experiment focused on performance comparisons for the accuracies of normal and anomaly judgments as well

TABLE 2. Detection capability comparison.

Type	Network	Result					
		NG Detection	OK Detection	Overkill	Escape	Heatmap Miss	Accuracy
CNN	ResNet	23850	69520	1	4	8651	90.73%
	SEResNet	23850	69518	3	4	7555	91.90%
	EfficientNet	23778	69459	62	76	14349	84.49%
Transformer	ViT	23818	69519	2	36	16543	82.24%
	SwinT	23826	69509	12	28	15960	82.86%
CNN + Transformer	CvT	23852	69519	2	2	6590	92.94%
	CMT	23848	69519	2	6	8166	91.25%
	CoAtNet	23851	69514	7	3	7250	92.22%
	CECvT	23852	69518	3	2	3572	96.17%

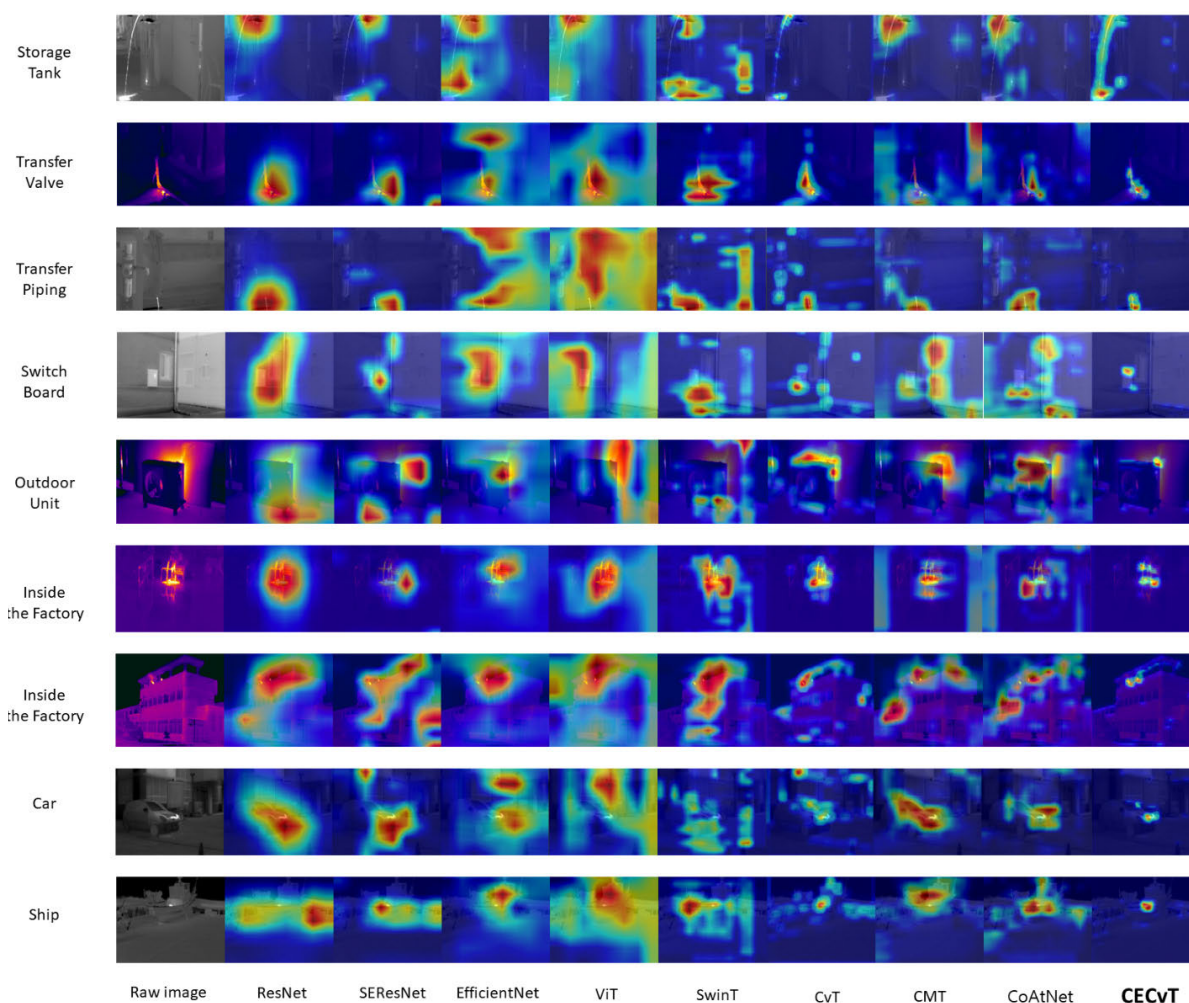


FIGURE 7. Heatmaps image of detection capability comparison with state-of-the-art studie.

as the occurrence of Heatmap Miss on an evaluation dataset consisting of nine classes. To compare the performances,

three CNN-based models (ResNet [32], SEResNet [33], and EfficientNet [34]), two Transformer-based networks

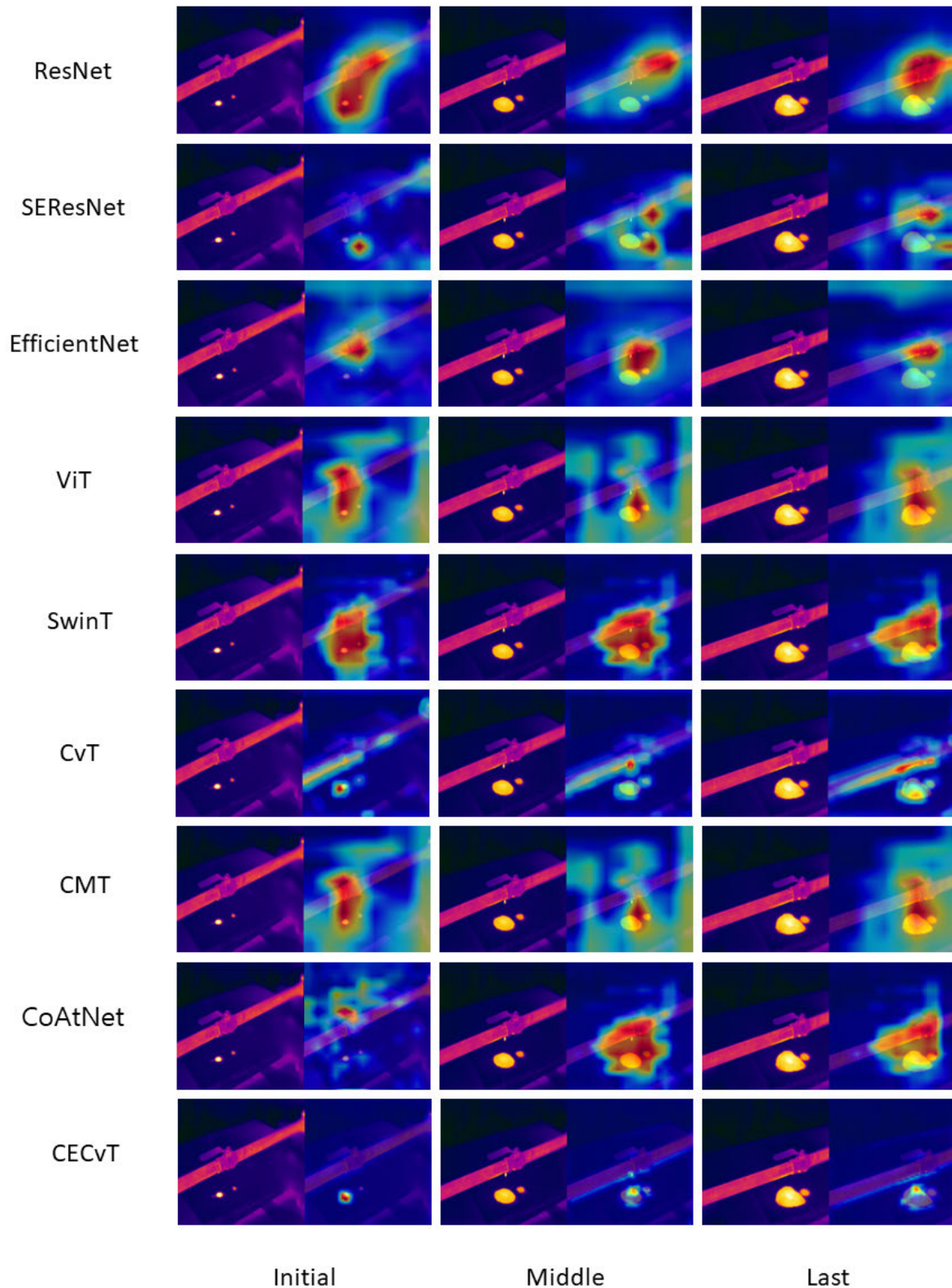


FIGURE 8. Heatmap images of detection capability when an initial abnormal occurs comparison with state-of-the-art studie.

(ViT [35] and SwinT [36]), and three hybrid networks combining the features of CNN and Transformer models (CvT [12], CMT [37], and CoAtNet [38]) were used. In total,

performance comparisons were conducted against eight networks. The results shown in Table 2 indicate that the proposed network CECvT exhibits the best performance.

From the numerical results in Table 2, among the CNN-based models, ResNet had 1 Overkill, 4 Escapes, 8651 Heatmap Miss, and 90.73% Accuracy. For SEResNet, the results showed 3 Overkill, 4 Escape, 7555 Heatmap Miss, and 91.90% Accuracy; EfficientNet showed numerical results of 62 Overkill, 76 Escape, 14349 Heatmap Miss, and 84.49% Accuracy. Among the Transformer-based networks, the numerical results were as follows: 2 Overkill, 36 Escape, 16543 Heatmap Miss, and Accuracy 82.24% for ViT; 12 Overkill, 28 Escape, 15960 Heatmap Miss, and 82.86% Accuracy for SwinT. For the combined CNN and Transformer models, the results were as follows: 2 Overkill, 2 Escape, 6590 Heatmap Miss, and 92.94% Accuracy for CvT; 2 Overkill, 6 Escape, 8166 Heatmap Miss, and 91.25% Accuracy for CMT; 7 Overkill, 3 Escape, 7250 Heatmap Miss, and 92.22% Accuracy for CoAtNet. The proposed network CECvT shows numerical results of 3 Overkill, 2 Escape, 3572 Heatmap Miss, and 96.17% Accuracy.

Based on the sum of Overkill and Escape, CvT displayed the best results, with a total of 4; however, the proposed network CECvT and ResNet both showed comparable performances with a combined total of 5 each. The CNN-based EfficientNet displayed the poorest performance, yielding the highest error total of 138. When comparing the CNN-based, Transformer-based, and combined CNN–Transformer models, it is evident that the ResNet-based and combined CNN–Transformer methods generally offer superior performances.

However, when considering the counts of Heatmap Misses, the proposed network CECvT showed outstanding performance with only 3,572 misses. In comparison, CvT, which had a favorable error performance, had 6,590 misses, while ResNet yielded 8,651. Based on the Accuracy metric, which incorporates the Heatmap Miss count, CECvT had 96.17%, outperforming the second-best CvT with 92.94% by a margin of 3.23%. Networks based on the Transformer architecture displayed the most inferior performances, but it is evident that the combined CNN–Transformer networks were generally excellent.

Upon examining the heatmaps in Figure 7, the visual findings align with the quantitative results in Table 2. Heatmaps from EfficientNet and the Transformer-based networks generally entail inaccuracies in pinpointing the exact locations. In stark contrast, CECvT, which displays the best performance, accurately detects the precise anomaly locations. Furthermore, CECvT accurately detects even minor anomalies, suggesting that it has been effectively trained on the most appropriate features for thermal-image-based anomaly detection.

B. DETECTION CAPABILITY FOR INITIAL ABNORMALITY OCCURRENCE AND COMPARISONS WITH STATE-OF-THE-ART STUDIES

The second experiment assessed the thermal images used for initial anomaly detection on the evaluation data by

segmenting the anomaly manifestations into three distinct phases: initial, middle, and last. This evaluation aimed to test the detection capability performance at each phase. The proposed network CECvT was benchmarked against the same set of eight networks used previously.

Figure 8 depicts the network-specific heatmaps for each of the three phases. For ResNet, the anomalies were correctly identified during the initial phase, but the subsequent middle and last phases were uniformly classified as normal. SEResNet consistently detected anomalies across all phases, but accurate heatmap localization was only evident in the initial phase; both middle and last phases exhibited heatmap misses. EfficientNet classified the images as normal across all phases, with every heatmap outcome demonstrating misses. ViT identified all phases as defects, and the heatmaps predominantly focused on inaccurate locations, which was a trend also observed for SwinT. The CvT heatmap, despite consistently diagnosing defects across all phases, was predominantly concentrated around pipelines. CMT identified defects in all phases, with its heatmap majorly emphasizing the actual anomaly locations; however, the heatmap's significant focus around nearby pipelines hints at reduced accuracy. CoAtNet classified the initial phase as normal and detected defects in the middle and last phases; the heatmaps for middle and last exhibited broader areas of concentration around the pipelines than for CMT, indicating lower accuracy. Finally, the proposed network CECvT consistently detected anomalies across all phases and accurately emphasized the exact anomaly locations via its heatmap; it is therefore evident that CECvT is the most suitable network for accurately diagnosing and pinpointing anomalies in each phase.

VI. CONCLUSION

This study presents the development of the CECvT network, which harnesses the structural advantages of both CNN and Transformer architectures. To address the challenge of edge information degradation commonly encountered in thermal images, a contrast enhancement module was introduced. Furthermore, a coupling module designed to leverage multiscale information was integrated to utilize the finer feature details. The CECvT network, which is anchored on thermal imagery, presents significant potential for the early diagnosis of anomalies in industrial equipment. Through performance comparisons, it was ascertained that the CECvT network outperformed the traditional CNN-based, Transformer-based, and hybrid networks CNN–Transformer architectures. Additionally, in performance comparisons across different anomaly emergence phases, CECvT consistently and accurately diagnosed anomalies at all stages, with its heatmaps effectively pinpointing the precise anomaly locations. The research team plans to further validate the efficacy of the proposed network by applying it to real-world industrial setups in subsequent studies.

Beyond thermal imaging, the authors also intend to test CECvT on standard camera imagery to examine its universality and ability to detect early-stage failures across a broad spectrum of data.

REFERENCES

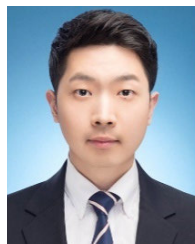
- [1] K. DeMedeiros, A. Hendawi, and M. Alvarez, "A survey of AI-based anomaly detection in IoT and sensor networks," *Sensors*, vol. 23, no. 3, p. 1352, Jan. 2023, doi: [10.3390/s23031352](https://doi.org/10.3390/s23031352).
- [2] L. Tang, H. Tian, H. Huang, S. Shi, and Q. Ji, "A survey of mechanical fault diagnosis based on audio signal analysis," *Measurement*, vol. 220, Oct. 2023, Art. no. 113294, doi: [10.1016/j.measurement.2023.113294](https://doi.org/10.1016/j.measurement.2023.113294).
- [3] J. S. Do, A. B. Kareem, and J.-W. Hur, "LSTM-autoencoder for vibration anomaly detection in vertical carousel storage and retrieval system (VCSRS)," *Sensors*, vol. 23, no. 2, p. 1009, Jan. 2023, doi: [10.3390/s23021009](https://doi.org/10.3390/s23021009).
- [4] B. Azzedine, R. Lias, Z. Zoubir, and G. Mounir, "The influence of the sensor position on the quality of the vibration measurement of rotating machinery on flexible supports," in *Proc. Int. Conf. Syst. Rel. Sci. (ICSRS)*, Nov. 2016, pp. 68–71, doi: [10.1109/ICSRS.2016.7815840](https://doi.org/10.1109/ICSRS.2016.7815840).
- [5] Y. Ota and M. Unoki, "Anomalous sound detection for industrial machines using acoustical features related to timbral metrics," *IEEE Access*, vol. 11, pp. 70884–70897, 2023, doi: [10.1109/ACCESS.2023.3294334](https://doi.org/10.1109/ACCESS.2023.3294334).
- [6] M. H. M. Ghazali and W. Rahiman, "An investigation of the reliability of different types of sensors in the real-time vibration-based anomaly inspection in drone," *Sensors*, vol. 22, no. 16, p. 6015, Aug. 2022, doi: [10.3390/s22166015](https://doi.org/10.3390/s22166015).
- [7] I. J. Aldave, P. V. Bosom, L. V. González, I. L. D. Santiago, B. Vollheim, L. Krausz, and M. Georges, "Review of thermal imaging systems in composite defect detection," *Infr. Phys. Technol.*, vol. 61, pp. 167–175, Nov. 2013, doi: [10.1016/j.infrared.2013.07.009](https://doi.org/10.1016/j.infrared.2013.07.009).
- [8] G. Park, M. Lee, H. Jang, and C. Kim, "Thermal anomaly detection in walls via CNN-based segmentation," *Autom. Construct.*, vol. 125, May 2021, Art. no. 103627, doi: [10.1016/j.autcon.2021.103627](https://doi.org/10.1016/j.autcon.2021.103627).
- [9] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, Mar. 2022, doi: [10.1145/3439950](https://doi.org/10.1145/3439950).
- [10] M. Filipiuk and V. Singh, "Comparing vision transformers and convolutional nets for safety critical systems," in *Proc. SafeAI@ AAAI*, 2022, pp. 1–5.
- [11] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Appl. Sci.*, vol. 13, no. 9, p. 5521, Apr. 2023, doi: [10.3390/app13095521](https://doi.org/10.3390/app13095521).
- [12] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31, doi: [10.1109/ICCV48922.2021.00009](https://doi.org/10.1109/ICCV48922.2021.00009).
- [13] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully convolutional cross-scale-flows for image-based defect detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1829–1838.
- [14] C. Henry, S. Poudel, S.-W. Lee, and H. Jeong, "Automatic detection system of deteriorated PV modules using drone with thermal camera," *Appl. Sci.*, vol. 10, no. 11, p. 3802, May 2020, doi: [10.3390/app10113802](https://doi.org/10.3390/app10113802).
- [15] S. Lee, Y. Chung, R. Shrestha, and W. Kim, "Automated defect detection using threshold value classification based on thermographic inspection," *Appl. Sci.*, vol. 11, no. 17, p. 7870, Aug. 2021, doi: [10.3390/app11177870](https://doi.org/10.3390/app11177870).
- [16] A. Glowacz, "Thermographic fault diagnosis of ventilation in BLDC motors," *Sensors*, vol. 21, no. 21, p. 7245, Oct. 2021, doi: [10.3390/s21217245](https://doi.org/10.3390/s21217245).
- [17] E. Resendiz-Ochoa, J. J. Saucedo-Dorantes, J. P. Benitez-Rangel, R. A. Osornio-Rios, and L. A. Morales-Hernandez, "Novel methodology for condition monitoring of gear wear using supervised learning and infrared thermography," *Appl. Sci.*, vol. 10, no. 2, p. 506, Jan. 2020, doi: [10.3390/app10020506](https://doi.org/10.3390/app10020506).
- [18] C. Kim, J.-S. Choi, H. Jang, and E.-J. Kim, "Automatic detection of linear thermal bridges from infrared thermal images using neural network," *Appl. Sci.*, vol. 11, no. 3, p. 931, Jan. 2021, doi: [10.3390/app11030931](https://doi.org/10.3390/app11030931).
- [19] K. Tong, Z. Wang, L. Si, C. Tan, and P. Li, "A novel pipeline leak recognition method of mine air compressor based on infrared thermal image using IFA and SVM," *Appl. Sci.*, vol. 10, no. 17, p. 5991, Aug. 2020, doi: [10.3390/app10175991](https://doi.org/10.3390/app10175991).
- [20] J. J. V. Díaz, M. Vlaminck, D. Lefkaditis, S. A. O. Vargas, and H. Luong, "Solar panel detection within complex backgrounds using thermal images acquired by UAVs," *Sensors*, vol. 20, no. 21, p. 6219, Oct. 2020, doi: [10.3390/s20216219](https://doi.org/10.3390/s20216219).
- [21] K. Stypułkowski, P. Gołda, K. Lewczuk, and J. Tomaszewska, "Monitoring system for railway infrastructure elements based on thermal imaging analysis," *Sensors*, vol. 21, no. 11, p. 3819, May 2021, doi: [10.3390/s21113819](https://doi.org/10.3390/s21113819).
- [22] B. Kim, S.-W. Choi, G. Hu, D.-E. Lee, and R. O. S. Juan, "Multivariate analysis of concrete image using thermography and edge detection," *Sensors*, vol. 21, no. 21, p. 7396, Nov. 2021, doi: [10.3390/s21217396](https://doi.org/10.3390/s21217396).
- [23] K.-H. Fanchiang and C.-C. Kuo, "Application of thermography and adversarial reconstruction anomaly detection in power cast-resin transformer," *Sensors*, vol. 22, no. 4, p. 1565, Feb. 2022, doi: [10.3390/s22041565](https://doi.org/10.3390/s22041565).
- [24] X. Gong, Q. Yao, M. Wang, and Y. Lin, "A deep learning approach for oriented electrical equipment detection in thermal images," *IEEE Access*, vol. 6, pp. 41590–41597, 2018, doi: [10.1109/ACCESS.2018.2859048](https://doi.org/10.1109/ACCESS.2018.2859048).
- [25] Z. Jia, Z. Liu, C.-M. Vong, and M. Pecht, "A rotating machinery fault diagnosis method based on feature learning of thermal images," *IEEE Access*, vol. 7, pp. 12348–12359, 2019, doi: [10.1109/ACCESS.2019.2893331](https://doi.org/10.1109/ACCESS.2019.2893331).
- [26] M. Siami, T. Barszcz, J. Wodecki, and R. Zimroz, "Automated identification of overheated belt conveyor idlers in thermal images with complex backgrounds using binary classification with CNN," *Sensors*, vol. 22, no. 24, p. 10004, Dec. 2022, doi: [10.3390/s222410004](https://doi.org/10.3390/s222410004).
- [27] M. Vlaminck, R. Heidbuchel, W. Philips, and H. Luong, "Region-based CNN for anomaly detection in PV power plants using aerial imagery," *Sensors*, vol. 22, no. 3, p. 1244, Feb. 2022, doi: [10.3390/s22031244](https://doi.org/10.3390/s22031244).
- [28] G. Chen, W. Wang, and S. Tan, "IRSTFormer: A hierarchical vision transformer for infrared small target detection," *Remote Sens.*, vol. 14, no. 14, p. 3258, Jul. 2022, doi: [10.3390/rs14143258](https://doi.org/10.3390/rs14143258).
- [29] T. Wu, B. Li, Y. Luo, Y. Wang, C. Xiao, T. Liu, J. Yang, W. An, and Y. Guo, "MTU-Net: Multilevel TransUNet for space-based infrared tiny ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601015, doi: [10.1109/TGRS.2023.3235002](https://doi.org/10.1109/TGRS.2023.3235002).
- [30] National Information Society Agency (NIA). *AI Hub*. Accessed: Sep. 4, 2023. [Online]. Available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=235>
- [31] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new image contrast enhancement algorithm using exposure fusion framework," in *Computer Analysis of Images and Patterns*. 2017, pp. 36–46, doi: [10.1007/978-3-319-64698-5_4](https://doi.org/10.1007/978-3-319-64698-5_4).
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [33] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: [10.1109/TPAMI.2019.2913372](https://doi.org/10.1109/TPAMI.2019.2913372).
- [34] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, May 2019, pp. 6105–6114.
- [35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

- [36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [37] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12175–12185.
- [38] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoatNet: Marrying convolution and attention for all data sizes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3965–3977.



DONGHYUN KIM received the B.S. and M.S. degrees in electronics and information engineering from Korea University, South Korea, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree in medical artificial intelligence with Eulji University, South Korea.

Since 2020, he has been the AI Team Leader with Synapse Imaging. His research interests include medical image processing, anomaly detection, computer vision, and artificial intelligence.



HOSEONG HWANG received the B.S. degree in radiological science from Eulji University, South Korea, in 2018, where he is currently pursuing the M.S. degree in medical artificial intelligence and the degree in machine intelligence convergence systems.

From 2018 to 2020, he was an Associate Research Engineer in medical application with Eulji University. From 2020 to 2022, he was a Radiologic Technologists in radiology with the Shihwa Medical Center. His research interest includes medical image processing with artificial intelligence.



HOCHUL KIM received the B.S. degree in applied electronics engineering and the M.S. and Ph.D. degrees in medical and biological engineering from Korea University, South Korea, in 2002, 2004, and 2009, respectively.

From 2009 to 2010, he was a Senior Researcher with the Korea Electrotechnology Research Institute, South Korea. From 2009 to 2012, he was a Research Assistant Professor with the College of Life Science and Biotechnology, Dongguk University, South Korea. Since 2012, he has been a Professor with the Department of Radiological Science, Eulji University, South Korea. His research interests include medical image processing, artificial intelligence, computer vision, radiation detection, and conservation voltage reduction.

• • •