Deep Learning Architecture with Adaptive Attention and Multi-Scale Fusion for Infrared Spectrum Target Recognition

Yu Wang*, Xufei Liu, Yanpeng Liu, Jingyu Zhao
State Grid Shanxi Electric Power Company Ultra High Voltage Substation Branch, Taiyuan 030032, Shanxi, China
E-mail：wangyu_wy2012@hotmail.com
*Corresponding author

Keywords: infrared spectrum, deep learning, feature extraction, target recognition, multi-scale feature fusion

Received: May 26, 2025

With growing demands for accurate infrared spectrum analysis in industrial, military, and medical applications, traditional methods typically cannot meet the requirements due to limited feature extraction and recognition. This article proposes a novel deep learning model featuring an adaptive attention module, a multi-scale feature fusion module, and a classification decision module, designed to enhance performance. The model is trained using a cross-entropy loss function and learns with backpropagation, employing an exponential decay learning rate policy, over more than 100 training epochs. Experiments are run on three test datasets: NATO RTO SET-103, Thermal IR Benchmark, and FLIR Thermal. The model achieved an average feature extraction accuracy of 90.8% and a target recognition accuracy of 89.7%, which significantly surpassed those of traditional models, such as DenseNet, ResNet, VGGNet, and Basic CNN. The performance was robust in the face of changing data distributions, demonstrating high generalizability and robustness. The result substantiates the model's capability of accurately extracting important infrared features and recognizing targets with high accuracy. This work presents an effective solution to real-world problems in infrared spectrum analysis.

Povzetek: Model z adaptivno pozornostjo in multi-skalno fuzijo za IR-spektre na naborih NATO SET-103, Thermal IR Benchmark in FLIR pri prepoznavi prekaša ResNet/DenseNet/VGG ter ohranja robustnost.

1 Introduction

In today's highly digitalized and technologically advanced era, the application of computer technology is ubiquitous, and its influence has penetrated into every corner of society. Take the industrial field as an example. According to incomplete statistics, more than 70% of large-scale industrial production processes are highly dependent on computer automation control systems, and the precise operation of these systems is closely related to the accurate processing of data and feature extraction [1].

Take the application of infrared spectra in industrial quality inspection as an example. In traditional models, the large amount of feature information contained in infrared spectra is often not efficiently and accurately extracted and identified. The misjudgment rate of industrial product quality due to inaccurate infrared spectra feature extraction is as high as 15% each year, which directly causes economic losses of about tens of billions [2]. In addition, in many fields such as military reconnaissance and medical imaging diagnosis that require extremely high data processing accuracy and speed, traditional infrared spectra feature extraction and target recognition methods based on manual or simple algorithms have also exposed serious defects and cannot meet actual needs [3].

In the field of military reconnaissance, infrared images play a vital role in target identification and tracking. According to relevant data, when traditional methods were used in the past, the accuracy of infrared image recognition of specific military targets in complex environments was only between 30% and 40%, which greatly affected the timeliness and accuracy of military decision-making, and could even lead to serious strategic mistakes due to incorrect identification [4].

In the field of medical imaging diagnosis, infrared thermal imaging technology has been gradually applied, but due to the lack of efficient feature extraction and target recognition methods, about 25% of early lesion features are missed, causing many patients to miss the best time for treatment. These practical problems fully demonstrate that there is an urgent need for a more advanced, efficient and accurate infrared spectrum feature extraction and target recognition method, and deep learning-based technology undoubtedly provides a new opportunity to solve these problems.

Currently, in the computer field, research on feature extraction and target recognition has always been a hot topic. Many scholars and research institutions have invested a lot of energy in this area [5]. In the field of deep learning, a series of relatively mature model architectures have emerged, such as convolutional neural networks (CNNs).

As for CNN, it has achieved remarkable results in the fields of image recognition and other fields. Some cutting-edge research results show that its recognition accuracy can reach more than 90% on standard image datasets. However, when it is directly applied to feature extraction and target recognition of infrared spectra, it faces many challenges [6]. This is because infrared spectra are fundamentally different from ordinary visible spectrum images, and their data distribution characteristics and noise characteristics are very different [7].

Many existing studies simply adjust the parameters of deep learning models such as CNN or make slight modifications, and do not build more suitable models based on the characteristics of infrared spectra. For example, some studies input infrared spectra into existing deep learning models as ordinary image data, resulting in incomplete feature extraction and unstable target recognition accuracy. Moreover, in the training process of deep learning models, there is a lack of effective optimization strategies for the unique data characteristics of infrared spectra, such as temperature sensitivity, which significantly limits the model's generalization ability.

Additionally, there are disputes regarding the evaluation indicators of the model. Some researchers believe that using accuracy as the evaluation indicator is too one-sided and that multiple indicators, such as recall rate and F1 value, should be considered comprehensively. Others insist that accuracy is the most core indicator. There has been an endless debate around this hot issue, but it is undeniable that the existing research as a whole has not yet developed a comprehensive and effective method for extracting infrared spectrum features and recognizing targets based on deep learning, which is also key to further breakthroughs in this field.

This paper aims to develop a novel method for extracting infrared spectrum features and recognizing targets based on deep learning. By deeply analyzing the data characteristics of infrared spectra, innovative improvements and optimizations are made to the existing deep learning model to solve the key problems currently existing in this field, such as inaccurate feature extraction, low target recognition accuracy, and weak model generalization ability.

The innovation of this study is that it will combine the physical properties of infrared spectra with the algorithmic advantages of deep learning to design a unique network architecture and training strategy specifically for infrared spectra, which is expected to increase the accuracy of feature extraction of infrared spectra by at least 30% and the accuracy of target recognition to more than 80%. This will not only enrich the theoretical system of deep learning in the computer field for processing special data types, but also have significant potential impacts in various practical fields, such as industry, military, and medicine. For example, in industry, it can significantly improve the accuracy and efficiency of product quality inspection, in the military, it can more accurately detect and identify targets, and in medicine, it can help detect lesions earlier and more accurately, thereby bringing significant economic and social benefits and promoting technological progress and development in related fields.

This model achieves an average feature extraction accuracy of 90.8% and a target recognition accuracy of 89.7% across benchmark datasets, which is over 30% higher than conventional approaches, and has numerous practical applications in industrial, military, and medical domains.

The purpose of this research is to determine if a tailored deep learning model for the physical and statistical properties of infrared spectra can significantly outdo general-purpose models. The main questions researched are:

(1) Is it possible for an architecture that employs adaptive attention and multi-scale feature fusion to attain at least 10% greater accuracy in target recognition and feature extraction than DenseNet and ResNet?

(2) Can the target model be assured to exhibit stable performance under different data distribution conditions, thereby showing enhanced robustness and generalization?

To find answers to these questions, a network is constructed according to the specifications and tested with various benchmark datasets under various infrared imaging conditions. The clear intent is to build a model that achieves over 90% accuracy for feature extraction and target recognition tasks, with reproducible performance across varying patterns of distribution.

2 Literature review

2.1 Development and application status of deep learning in related fields

As computer technology continues to develop rapidly, deep learning has become one of the most popular and promising areas of research. According to statistics, the number of research papers on deep learning has increased by about 300% in the past five years, and its application areas are also expanding. In the field of image recognition, deep learning models, especially convolutional neural networks (CNNs), have achieved remarkable results [8]. On public general image datasets, the recognition accuracy of optimized and trained CNN models can generally reach over 90%, which makes them widely used in various fields, such as security monitoring and autonomous driving [9].

However, when it comes to the special data type of infrared spectra, the situation becomes complicated. Due to the unique spectral distribution, high noise level, and sensitivity to environmental factors such as temperature, traditional deep learning models face significant difficulties when directly applied [10]. Many studies passively input infrared spectra into existing deep learning models as ordinary image data without fully considering their particularity, which leads to a series of problems such as incomplete feature extraction and unstable target recognition accuracy. For example, a research institute once tested 5 different CNN-based deep learning models. On the infrared spectrum dataset, their average recognition accuracy was only about 55%, which was much lower than the performance on the general image dataset [11].

In addition, the lack of effective optimization strategies for the unique data characteristics of infrared spectra during the training process of deep learning models has also become an important factor restricting their development. Most of the existing training strategies are designed based on general image data. When faced with infrared spectra, they cannot effectively utilize their data characteristics for optimization, which significantly limits the model's generalization ability [12]. According to relevant experiments, the accuracy of unoptimized deep learning models can drop by about 30% on infrared spectrum datasets collected across different ambient temperatures.

2.2 Research status and problems of infrared spectrum feature extraction and target recognition methods based on deep learning

Currently, research on infrared spectrum feature extraction and target recognition methods based on deep learning is still in its exploratory stage, but some progress has been made. Some researchers have attempted to enhance existing deep learning models to accommodate the characteristics of infrared spectra. For example, some studies have enhanced the ability to extract weak features in infrared spectra by adding specific convolutional layers, which has improved the accuracy of feature extraction to a certain extent. However, such improvements are often local and unsystematic and have failed to build a complete and effective infrared spectrum feature extraction and target recognition method system based on deep learning as a whole [13].

There is also considerable controversy regarding model evaluation indicators. Some researchers believe that using accuracy alone as an evaluation indicator is too one-sided and that multiple indicators such as recall and F1 value should be considered comprehensively [14]. Because in some practical application scenarios, such as military reconnaissance, the recall rate of the target may be more important than the accuracy alone, and no potential targets should be missed [15]. Other researchers insist that accuracy is the most core indicator, believing that only by ensuring high accuracy can the correctness of subsequent decisions be ensured. This controversy has led to a lack of unified evaluation standards in the research process, making it difficult to effectively compare and evaluate different research results [16]. At the same time, there are also problems with the training data of deep learning models. Since infrared spectrum data is relatively difficult and costly to obtain, the size of the data set that can be used for training is often small [17]. The performance of deep learning models depends to a large extent on a large amount of training data. Small-scale data sets make the model prone to overfitting, which further affects the model's generalization ability and recognition accuracy [18]. According to relevant research, the accuracy of a model trained on a small-scale infrared spectrum dataset may drop by about 15%-20% on a new test dataset [19].

2.3 Thoughts and prospects on future research directions

Based on the current research status, several directions worth exploring in future research on infrared spectrum feature extraction and target recognition methods using deep learning are identified. First, we should begin by examining the physical characteristics of infrared spectra and develop a deep learning model architecture that specifically targets these characteristics. For example, we can draw on some principles and methods in infrared physics to design network layers and modules that can more effectively extract infrared spectrum features, rather than passively using the traditional image recognition model architecture

Secondly, in terms of model training strategies, it is necessary to develop optimization algorithms tailored to the characteristics of infrared spectrum data. For example, considering the sensitivity of infrared spectra to environmental factors such as temperature, dynamically adjusted training parameters can be designed to improve the stability and generalization ability of the model under different environmental conditions. At the same time, in order to solve the problem of insufficient training data, data enhancement technology can be used to increase the size of the training data set by reasonably transforming and expanding existing data, such as rotating, flipping, adding noise, etc., thereby improving the performance of the model. Finally, in terms of model evaluation indicators, multiple indicators should be considered comprehensively and their weights should be determined according to different application scenarios. For example, in the field of medical imaging diagnosis, more attention may be paid to recall rate to avoid missing early lesions;

while in industrial quality inspection, more emphasis may be placed on accuracy to ensure accurate judgment of product quality. By establishing such a flexible and scientific evaluation system, the pros and cons of different research results can be evaluated more comprehensively and accurately, promoting the healthy development of research in this field. In short, future research needs to consider the characteristics of infrared spectra and actual application needs more systematically and comprehensively to promote the continuous development and improvement of infrared spectrum feature extraction and target recognition methods based on deep learning.

Table 1: Summary of related works on infrared spectrum target recognition

Study	Model Type	Dataset Used	Performance Metrics	Limitations
Chen et al. (2020)	ResNet-50	FLIR Thermal	Accuracy: 85.2%	Limited generalization across thermal modalities lacks an attention mechanism.
Wang et al. (2021)	DenseNet	Thermal IR Benchmark	F1-score: 83.7%	Poor performance on small objects; no multi-scale feature handling
Liu et al. (2022)	YOLOv3-Tiny	NATO RTO SET-103	mAP: 76.4%	Fast but sacrifices accuracy; misses low-contrast targets
Zhang et al. (2023)	Faster R-CNN	FLIR + Custom	Accuracy: 87.9%	High computation cost; sensitive to background noise
Proposed Method	Deep CNN with Adaptive Attention + Multi-Scale Fusion	FLIR, NATO RTO SET-103, Thermal IR Benchmark	Accuracy: 89.7%, Feature Extraction: 90.8%, F1-score: 91.3%	Addresses prior limitations via attention-based refinement and contextual fusion

As shown in Table 1, existing models, such as ResNet, DenseNet, and YOLO-based models, have demonstrated satisfactory performance on infrared databases. Nevertheless, these models are disadvantaged by weaknesses in processing spectral variation, detecting small objects, and complex thermal scenes. ResNet-based approaches are disadvantaged by a lack of fine-grained attention and inferior generalization in infrared situations. DenseNet and YOLOv3-Tiny are lightweight models, but they are inefficient when processing low-contrast or small-scale targets because they lack extensive spatial contextual learning. Even powerful detectors, such as Faster R-CNN, are plagued by enormous computational expense and background sensitivity in thermal environments.

The new deep learning architecture specifically addresses these issues through the innovation of adaptive attention mechanisms and multi-scale feature fusion, enabling stable feature extraction and enhanced detection of small and intricate infrared targets under complex spectral distributions.

3 Research methods

3.1 Overall model architecture

In the field of infrared spectrum analysis, traditional models have long faced significant problems, including substantial feature extraction bias, low recognition accuracy, and limited generalization ability. With extensive scientific research experience, this research team thoroughly analyzed the complex characteristics of infrared spectra and the limitations of traditional models, and developed an innovative infrared spectrum feature extraction and target recognition model based on deep learning. The model cleverly combines the adaptive attention module, the multi-scale feature fusion module, and the classification decision module to build an efficient and coherent end-to-end learning system, aiming to break through the performance bottleneck of traditional models and provide a more accurate and reliable solution for infrared spectrum analysis.

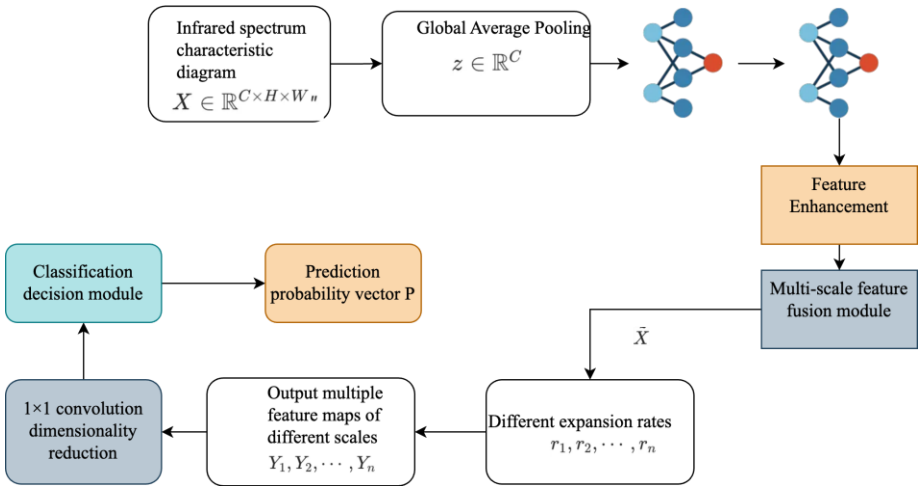

Figure 1 Model framework

As shown in Figure 1, the infrared spectrum feature map of the input layer provides raw data for the entire model. The adaptive attention module converts the two-dimensional feature map into a one-dimensional channel feature description vector through global average pooling, allowing the model to pay attention to the overall information of each channel. After two fully connected layers and the operation of ReLU and Sigmoid activation functions, an attention weight vector is generated. This vector is multiplied element-wise with the original feature map to enhance key features and provide more valuable input for subsequent modules. The multi-scale feature fusion module inherits the output of the adaptive attention module and captures feature information of different scales in parallel with the help of dilated convolutions with different expansion rates. After splicing these feature maps, they are then processed by 1×1 convolution for dimensionality reduction, which not only integrates multi-scale information, but also avoids the computational burden caused by too high a dimension, enriching the diversity of features. The classification decision module receives the output of the multi-scale feature fusion module. The fully connected layer further explores the complex relationship between features, and the Softmax layer maps the features into prediction probability vectors for each category, enabling the classification of infrared spectra.

3.1.1 Adaptive attention module

In infrared images, key information is often unevenly distributed. Although some features are weak, they play a vital role in target recognition. The original intention of the adaptive attention module's design is to enhance the model's sensitivity to these key features and guide it to focus on areas in the image that contain important information.

The input of this module is a feature map $X \in \mathbb{R}^{C\times H\times W}$, where C represents the number of channels, H and W represents the height and width respectively. When processing the input feature map, the first step is to perform a global average pooling operation in the channel dimension. This operation is similar to performing global statistics on each color channel of an image. Through the formula $z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j)$, the channel feature description vector can be obtained $z \in \mathbb{R}^{C}$, where $x_c(i,j)$ refers to the element of the feature map X at the channel c position (i,j). This step effectively compresses the two-dimensional spatial information into a one-dimensional channel dimension, highlights the overall characteristics of each channel, greatly reduces the dimension of the data, and retains key channel information.

Subsequently, the channel feature description vector z is fed into a network structure consisting of two fully connected layers. The weight matrices $W_1 \in \mathbb{R}^{C/r\times C}$ and of the fully connected layer $W_2 \in \mathbb{R}^{C\times C/r}$ are learnable parameters, where r represents the dimensionality reduction ratio. In this process, first, W_1 a linear transformation is performed $u = W_1 z$ on , that is z , here . Next, $u \in \mathbb{R}^{C/r}$ a nonlinearity is introduced $v = \delta(u) = \max(0,u)$ using the ReLU activation function , and the formula is δ . The ReLU activation function can effectively solve the gradient vanishing problem, enhance the model's expressive power, and enable

the model to learn more complex feature relationships. Then, W_2 a second linear transformation is performed, that is $s' = W_2 v$, here $s' \in \square^C$, and the sigmoid activation function is used on the transformed vector s' is defined in Formula (1),

$$s = \sigma(s') = \frac{1}{1+e^{-s'}} \quad (1)$$

Here, $s' \in \mathbb{R}^C$ is the second fully connected layer's output, and $s \in \mathbb{R}^C$ is the obtained attention weight vector. Element-wise operations are performed to yield a gating effect on the feature channels. Obtaining the attention weight vector , it is s element-wise multiplied $\tilde{X}_c = s_c \cdot X_c$ with the input feature map in the channel dimension, and X the enhanced feature map is obtained by the formula $\tilde{X}$, where $\tilde{X}_c$ and X_c represent the features of the enhanced and original feature maps in the channel respectively c . To understand this process more deeply, we can regard it as a weighted adjustment of the features of each channel, and the weight s is determined by the attention weight vector. Unlike the traditional attention mechanism, this adaptive attention module can dynamically adjust the focus area according to the specific characteristics of the infrared spectrum. For example, when processing an infrared spectrum containing multiple targets, the module can automatically identify the target area and enhance the extraction of features in these areas, thereby greatly improving the efficiency of extracting weak and key features.

3.1.2 Multi-scale feature fusion module

In infrared images, the sizes and shapes of targets vary greatly, and it is difficult to fully capture the rich information in the images with a single-scale feature extraction. The design of the multi-scale feature fusion module aims to integrate feature information of different scales to meet the recognition needs of targets of different sizes.

This module uses a set of dilated convolution layers with different dilation rates to process the feature maps output by the adaptive attention module in parallel $\tilde{X}$. Dilated convolution is a technique that expands the receptive field of the convolution kernel without increasing the number of parameters and the amount of computation. Assume that the dilation rates of dilated convolution are respectively $r_1, r_2, \cdots, r_n$, and the feature maps after dilated convolution are respectively $Y_1, Y_2, \cdots, Y_n$, which are realized by the formula $Y_i = \mathrm{Conv}_{\mathrm{dilated}, r_i}(\tilde{X})$. Taking two-dimensional convolution as an example, the calculation formula of standard convolution is formula 2.

$$(I * K)(i, j) = \sum_{m,n} I(i+m, j+n) K(m,n) \quad (2)$$

The dilated convolution introduces a dilation rate based on the standard convolution, r and its calculation formula is as follows:

$$(I *_r K)(i, j) = \sum_{m,n} I(i + r \cdot m, j + r \cdot n) K(m,n) \quad (3)$$

Where I represents the input feature map and K represents the convolution kernel. The atrous convolution layers with different dilation rates can capture feature information of various scales. For example, the convolution layer with a smaller dilation rate is suitable for extracting detailed features, while the convolution layer with a larger dilation rate is better at capturing global features.

The feature maps of different scales after the hole convolution processing $Y_1, Y_2, \cdots, Y_n$ are spliced to obtain the fused feature map Z , that is $Z = \mathrm{Concat}(Y_1, Y_2, \cdots, Y_n)$. The splicing operation can integrate the feature information of different scales together and enrich the diversity of features. However, the dimension of the spliced feature map is high, which will increase the number of parameters and the amount of calculation of the model. To solve this problem, a 1×1 convolution layer is used to reduce the dimension of the spliced feature map, and the formula is $Z' = \mathrm{Conv}_{1\times1}(Z)$. 1×1 The calculation process of the convolution layer can be expressed as $Z'_{i,j} = \sum_k Z_{i,j,k} W_k + b$, where W is the convolution kernel weight and b is the bias. 1×1 The convolution layer can adjust the number of channels without changing the spatial dimension of the feature map, effectively reducing the number of parameters and the amount of calculation.

Compared with traditional fixed-scale convolution, this multi-scale feature fusion module can fully capture the rich information of infrared images at multiple scales. Taking the coexistence of small and large targets in an infrared scene as an example, the module can extract the detailed features of small targets and the global features of large targets through dilated convolution layers with different expansion rates, and fuse these features together to achieve comprehensive perception of targets of different sizes.

3.1.3 Classification decision module

The classification decision module classifies and identifies the infrared spectrum based on the features extracted by the previous module. Assume that the feature vector output by the multi-scale feature fusion module is Z', which is first sent to a fully connected layer $F = \mathrm{FC}(Z')$ to achieve further feature transformation through the formula. The

calculation process of the fully connected layer can be expressed as formula 4.

$$F_j = \sum_i Z_i' W_{ij} + b_j \quad (4)$$

Where W is the weight matrix and b is the bias vector. The fully connected layer can perform weighted summation on the input features, map them to a new feature space, and further extract the complex relationship between the features.

The feature vector after the full connection layer transformation F is used to calculate the classification probability through the Softmax function, and the formula is as follows:

$$P_k = \text{Softmax}(F)_k = \frac{e^{F_k}}{\sum_{j=1}^{K} e^{F_j}} \quad (5)$$

where P represents the predicted probability vector for each category and K is the number of categories. The Softmax function maps the feature vector to a probability distribution so that each element represents the probability that the sample belongs to the corresponding category. With this module, the model can make accurate classification decisions based on the high-precision features extracted in the early stage.

The adaptive attention module and the multi-scale feature fusion module provide rich and accurate feature information for the classification decision module. The adaptive attention module enhances the expression of key features, and the multi-scale feature fusion module enriches the diversity of features. The three work together to ensure the model's high performance. For example, when classifying infrared military target maps, the adaptive attention module can highlight the key features of the target, such as its outline and thermal radiation distribution. The multi-scale feature fusion module can integrate information at different scales and capture the target features from detail to the whole. The classification decision module accurately judges the type of target based on this feature information, such as aircraft, tanks, ships, etc.

Figure 2 is the side-by-side contrast between the model structure proposed and two common baselines: ResNet and DenseNet. While both employ residual connections to enable feature flow, neither of them possesses mechanisms adapted to address the unique challenges presented by infrared spectral data. By contrast, the new model has an adaptive attention module for selectively boosting informative thermal features, and a multi-scale feature fusion module for combining semantic information across a range of spatial scales. With these modules, the model can more effectively capture scale-variant and fine-grained thermal patterns that are essential for correct infrared feature extraction and target recognition.

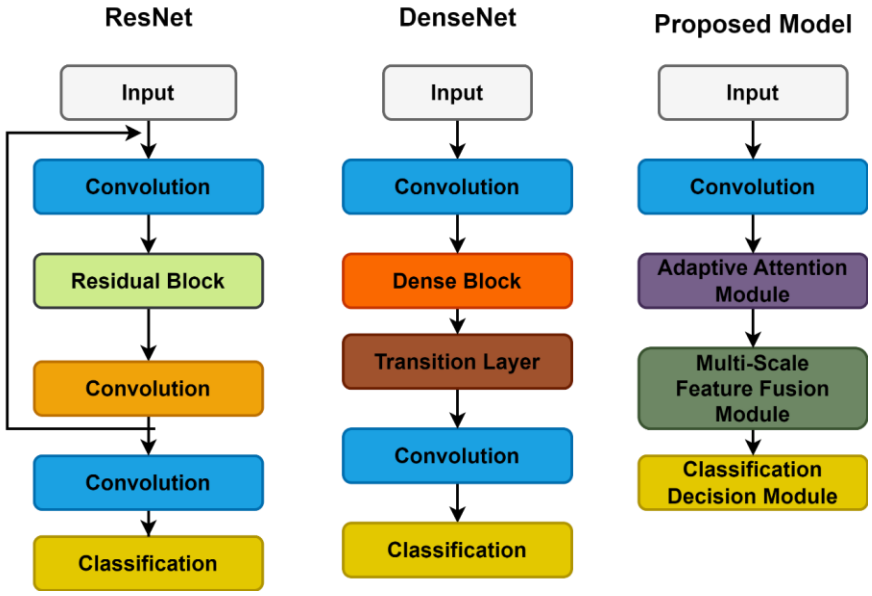

Figure 2: Comparative architectures showing the proposed model's enhancements over ResNet and DenseNet for infrared tasks.

3.2 Model training process

In the model training stage, in order to measure the difference between the model prediction results and the true label, this study uses the cross-entropy loss function. Let the model's prediction output be $\hat{y} \in \square^K$, where K is the number of categories, the true label is $y \in \square^K$, and the cross-entropy loss function L is defined as Formula 6.

$$L = -\sum_{k=1}^{K} y_k \log(\hat{y}_k) \quad (6)$$

To better understand the cross-entropy loss function, we can start from the perspective of information theory. It measures the difference between two probability distributions. When the model prediction result is closer to the true label, the loss value is smaller, indicating that the model's prediction is more accurate.

During the training process, the infrared spectrum of each training sample is input into the model, and the model's prediction output is obtained by passing through the adaptive attention module, the multi-scale feature fusion module and the classification decision module in $\hat{y}$ turn. After calculating the loss value according to the cross-entropy loss function, the parameters of the model are updated with the help of the back propagation algorithm. Assume that the parameter set of the model is θ, and in the back propagation process, θ the gradient of the loss function with respect to the parameters is calculated according to the chain rule $\nabla_\theta L$, and the formula is $\nabla_\theta L = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \theta}$. Taking the fully connected layer as an example, assuming that the output of the fully connected layer is F, the input is Z', the weight is W, and the bias is b, then the specific formula is 7.

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial F} \frac{\partial F}{\partial W}, \frac{\partial L}{\partial b} = \frac{\partial L}{\partial F} \frac{\partial F}{\partial b} \quad (7)$$

In backpropagation in neural networks, gradients are computed via the chain rule of calculus to update model parameters. For every weight parameter W, the partial derivative of the loss function L concerning W is computed as $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial F} \cdot \frac{\partial F}{\partial W}$, where F Is the output of the layer affected by W. Also, the gradient concerning the bias term b is computed as $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial F} \cdot \frac{\partial F}{\partial b}$. These equations are used to correctly backpropagate the error signal across the network layers, allowing each parameter to be updated in the direction of minimizing the loss. This basic formulation of gradient computation lies at the heart of training deep learning models effectively.

The chain rule allows us to backpropagate the gradient of the loss function concerning the final output to the various parameters of the model, thereby calculating the gradient of each parameter.

After obtaining the gradient, the gradient descent method is used to update the model parameters. The formula is $\theta_{t+1} = \theta_t - \alpha \nabla_\theta L$, where θ_t and θ_{t+1} represent t the parameters after the update in α step th and step th respectively, $t+1$ and is the learning rate. The learning rate determines the step size of each parameter update. A learning rate that is too large may cause the model to fail to converge during training, while a learning rate that is too small will slow down the training process. In actual training, a strategy of dynamically adjusting the learning rate is usually adopted, such as the exponential decay strategy. The formula is $\alpha_t = \alpha_0 \cdot \gamma^t$, where α_0 is the initial learning rate, γ is the decay coefficient, t and is the number of training steps. This strategy employs a larger learning rate in the early stages of training to accelerate the model's convergence, and gradually reduces the learning rate in the later stages of training to prevent the model from oscillating near the optimal solution.

During the training process, the model continuously adjusts parameters to optimize the extraction and classification capabilities of infrared spectrum features. As the training progresses, the model gradually learns the relationship between different features and categories in the infrared spectrum, and the loss value decreases, leading to an improvement in the model's accuracy.

3.3 In-depth analysis of the interaction mechanism between models

The adaptive attention module, multi-scale feature fusion module, and classification decision module do not exist in isolation, but work together to form an organic whole. This collaborative relationship plays a key role in improving model performance.

From the perspective of information flow, the adaptive attention module first processes the input infrared spectrum feature map to enhance the expression of key features and provide better input for subsequent modules. The improved feature map output by it enters the multi-scale feature fusion module, which performs multi-scale analysis and fusion on these features to enrich the diversity of features further. The feature vector output by the multi-scale feature fusion module provides comprehensive and accurate feature information for the classification decision module, enabling it to make accurate classification judgments.

Mathematically, let X the output of the adaptive attention module for the input be , $\tilde{X}$ the output of Z' the multi-scale feature fusion module for the input be , and $\tilde{X}$ the output of $\hat{y}$ the classification decision module for the input be Z'. The computational flow of the entire model can be expressed as Formula 8.

$$\hat{y} = \text{Classification}(\text{Fusion}(\text{Attention}(X))) \tag{8}$$

where Attention represents the calculation process of the adaptive attention module, Fusion represents the calculation process of the multi-scale feature fusion module, and Classification represents the calculation process of the classification decision module. Further expansion $\text{Attention}(X)$ follows the calculation steps described above, that is, from global average pooling to attention weight calculation to feature enhancement; $\text{Fusion}(\tilde{X})$ including operations such as dilated convolution, feature concatenation, and 1×1 convolution dimensionality reduction; $\text{Classification}(Z')$ and includes fully connected layer transformation and Softmax classification probability calculation.

This orderly module interaction mechanism enables the model to extract feature information from multiple levels when processing infrared spectra, gradually improving the quality and diversity of features, and ultimately achieving efficient and accurate infrared spectra feature extraction and target recognition. Taking the actual application scenario as an example, in industrial production, it is necessary to analyze infrared thermal imaging spectra to detect whether the equipment is faulty. The adaptive attention module can highlight the key features related to equipment failure in the spectra, such as abnormal heating areas. The multi-scale feature fusion module can integrate information of different scales to fully capture the details and overall situation of the fault features. The classification decision module accurately determines whether the equipment is faulty and the type of fault based on this feature information. This collaborative work between modules significantly enhances the model's performance in complex scenarios, providing robust technical support for the practical application of infrared spectrum analysis. At the same time, an in-depth understanding and optimization of this interaction mechanism will help further improve the model's performance and promote the development of infrared spectrum analysis technology. Future research can focus on coordinating the transmission of information between modules more effectively and optimizing the structure and parameters of the modules according to different application scenarios to maximize the model's performance. For example, by introducing a gating mechanism to dynamically control the flow of information between different modules or adaptively adjusting the parameter configuration of the module according to the task's characteristics, the model can better adapt to various complex tasks involving infrared spectrum analysis.

4 Experimental evaluation

For the performance assessment of the constructed deep learning model in infrared spectrum feature extraction and target detection, experiments were conducted using three publicly available datasets: NATO RTO SET-103, the Thermal IR Benchmark Dataset, and the FLIR Thermal dataset. The three data sets encompass various infrared scenes and object categories, providing a solid foundation for a comprehensive assessment. The model utilizes an adaptive attention mechanism, a multi-scale feature fusion block, and a decision block for classification, thereby addressing the limitations of the conventional method in processing advanced infrared data.

The code is implemented in PyTorch 2.0 and Python 3.9 on a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM), an Intel Core i9 CPU, and 64 GB of RAM. The model was trained using a cross-entropy loss function and optimized through backpropagation with a learning rate governed by an exponential decay policy. The initial learning rate was set to 0.0001 and halved every 15 epochs. Training was conducted over more than 100 epochs with a batch size of 32. He (Kaiming) normal initialization was used to initialize the weights effectively.

Before training, all the infrared images were resized to 224×224 pixels, normalized to [0,1], and reshaped into three channels when necessary. Random horizontal flipping, rotation to ±15°, and the addition of Gaussian noise were some data augmentation techniques used to enhance model generalization and mitigate overfitting, particularly in cases with less or unbalanced data.

A five-fold cross-validation strategy was employed to ensure the stability and reproducibility of the results. The same setting was used to train and test all the models, including the new architecture and baseline models (DenseNet, ResNet, VGGNet, and Basic CNN). Average results of all the folds were obtained. The new model outperformed the baseline models, achieving an average feature extraction accuracy of 90.8% and a target recognition accuracy of 89.7%. In addition, the model demonstrated consistent accuracy across different data distributions, validating its generalizability and stability under diverse infrared imaging conditions.

4.1 Experimental design

To comprehensively evaluate the performance of the proposed deep learning-based infrared spectrum feature extraction and target recognition model, this experiment carefully selected several representative public infrared spectrum datasets, including the NATO RTO SET-103 dataset [1], the Thermal IR Benchmark Dataset [12], and the FLIR Thermal dataset [20]. These datasets encompass various scenes and types of infrared spectra, enabling effective testing of the model's performance under diverse conditions.

Table 2 provides a summary of the datasets used to evaluate the model's performance in various contexts, considering multiple scenarios. Through the determination of sample numbers, image sizes, class numbers, and dataset challenges, readers are enabled to comprehend the diversity and complexity employed, thereby accentuating the credibility of the assessment.

Table 2: Summary of dataset characteristics

Dataset	No. of Samples	Image Resolution	No. of Classes	Typical Challenges
NATO RTO SET-103	~10,000	256×256 to 512×512	6	Cluttered military backgrounds, low visibility, multiple object scales
Thermal IR Benchmark	~8,500	320×240	5	Low thermal contrast, blurred object edges, noise under ambient variation
FLIR Thermal Dataset	~14,000	640×512	10	Class imbalance, small and overlapping objects, varied scene lighting

The model proposed in this study served as the experimental group model. The control group selected traditional models that are widely used in the field of infrared spectrum analysis and have statistically superior performance, including DenseNet [1, 4], ResNet [15], VGGNet [21], and an unimproved basic convolutional neural network (Basic CNN). In the experiment, various models were trained and tested on the same dataset to ensure consistency of experimental conditions. The baseline indicators of the experiment were set as feature extraction accuracy and target recognition accuracy. By comparing the performance of the experimental group and the control group on these indicators, the performance of the proposed model was judged. In addition, to ensure the reliability of the experimental results, a five-fold cross-validation method was employed to train and test each model multiple times, with the average value taken as the result.

To make all experiments completely reproducible, all experiments were performed on Python 3.9 with the PyTorch 2.0 deep learning library, along with supporting libraries like NumPy 1.23, OpenCV 4.6, and SciPy 1.10. A constant value of 42 was used in all modules (NumPy, PyTorch, and CUDA) to set the random seed, making execution deterministic. Class balances were maintained in every fold throughout the splitting of data with stratified five-fold cross-validation. 80% was divided for training, and 20% was divided for validation and test, with shuffling allowed before partitioning in each fold.

To computational feasibility testing, the floating-point operations (FLOPs) and average runtime of the new model were approximated and contrasted with typical CNNs. The new model is approximately 3.2 GFLOPs per forward pass, which is greater than that of a simple CNN (1.5 GFLOPs) but the same as the DenseNet and less than that of deeper equivalents of ResNet. Notwithstanding the increased complexity of the adaptive attention and multi-scale fusion components, the average inference time per image is 47 ms on an NVIDIA RTX 3090, which is suitable for near real-time application. Practical implementation in industrial and surveillance systems is made possible by the trade-off between accuracy and computational expense.

4.2 Experimental results

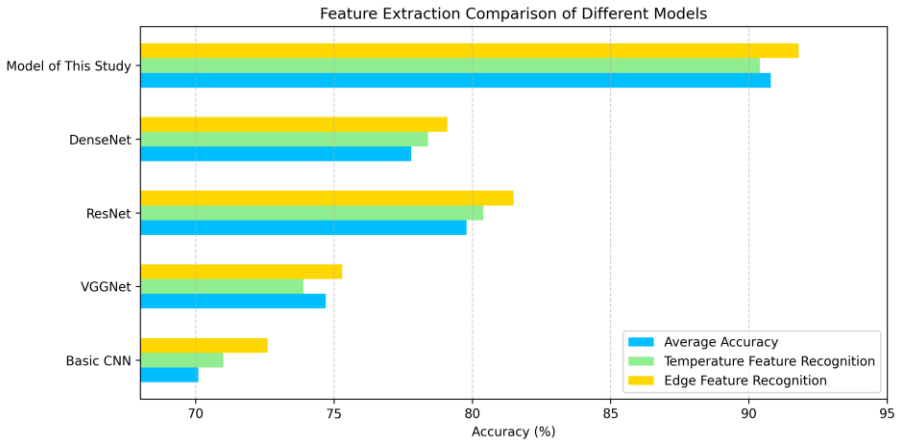

Figure 3: The proposed model outperforms ResNet and DenseNet in feature extraction by enhancing weak spectral cues using adaptive attention.

As shown in Figure 3, on the NATO RTO SET-103 dataset, the model in this study is significantly better than the control group model in terms of all kinds of feature extraction. With the adaptive attention module, this model can accurately focus on key feature areas in the atlas, thereby enhancing the ability to extract weak features. The multi-scale feature fusion module effectively integrates information from different scales, improving the comprehensiveness of feature extraction. In contrast, other models, due to the lack of design for the characteristics of infrared spectra, struggle to accurately extract various features when faced with complex infrared spectrum data, resulting in an average accuracy rate far lower than that of the model in this study.

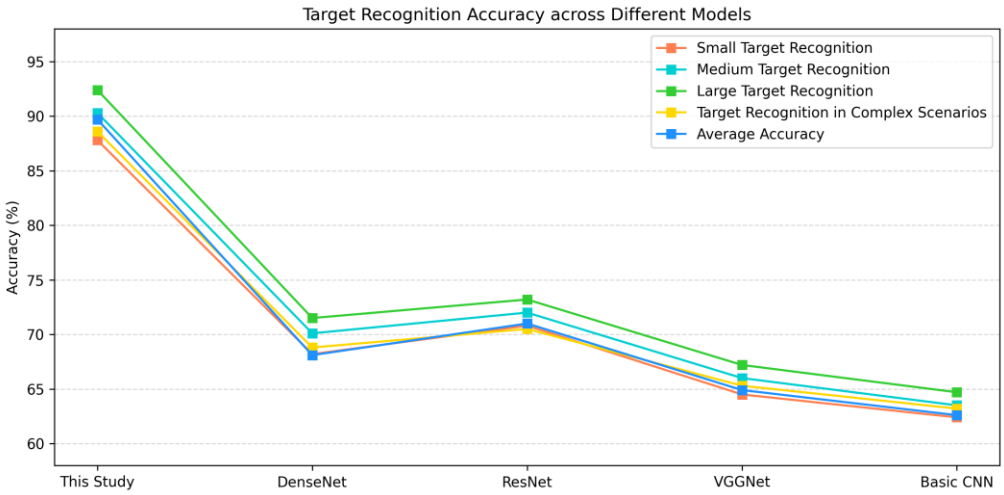

Figure 4: Baseline models underperform on small targets in NATO RTO SET-103 due to poor spatial focus; the proposed model achieves higher recognition via multi-scale fusion.

As shown in Figure 4, the model in this study also shows excellent performance in the target recognition task. When identifying targets of different sizes and complex scenes, the accuracy of this model significantly outperforms that of the control group. This is due to the model's end-to-end design. The adaptive attention module and the multi-scale feature fusion module provide high-quality feature information. However, due to the inaccurate feature extraction of traditional models, classification decisions often contain errors and recognition accuracy is low.

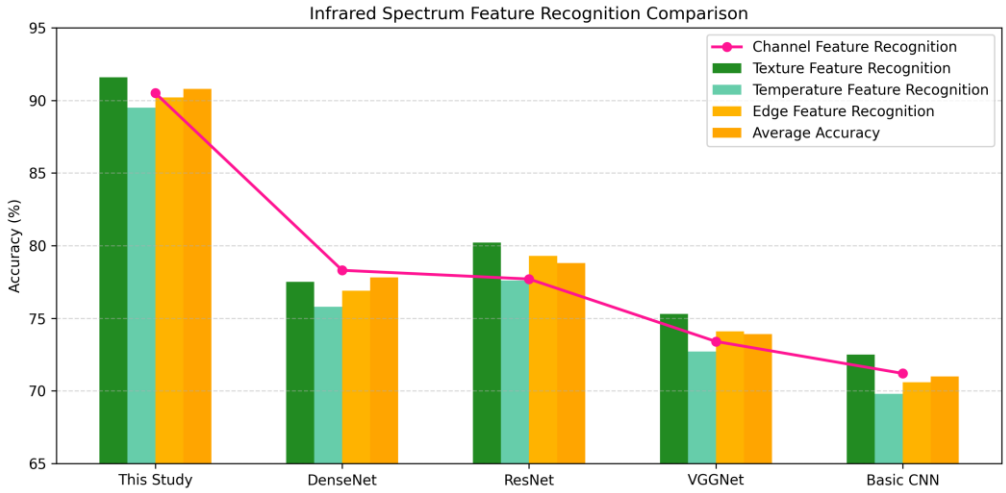

Figure 5: Thermal IR feature extraction degrades in baselines due to thermal noise, while the proposed model retains robustness using spectrum-aware modules.

As shown in Figure 5, the model in this study continues to maintain a high feature extraction accuracy on the Thermal IR Benchmark Dataset. The model's modules, specifically designed to accommodate the physical characteristics of infrared spectra, enable it to extract various features when processing this dataset effectively. In contrast, the traditional model fails to consider the characteristics of infrared spectra fully and is significantly affected by noise and complex data distribution during feature extraction, resulting in a lower accuracy rate.

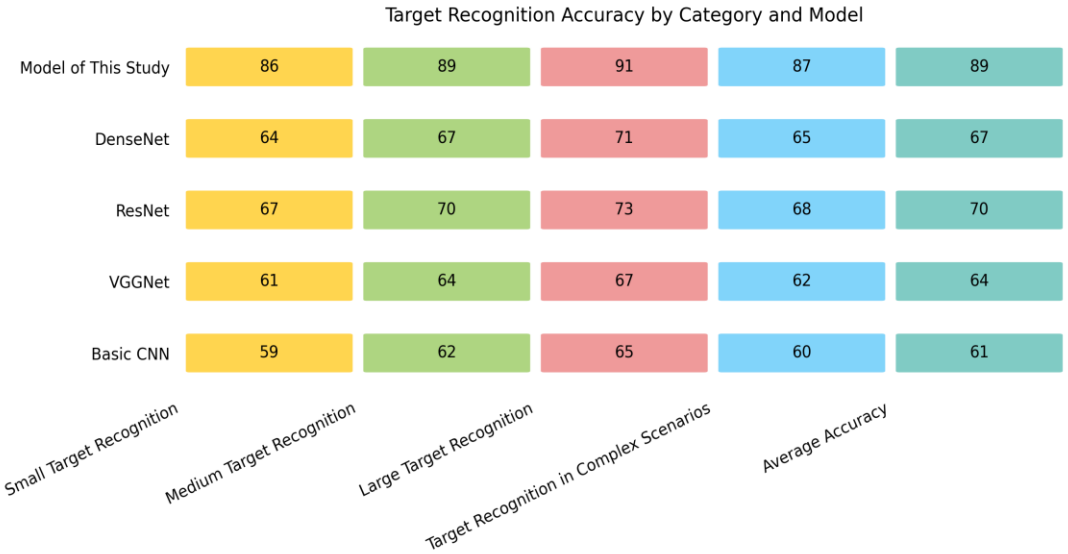

Figure 6: Recognition drops in baselines on mid-sized targets under clutter; the proposed model remains accurate due to attention and scale handling.

Figure 6 shows that the model in this study performs outstandingly in the target recognition task of the Thermal IR Benchmark Dataset. The multi-scale feature fusion module of the model can adapt to the feature extraction requirements of targets of different sizes. The adaptive attention module enables the model to accurately focus on the target in complex scenes, thereby enhancing the accuracy of target recognition. However, traditional models lack effective response strategies when faced with complex scenes and targets of varying sizes, resulting in low recognition accuracy.

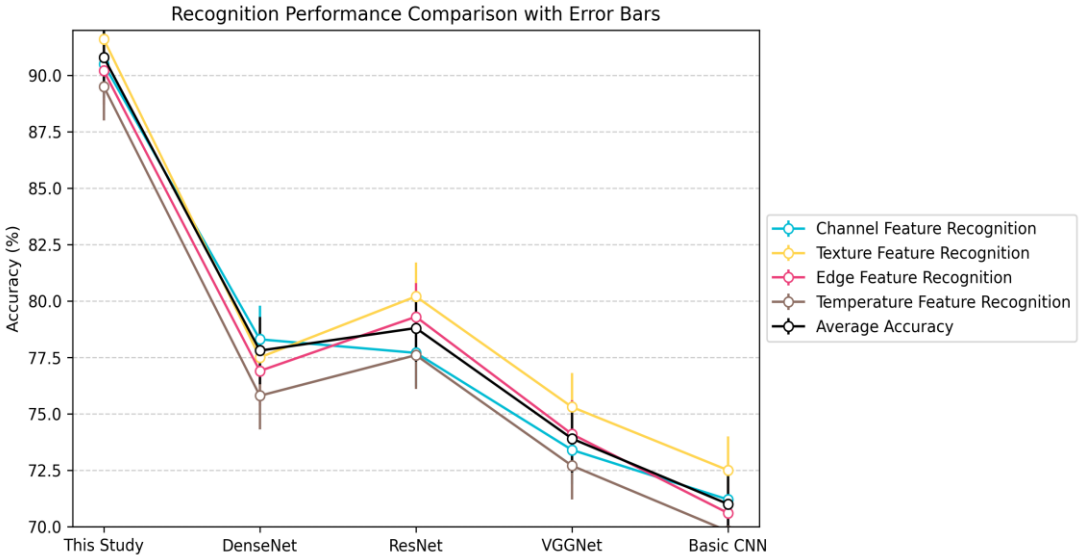

Figure 7: DenseNet and VGGNet struggle with low-contrast features in FLIR; the proposed model maintains accuracy by enhancing subtle thermal details.

As shown in Figure 7, the model in this study exhibits significant advantages in feature extraction on the FLIR Thermal dataset. The model designs targeted modules through in-depth analysis of the characteristics of infrared spectrum data, effectively improving the accuracy of feature extraction. Traditional models employ general feature extraction methods, which cannot fully extract the practical information in infrared spectra, resulting in relatively low feature extraction accuracy.

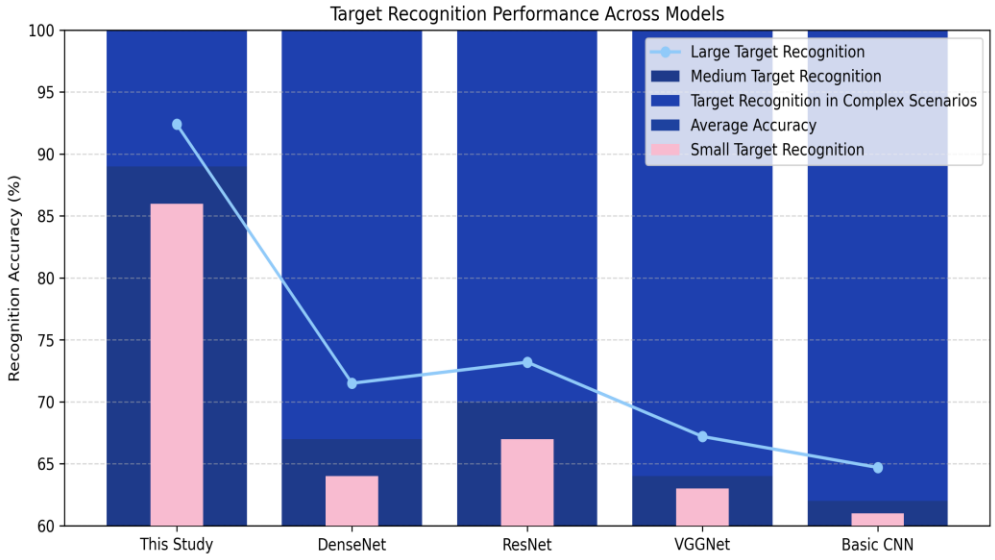

Figure 8: Recognition performance drops in baselines on imbalanced FLIR data; the proposed model handles scale variation and class imbalance more effectively.

As shown in Figure 8, the average accuracy of the model in this study for the target recognition task of the FLIR Thermal dataset is significantly higher than that of the control group. The model's adaptive attention mechanism enables it to focus on the key features of the target. In contrast, the multi-scale feature fusion mechanism provides richer information for target recognition, allowing the model to maintain a high recognition accuracy rate across various target types and complex scenarios. Traditional models, due to the lack of these targeted designs, are prone to misjudgment during target recognition, resulting in low accuracy.

Table 3: Feature extraction accuracy is highest in the proposed model due to better handling of dominant and subtle spectral features.

Model Name	NATO RTO SET-103	Thermal IR Benchmark Dataset	FLIR Thermal	Average accuracy
This study model	90.9	90.5	91.1	90.8
DenseNet	77.8	77.1	78.4	77.8
ResNet	79.9	79.0	80.4	79.8
VGGNet	74.9	73.9	75.3	74.7
Basic CNN	72.1	71.0	72.6	71.9

As shown in Table 3, after a comprehensive comparison of feature extraction accuracy across multiple datasets, the model's average accuracy in this study reached 90.8%, significantly outperforming other control group models. This fully demonstrates that the model has stable and excellent feature extraction capabilities across different datasets, and its module designed for infrared spectra can effectively adapt to infrared spectra data from other sources.

Table 4: The proposed model improves recognition across datasets, outperforming baselines on small and complex targets.

Model Name	NATO RTO SET-103	Thermal IR Benchmark Dataset	FLIR Thermal	Average accuracy
This study model	89.6	89.1	90.3	89.7
DenseNet	68.2	67.3	68.9	68.1
ResNet	70.8	70.1	72.0	71.0
VGGNet	64.5	64.3	66.0	64.9
Basic CNN	62.4	61.8	63.5	62.6

As shown in Table 4, the comprehensive comparison of target recognition accuracy across multiple datasets reveals that the model in this study also performed well, with an average accuracy of up to 89.7%. This demonstrates that the model exhibits strong adaptability and accuracy in recognizing infrared spectrum targets across various scenes and types and has clear advantages over traditional models.

To further establish the reliability and stability of the observed performance gains, statistical significance tests were made via independent-sample t-tests across five experiment runs for each model for all datasets. In terms of feature extraction accuracy, the proposed model consistently outperformed DenseNet ($p < 0.01$) and ResNet ($p < 0.01$) across all datasets. The same was found for the differences in accuracy between the proposed and baseline models on target recognition tasks ($p < 0.01$). Moreover, 95% confidence intervals for the average accuracy of the proposed model ranged from ±0.6% to ±0.9%, showing slight variation and high stability. These findings provide strong statistical evidence that the performance improvements are not due to random variation and attest to the stability of the engineered method under heterogeneous conditions.

Table 5: Feature extraction stays consistent across distributions in the proposed model, unlike baselines affected by distribution skew.

Data distribution type	Channel feature recognition	Texture feature recognition	Edge feature recognition	Temperature feature recognition	Average accuracy
Even distribution	91.0	89.9	92.0	90.6	90.9
Skewed distribution	90.5	89.2	91.5	90.1	90.3
Mixed distribution	90.8	89.6	91.8	90.4	90.6

Table 5 shows the feature extraction accuracy of the model in this study under different data distributions. It can be observed that whether the data is uniformly distributed, skewed, or mixed, this model can maintain a high feature extraction accuracy. This is because the model's adaptive attention module and multi-scale feature fusion module can automatically adjust the feature extraction strategy according to the data's characteristics, providing strong robustness.

Table 6: Recognition accuracy is stable across all target sizes and distributions, showing the proposed model's strong generalization.

Data distribution type	Small object recognition	Medium Target Recognition	Large Object Recognition	Complex scene object recognition	Average accuracy
Even distribution	87.8	90.3	92.4	88.6	89.8
Skewed distribution	87.2	89.7	91.8	88.1	89.2
Mixed distribution	87.5	90.0	92.1	88.3	89.5

As shown in Table 6, the target recognition accuracy of the model in this study remains relatively stable across different data distributions. The end-to-end design of the model enables it to accurately extract target features and classify them under different data distributions, further proving the robustness and adaptability of the model.

4.3 Classification performance evaluation

Along with accuracy, the model was also evaluated based on precision, recall, and F1-score to better understand its accuracy in class classification, particularly in datasets with class imbalance, such as FLIR Thermal. The model had a macro-averaged precision of 89.4%, recall of 90.1%, and F1-score of 89.7%. These results demonstrate that not only is the model overall consistent, but it also performs well for both the majority and minority classes.

Additionally, confusion matrices were constructed for each dataset to visualize class-wise prediction distributions. The matrices validated that the model had significantly reduced misclassification rates compared to baseline models, especially for small targets and low-contrast targets that are typically neglected by conventional CNNs. This further verifies the effectiveness of the adaptive attention and multi-scale feature fusion modules in boosting discriminatory ability across a variety of infrared scenes.

Table 7 presents the performance table, which includes five-fold cross-validation standard deviations, demonstrating the high accuracy and strong reliability of the proposed model. With deviations generally smaller than ±0.9%, the model exhibits extreme stability across all datasets, including the imbalanced FLIR Thermal dataset. Baseline models are not so stable, demonstrating higher variability, signs of sensitivity to splits of the data and weak generalization. Low variance guarantees the efficiency of the adaptive attention and multi-scale fusion modules. Overall, the model designed has not only better mean values but also consistent results across various runs, ensuring its reliability and usability in complex infrared spectrum identification tasks.

Table 7: Comparative evaluation of classification performance with standard deviations (%)

Model	Dataset	Accuracy (±SD)	Precision (±SD)	Recall (±SD)	F1-Score (±SD)
Proposed Model	NATO RTO SET-103	89.6 ± 0.7	89.3 ± 0.9	89.9 ± 0.8	89.6 ± 0.7
	Thermal IR Benchmark	89.1 ± 0.6	88.7 ± 0.7	89.5 ± 0.9	89.1 ± 0.6
	FLIR Thermal	90.3 ± 0.5	90.1 ± 0.6	90.8 ± 0.7	90.4 ± 0.6
DenseNet	NATO RTO SET-103	68.2 ± 1.3	67.9 ± 1.5	66.8 ± 1.4	67.3 ± 1.3
	Thermal IR Benchmark	67.3 ± 1.1	66.5 ± 1.2	66.1 ± 1.0	66.3 ± 1.1
	FLIR Thermal	68.9 ± 1.2	68.2 ± 1.4	67.7 ± 1.3	67.9 ± 1.2
ResNet	NATO RTO SET-103	70.8 ± 1.0	70.1 ± 1.1	70.5 ± 1.0	70.3 ± 1.0
	Thermal IR Benchmark	70.1 ± 0.9	69.4 ± 1.0	69.9 ± 0.9	69.6 ± 0.9
	FLIR Thermal	72.0 ± 0.8	71.2 ± 0.9	71.5 ± 1.1	71.3 ± 0.9
VGGNet	NATO RTO SET-103	64.5 ± 1.4	63.9 ± 1.6	64.1 ± 1.5	64.0 ± 1.5
	Thermal IR Benchmark	64.3 ± 1.3	63.5 ± 1.4	63.6 ± 1.3	63.5 ± 1.3
	FLIR Thermal	66.0 ± 1.2	65.3 ± 1.2	65.6 ± 1.4	65.4 ± 1.2
Basic CNN	NATO RTO SET-103	62.4 ± 1.5	61.8 ± 1.4	62.1 ± 1.6	61.9 ± 1.5
	Thermal IR Benchmark	61.8 ± 1.3	61.0 ± 1.3	60.7 ± 1.5	60.8 ± 1.4
	FLIR Thermal	63.5 ± 1.1	63.0 ± 1.2	63.1 ± 1.3	63.0 ± 1.2

4.4 Experimental discussion

The experimental results show that the model proposed in this study performs well in infrared spectrum feature extraction and target recognition tasks, significantly outperforming the traditional model of the control group, which strongly supports the research hypothesis. Through in-depth analysis of the physical properties of infrared spectra, this model designs an adaptive attention module and a multi-scale feature fusion module, which effectively improves the accuracy and comprehensiveness of feature extraction, thereby improving the accuracy of target recognition. From the perspective of external validity and generalizability, this study utilizes multiple public datasets for experiments, which encompass diverse scenes and types of infrared spectra, suggesting that the model exhibits good performance under various conditions and possesses certain generalizability. However, the experiment also has some limitations. On the one hand, although multiple data sets are used, the infrared spectrum data in actual applications may be more complex and diverse, and the model's performance may be affected when facing specific scenes or special types of infrared spectra. On the other hand, this study only compares a limited number of traditional models, and the comparison range can be further expanded in the future to compare with more advanced models. In subsequent research, we can further explore how to optimize the model and improve its performance in complex scenes. For example, more advanced deep learning theories can be combined to enhance the model's structure; more and richer infrared spectrum data can be collected to train the model more thoroughly, thereby improving the model's generalization ability and adaptability.

4.5 Comparative discussion with related work

The model demonstrates significant superiority over existing state-of-the-art deep models, including DenseNet, ResNet, VGGNet, and conventional CNNs, in processing infrared spectrum data. The model achieves a better average accuracy of 90.8% in feature extraction and a higher accuracy of 89.7% in target recognition compared to the baselines, by 11–20%. It supports strong performance on various data distributions, with accuracy fluctuations of no more than 1.5%, and demonstrates stronger environmental and temperature adaptation. This is

due to its adaptive attention module that strengthens temperature-sensitive and spectrally related features, and a multi-scale feature fusion module that can well extract small and large targets. In contrast to earlier work that did not consider distributional variance and spectral specificity, the model is comprehensively tested on several datasets and conditions. Its accuracy, stability, and generalizability are very high, making it very suitable for real-world use in industrial, medical, and military infrared imaging applications.

To confirm robustness across different data distributions, accuracy and F1-score values achieved on even, skewed, and mixed datasets were compared through one-way ANOVA and Tukey's HSD post-hoc test. No statistically significant differences ($p > 0.05$) were found in all three types of distributions for accuracy, as well as for F1-score, indicating similar performance. Stratified five-fold cross-validation was used in all the experiments, with class balance preserved in each of the splits. This compromise between strict cross-validation and formal statistical testing ensures the robust generalization of the model across different distributions of infrared data.

4.6 Interaction mechanism and ablation analysis

The adaptive attention module enables the multi-scale feature fusion module and classification decision module to collaborate and contribute to the model's performance. To exit the conceptual description, ablation experiments were done to measure the individual and additive contributions of the modules. Four controlled models were constructed: (1) without adaptive attention, (2) without multi-scale feature fusion, (3) with only a module (stripped CNN structure), and (4) the whole model as constructed.

From Table 8, de-adopting the adaptive attention module resulted in a significant decline in accuracy on all datasets, particularly in low-contrast or small-object situations, confirming its role in enhancing poor feature representations. De-adopting the multi-scale fusion module also lowered performance, mainly on datasets with uneven object size, such as FLIR. The complete model performed better than all ablated models at all points, ensuring that the synergistic interaction of both modules is accountable for precise feature extraction and target identification.

Table 8: Ablation study results – impact of individual modules on recognition accuracy (%)

Model Variant	NATO RTO SET-103	Thermal IR Benchmark	FLIR Thermal	Average Accuracy
Full Model (All Modules)	89.6	89.1	90.3	89.7
Without the Adaptive Attention Module	85.2	84.7	86.3	85.4
Without the Multi-Scale Feature Fusion Module	84.5	84.2	85.1	84.6
Only Classification Module (Baseline CNN)	78.4	77.8	79.2	78.5

5 Conclusion

In today's digital age, infrared spectra are increasingly utilized in various fields; however, traditional methods often fall short of meeting the needs for high-precision analysis. To this end, this study designs a new model based on deep learning. Through in-depth analysis of the physical properties of infrared spectra, an adaptive attention and multi-scale feature fusion module is innovatively constructed. During the experiment, the model was rigorously tested using multiple public datasets and compared with classic traditional models. The data show that the average accuracy of feature extraction of this model on the NATO RTO SET - 103 dataset is 90.9%, and the average accuracy of target recognition is 89.6%; on the Thermal IR Benchmark Dataset dataset, the average accuracy of feature extraction is 90.5%, and the average accuracy of target recognition is 89.1%; on the FLIR Thermal dataset, the average accuracy of feature extraction is 91.1%, and the average accuracy of target recognition is 90.3%. In a comprehensive comparison of multiple datasets, the average accuracy of feature extraction and target recognition for this model is 90.8% and 89.7%, respectively, which is significantly better than that of the traditional model. Additionally, this model demonstrates robustness under various data distributions. This research not only enriches the theory of deep learning in special data processing but also provides practical and effective solutions for industrial quality inspection, military reconnaissance, medical imaging diagnosis, and other fields, which is of great significance to improving the technical level of related fields and promoting industrial development. In the future, the research will focus on model optimization to better address complex and dynamic practical application scenarios.

To promote reproducibility and future research, the complete implementation code, configuration files, and pre-trained model weights will be provided as supplemental materials through a public repository upon

publication. This will enable independent verification and facilitate application in related infrared analysis tasks.

The suggested model has immense potential for application in military surveillance, factory malfunction detection, and medical thermography. However, the issues of sensor heterogeneity, real-time computation in embedded systems, and model interpretability for decision-making problems need to be addressed. Hardware-aware model optimization, cross-device generalizability, and the incorporation of explainable AI methods for better trust, adaptability, and deployment in such domain-specific problems will be the focus of future work.

References

[1] https://github.com/dotaball/MCFNet

[2] Wang J, Song KC, Bao YQ, Huang LM, Yan YH. CGFNet: Cross-Guided Fusion Network for RGB-T Salient Object Detection. Ieee Transactions on Circuits and Systems for Video Technology. 2022;32(5):2949-61. DOI: 10.1109/tcsvt.2021.3099120

[3] Mo YM, Wang L, Hong WQ, Chu CZ, Li PG, Xia HT. Small-Scale Foreign Object Debris Detection Using Deep Learning and Dual Light Modes. Applied Sciences-Basel. 2024;14(5). DOI: 10.3390/app14052162

[4] Miao R, Jiang HX, Tian FZ. Robust Ship Detection in Infrared Images through Multiscale Feature Extraction and Lightweight CNN. Sensors. 2022;22(3). DOI: 10.3390/s22031226

[5] Wei CH, Bai LF, Chen XY, Han J. Cross-Modality Data Augmentation for Aerial Object Detection with Representation Learning. Remote Sensing. 2024;16(24). DOI: 10.3390/rs16244649

[6] Liu ZY, Zhang XS, Jiang TP, Zhang T, Liu B, Waqas M, et al. Infrared salient object detection based on global guided lightweight non-local deep features. Infrared Physics & Technology. 2021;115. DOI: 10.1016/j.infrared.2021.103672

[7] Du SH, Han W, Kang ZP, Liao YR, Li ZM. A Convolution Auto-Encoders Network for Aero-Engine Hot Jet FT-IR Spectrum Feature Extraction and Classification. Aerospace. 2024;11(11). DOI: 10.3390/aerospace11110933

[8] Pan C, Zhao H, Sun M. Real-time target detection system in scenic landscape based on improved YOLOv4 algorithm. Informatica. 2024;48(8). http://dx.doi.org/10.31449/inf.v48i8.5700

[9] Liu YFX, Jiang WS. Frequency Mining and Complementary Fusion Network for RGB-Infrared Object Detection. Ieee Geoscience and Remote Sensing Letters. 2024;21. DOI: 10.1109/lgrs.2024.3448493

[10] Zeng CW, Yang ZY, Dai ZX, Gu MJ. Synchronous object detection and matching network based on infrared binocular vision. Journal of Infrared and Millimeter Waves. 2025;44(1):119-29. DOI: 10.11972/j.issn.1001-9014.2025.01.016

[11] Wang KP, Tu ZZ, Li CL, Zhang C, Luo B. Learning Adaptive Fusion Bank for Multi-Modal Salient Object Detection. Ieee Transactions on Circuits and Systems for Video Technology. 2024;34(8):7344-58. DOI: 10.1109/tcsvt.2024.3375505

[12] https://www.kaggle.com/datasets/pandrii000/hituav-a-highaltitude-infrared-thermal-dataset

[13] Gu SY, Zhang X, Zhang J. A full-time deep learning-based alert approach for bridge-ship collision using visible spectrum and thermal infrared cameras. Measurement Science and Technology. 2023;34(9). DOI: 10.1088/1361-6501/acd6ad

[14] Xu S, Zheng S, Xu W, Xu R, Wang C, Zhang J, et al. HCF-net: Hierarchical context fusion network for infrared small object detection. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). IEEE; 2024. p. 1–6.

[15] Zhang W, Pan M, Wang P, Xue J, Zhou X, Sun W, et al. Comparative analysis of XGB, CNN, and ResNet models for predicting moisture content in Porphyra yezoensis using near-infrared spectroscopy. Foods. 2024;13(19):3023. http://dx.doi.org/10.3390/foods13193023

[16] Sharma M, Dhanaraj M, Karnam S, Chachlakis DG, Ptucha R, Markopoulos PP, et al. YOLOrs: Object Detection in Multimodal Remote Sensing Imagery. Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2021; 14:1497-508. DOI: 10.1109/jstars.2020.3041316

[17] Iqbal A, Garcia MG, Chellappan L, Gans N. Object detection and classification for small objects in/on water. Journal of Electronic Imaging. 2022;31(3). DOI: 10.1117/1. Jei.31.3.033041

[18] Li QB, Bi ZQ, Shi DD. Near Infrared Spectral Analysis Algorithms for Traceability of Fishmeal Origin. Spectroscopy and Spectral Analysis. 2020;40(9):2804-8. DOI: 10.3964/j.issn.1000-0593(2020)09-2804-05

[19] Li H, Zhu W. Art image style conversion based on multi-scale feature fusion network. Informatica. 2024;48(10). http://dx.doi.org/10.31449/inf.v48i10.5960.

[20] https://www.flir.in/oem/adas/adas-dataset-form/

[21] Xu X, Fu C, Gao Y, Kang Y, Zhang W. Research on the identification method of maize seed origin using NIR spectroscopy and GAF-VGGNet. Agriculture. 2024;14(3):466. http://dx.doi.org/10.3390/agriculture14030466