文章编号：1001-9014(2025)01-0103-09

DOI:10.11972/j.issn.1001-9014.2025.01.014

Infrared aircraft few-shot classification method based on cross-correlation network

HUANG Zhen[1,2,3], ZHANG Yong[1,3*], GONG Jin-Fu[1,2,3]

(1. Key Laboratory of Infrared System Detection and Imaging Technology, Chinese Academy of Sciences, Shanghai 200083, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China)

Abstract: In response to the scarcity of infrared aircraft samples and the tendency of traditional deep learning to overfit, a few-shot infrared aircraft classification method based on cross-correlation networks is proposed. This method combines two core modules: a simple parameter-free self-attention and cross-attention. By analyzing the self-correlation and cross-correlation between support images and query images, it achieves effective classification of infrared aircraft under few-shot conditions. The proposed cross-correlation network integrates these two modules and is trained in an end-to-end manner. The simple parameter-free self-attention is responsible for extracting the internal structure of the image while the cross-attention can calculate the cross-correlation between images further extracting and fusing the features between images. Compared with existing few-shot infrared target classification models, this model focuses on the geometric structure and thermal texture information of infrared images by modeling the semantic relevance between the features of the support set and query set, thus better attending to the target objects. Experimental results show that this method outperforms existing infrared aircraft classification methods in various classification tasks, with the highest classification accuracy improvement exceeding 3%. In addition, ablation experiments and comparative experiments also prove the effectiveness of the method.

Key words: infrared imaging, aircraft classification, few-shot learning, parameter-free attention, cross attention

基于交叉相关网络的少样本红外空中目标分类方法

黄 臻[1,2,3], 张 湧[1,3*], 公劲夫[1,2,3]

(1. 中国科学院红外探测与成像技术重点实验室,上海 200083;
2. 中国科学院大学,北京 100049;
3. 中国科学院上海技术物理研究所,上海 200083)

摘要:针对红外空中目标样本匮乏、传统深度学习易产生过拟合等问题,提出一种基于交叉相关网络的少样本红外目标分类方法。该方法结合简单无参数自注意力和交叉注意力两个核心模块,通过分析支持图像和查询图像之间的自相关性和互相关性,实现少样本条件下红外目标的有效分类。所提出的交叉相关网络结合了这两个模块,以端到端的方式进行训练。其中,简单无参数自注意力负责提取图像内部结构,交叉注意力可以计算图像之间的互相关,进一步提取并融合图像之间的特征。与现有的小样本红外目标分类模型相比,该模型通过建模支持集和查询集之间特征的语义相关性,聚焦红外图像的几何结构和纹理信息,从而更好地关注目标对象。实验结果表明,该方法在各项分类任务中性能均优于现有的红外空中目标分类方法,且分类准确率最高提升超过3%。此外,消融实验和对比实验也证明了该方法的有效性。

关 键 词:红外成像;空中目标分类;少样本学习;无参数注意力;交叉注意

中图分类号:TP391.4　　**文献标识码**:A

Received date: 2024-03-29, **revised date**: 2024-06-05　　**收稿日期**:2024-03-29,**修回日期**:2024-06-05

Foundation items: Supported by the National Pre-research Program during the 14th Five-Year Plan(514010405)

Biography: HUANG Zhen(1997-), male, Chongqing, China. Ph. D. The research area involves image processing and object detection. E-mail: huangzhen@mail.sitp.ac.cn

* **Corresponding author**: E-mail: zybxy@sina.com

Introduction

In recent years, the rapid advancement of infrared detection and imaging technology has led to an expanding application scope for infrared images[1]. Infrared detection technology, characterized as a passive detection technique, offers advantages such as long-range detection, high concealment and robust all-weather capabilities[2]. Recognition of infrared targets, as a crucial component of infrared detection imaging systems, is paramount for enhancing system performance and expanding application domains. Presently, deep learning models demonstrate remarkable performance in visual recognition tasks such as image classification; however, this significant performance heavily relies on the utilization of large quantities of labeled image data for training[3]. Nevertheless, due to factors such as the high cost and military sensitivity of infrared equipment, acquiring samples of airborne infrared targets poses significant challenges. For certain rare aircraft models, sample data may be limited to only tens or even units[4], implying that even with the adoption of methods such as data augmentation or transfer learning, serious overfitting issues can easily arise.

In contrast, the human visual system has the remarkable ability to rapidly form cognitive frameworks for new entities based on a few examples[5]. This capability enables humans to leverage prior knowledge and experience to quickly learn new tasks without the need for extensive data and time. Inspired by this, methods for few-shot learning to have emerged. The goal of few-shot learning is to design and train models capable of identifying new classes with only a few of annotated examples, akin to the human visual system. Currently, few-shot learning mainly encompasses meta-learning, transfer learning, and metric learning[6]. Meta-learning involves training a meta-learner across various classification tasks to extract generalizable knowledge. Transfer learning assumes shared knowledge between a source domain and a target domain, pre-training the model on a large amount of source data, and then fine-tuning it on the target domain to adapt to its data distribution. Metric learning aims to learn a discriminate distance metric, ensuring that samples from different classes have a large distance in the embedding space, while the distance between samples from the same class is minimized as much as possible.

Infrared images possess unique characteristics, such as low contrast and low signal-to-noise ratio. Moreover, apart from the target objects, infrared images may also contain various background interference, such as buildings and clouds. Therefore, designing a network model that can focus more on the target objects in infrared images under the constraint of extremely limited samples is crucial for our research. Recent advancements in few-shot learning have seen widespread application of meta-learning and transfer learning. Chen *et al.* [2] and Jin *et al.* [4] have employed meta-learning and improved relation networks techniques for infrared aircraft classification. They leverage the metric learning capability of relation networks and the rapid adaptation ability of meta-learning to enhance the accuracy of infrared airborne target classification. However, these methods often independently extract features from the support set and query samples, which can lead to less discriminative feature representations. Specifically, the meta-learning methods, while capable of rapidly adapting from the support set to the query set, lacks modeling of the semantic correlation between the support and query samples, failing to fully utilize the discriminative information shared across the two sets. Moreover, relation network method emphasizes modeling the relationship between support and query samples, but still extracts features from the two sets independently, unable to fully leverage their inherent semantic association, thereby limiting the robustness and accuracy of recognition. Consequently, these approaches might overlook critical inter-sample relationships, particularly in the context of infrared images where distinguishing between the target and background is more challenging due to lower contrast and resolution. It is noteworthy that infrared images are single-channel images lacking color information and have strong spatial correlation; thus, geometric structures and texture details become crucial features for infrared image recognition. However, existing few-shot learning methods for infrared images have not explicitly focused on these aspects.

In the task of few shot classification, test images in the query set come from novel classes, making it challenging for the extracted features to focus on the target objects[7]. For instance, in a test image containing multiple objects, the extracted features may only focus on objects from seen classes with a large number of labels in the training set, while ignoring target objects from unseen classes[8]. Kang et al. [9] proposed relation embeddings based on self-correlation representation and cross-correlation attention to model features within images and between images, effectively alleviating the aforementioned issue. However, their proposed self-correlation representation module primarily focuses on channel-wise correlations within images, overlooking spatial correlations. However, for infrared images, emphasizing spatial correlations of the image might be more valuable than channel autocorrelation because it can reflect local structures and texture information within the image.

We propose a few-shot infrared aircraft classification method based on cross-correlation networks, which integrates two crucial attention modules and is trained in an end-to-end manner. Firstly, by utilizing the parameter-free self-attention module(SAM), we extract the intra-correlation within each image to acquire feature representations in both spatial and channel dimensions. Subsequently, the cross-attention module(CA) is employed to generate cross-attention between support and query images, thereby enhancing the model's generalization capability. By efficiently fusing features within and between images with minimal parameters, the model reduces computational complexity. In contrast to current models for few-shot infrared aircraft classification, our approach enhances the focus on the infrared imagery's geometric and

textural details. It achieves this by establishing a semantic connection between the feature sets of the support and query samples, thereby improving the model's ability to accurately identify target objects. The proposed model receives robust support and validation from classification experiments and ablation studies, all achieved without the introduction of excessive parameters.

1 Method

In this section, we provide a detailed introduction to the Cross-Correlation Network (CCNet) proposed in this paper for few-shot infrared aircraft classification. The overall architecture of CCNet is illustrated in Fig. 1, comprising the simple parameter-free attention module and cross attention modules. For each pair of support classes and query samples, appropriate feature representations are obtained through the backbone network. Recent works[10-14] have utilized self-similarity as an intermediate feature transformation for deep neural networks, demonstrating its crucial role in learning effective representations of semantic correspondences in network learning. In this study, we introduce the SAM module, which learns the structural layout of images by computing the similarity of internal regions within infrared images. On the other hand, to fully exploit the semantic correlations between support and query features, we design the CA module to compute the cross-correlation between two image representations and learn to generate co-attention from it.

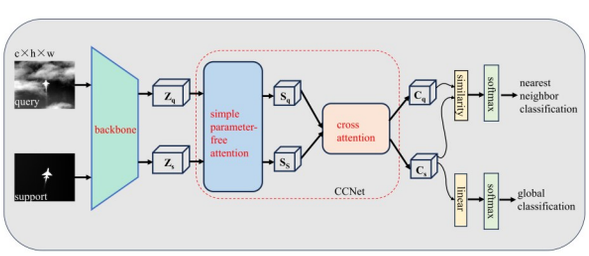


Fig. 1　The overall architecture of CCNet model
图1　CCNet模型整体架构

1.1 Parameter-free self-attention

Attention mechanisms allocate different weights to the importance of key information contained within channels, thereby enhancing the network's focus on important information. Common attention mechanisms are typically composed of convolutional layers, pooling layers, activation functions, etc., introducing additional parameters to the network. To improve network performance without increasing computational complexity, we introduce a simple, parameter-free attention mechanism module called SAM[15] into CCNet. The general structure and computation of SAM is shown in Fig. 2. SAM adopts an idea based on human visual processing, combining feature and spatial attention mechanisms, and designs a "plug-and-play" three-dimensional weight self-attention mechanism[16]. SAM calculates weights through an energy function, assigning unique weights to each neuron. Without adding any parameters to the original network, the three-dimensional attention weights can be used to quickly infer feature maps in each layer[17]. Specifically, information-rich neurons typically exhibit significant activation differences from surrounding neurons, and activated neurons may inhibit the activity of surrounding neurons, known as spatial suppression[18].

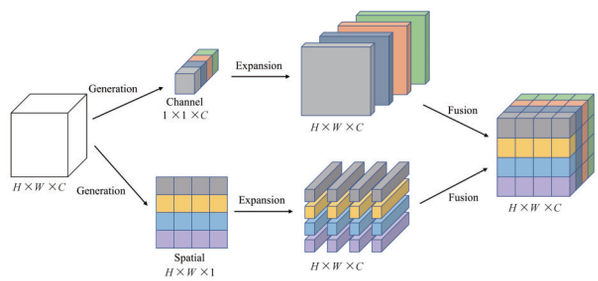


Fig. 2　Parameter-free self-attention model
图2　无参数自注意力模块

Building upon this, SAM defines an energy function to measure the difference between each feature and other features, thereby evaluating the importance of each feature. The definition of the energy function is as Equation (1):

$$e_t(\omega_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_o - \hat{x}_i)^2 \quad , \quad (1)$$

where $\hat{t} = \omega_t t + b_t$ and $\hat{x}_i = \omega_t x_i + b_t$, t and x_i represent the target neuron and other neurons within a single channel of input feature $X \in R^{C\times H\times W}$. Here, i denotes the index of spatial dimension, $M=H\times W$ represents the number of neurons in that channel, while ω_t and b_t stand for the weight and bias, respectively. As shown in Equation (2), this equation admits an optimal closed-form solution, enabling us to obtain the minimal energy expression:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\sigma^2 + 2\lambda} \quad , \quad (2)$$

where $\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$ and $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu})^2$ represent the mean and variance of all neurons in the channel, λ is a hyperparameter used for balancing. Equation (2) implies that as the energy decreases, the difference between neuron t and neighboring neurons becomes greater, making it more critical in image processing. Finally, the input features undergo enhancement processing to implement the attention mechanism through Equation(3):

$$\tilde{X} = \mathrm{sigmoid}\left(\frac{1}{E}\right) \odot X \quad , \quad (3)$$

where E represents the grouping of all e_t^* in both channel and spatial dimensions, and $\odot$ denotes the calculation of the Hadamard product. Adding a sigmoid function aims to constrain excessively large values in E, thereby ensuring the relative importance of each neuron. Therefore, employing the parameter-free self-attention as a three-dimensional weighting module allows for the assign-

ment of unique weights to each neuron, enhancing the attentional importance of each neuron in the feature maps. This three-dimensional weighting mechanism effectively exploits and highlights the structural features and background information within infrared images, facilitating efficient and reliable identification tasks for infrared targets.

1.2 Cross attention

In contrast to previous methods that independently extract features from support sets and query samples, we introduce a cross attention module to compute the cross-correlation between support and query images. The CA module enhances the model's focus on the target object by modeling the semantic relevance between class features and query features, thereby improving the efficiency and accuracy of the subsequent matching process. The cross attention module first takes the self-correlation representations of the support set and query samples (S_q and S_s) as inputs, then produces the corresponding cross-correlation representations (C_q and C_s). The architecture of the cross attention module, as shown in Fig. 3, primarily comprises three operations: cosine similarity computation, convolutional fusion and joint attention calculation. These operations will be introduced separately below.

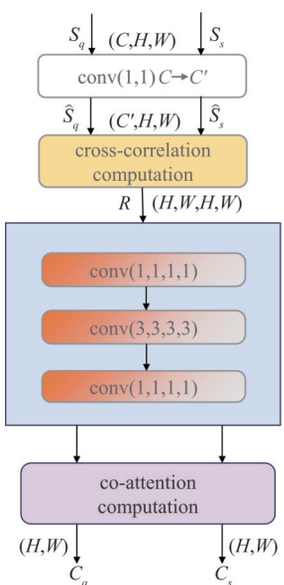


Fig. 3 The architecture of cross attention
图3 交叉注意力架构

In order to reduce computational complexity and obtain a more effective feature representation, we first employ a 1 × 1 convolutional layer to decrease the channel dimensions of S_q and $S_s \in R^{C\times H\times W}$ from C to C', resulting in the outputs S_q and $S_s \in R^{C'\times H\times W}$. Subsequently, the cross-correlation representation of S_q and S_s is computed using Equation (4):

$$\boldsymbol{R}\left(x_{q,}\ x_{\mathrm{s}}\right)=\left(\frac{\hat{\boldsymbol{S}}_{\mathrm{s}}\left(x_s\right)}{\left\|\hat{\boldsymbol{S}}_s\left(x_s\right)\right\|_2}\right)^T\left(\frac{\hat{S}_q\left(x_q\right)}{\left\|\hat{\boldsymbol{S}}_q\left(x_q\right)\right\|_2}\right) \quad , (4)$$

where x denotes a spatial location in the feature map, T signifies matrix transposition, and $R(x_q, x_s) \in R^{H\times W\times H\times W}$ represents the four-dimensional cross-correlation tensor. This computation method not only yields a reliable cross-correlation representation but also reduces the computational load.

In the process of fine-grained classification of infrared aircraft, due to the similarity of some target shapes, the cross-correlation tensor may contain unreliable correlations. Therefore, we adopt a convolution matching process to obtain a more reliable cross-correlation representation. Specifically, we use four-dimensional convolution, which enhances the expression ability of target features and improves the accuracy of classification by analyzing the consistency of adjacent matches in the four-dimensional space and achieving geometric matching on the tensor[9]. As shown in the blue box in the middle of Fig. 2, the convolution matching block consists of three 4D convolution layers. Firstly, the first convolution layer is responsible for increasing the number of channels to provide a richer feature representation in subsequent processing. Next, the second convolution layer generates multiple correlation tensors and aggregates them into a four-dimensional correlation tensor, achieving geometric matching of the tensor. Finally, the third convolution layer is responsible for restoring the number of channels. Between these three convolution layers, batch normalization layers and ReLU layers are inserted to enhance the stability of the network and improve the non-linearity of the activation function. This design allows the convolution matching block to effectively handle high-dimensional data, extract complex features, and provide more reliable input for subsequent generation of cross-attention.

After obtaining the reliable cross-correlation tensor, it is necessary to generate the common attention maps C_s and C_q for the support set and query set. Taking the calculation of the query attention map C_q as an example, the calculation method is shown in Equation (5):

$$C_q(x_q)=\frac{1}{HW}\sum_{x_s}\frac{\exp\left(\hat{R}\left(x_q,x_s\right)/\sigma\right)}{\sum_{x_q^{'}}\exp\left(\hat{R}\left(x_q^{'},x_s\right)/\sigma\right)} \quad , (5)$$

where x represents the position in the feature map, σ is the temperature factor, lower temperature will lead to lower entropy, making the distribution concentrate on a few positions with higher confidence[8], H and W represent the height and width of the feature map, and $R(x_q, x_s)$ is the matching score between positions x_q and x_s. Therefore, the attention map $C_q(x_q)$ in Equation (5) can be understood as converting the matching score of position x_q on the query image into the average probability of matching with position x_s on the support image. The calculation method of attention map C_s is similar to C_q. These co-attention maps improve the accuracy of infrared target classification by adjusting the attention position according to the images provided in the testing phase through the meta-learning cross-correlation pattern.

1.3 Loss function

Unlike many recent few-shot learning methods that adopt a 'pre-training + fine-tuning ' two-stage training

scheme, we propose an end-to-end training strategy for CCNet. This strategy jointly trains the designed modules and the backbone network by combining the metric loss L_{metric} and the global classification loss L_{label}. In this process, the calculation of L_{metric} is based on the cosine similarity between the query prototype feature vector and the support prototype feature vector. The calculation method of the metric loss is shown in Equation (6). This design of the metric loss helps guide the model to map the query embedding to the neighboring position of the prototype embedding of the same category:

$$L_{\text{metric}} = -\log \frac{\exp(\text{sim}(\bar{s}^{(n)}, \bar{q}^{(n)})/\tau)}{\sum_{N'=1}^{N} \exp(\text{sim}(\bar{s}^{(n')}, \bar{q}^{(n')})/\tau)} \quad , (6)$$

where $sim()$ denotes the calculation of cosine similarity, $\bar{s}^{(n)}$ and $\bar{q}^{(n)}$ represent the prototype vectors of the nth category, N indicates the total number of categories, and τ is the temperature factor.

The global classification loss L_{label} is computed using a fully connected layer followed by a softmax function, in order to classify each query sample among all available training categories. The specific calculation method is shown in Equation (7):

$$L_{\text{label}} = -\log \frac{\exp(w_c^T z_q + b_c)}{\sum_{c'=1}^{|C_{\text{train}}|} \exp(w_{c'}^T z_q + b_{c'})} \quad , (7)$$

where w_c^{T} represents the weights of the fully connected layer, b_c represents the corresponding bias, and $|c_{\text{train}}|$ represents the number of training categories. The overall classification loss is defined as:

$$L = L_{\text{label}} + \lambda L_{\text{metric}} \quad , (8)$$

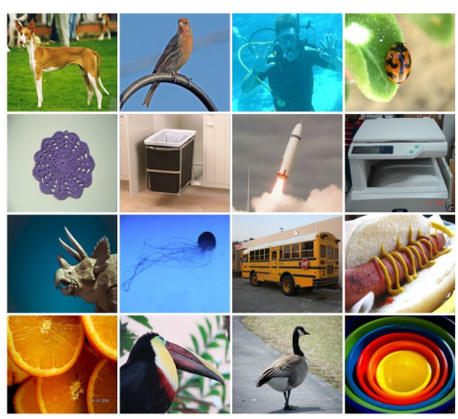
where λ is the weight that balances the effects of different losses. By optimizing the overall loss L, the network can be trained end-to-end using the gradient descent algorithm.

2 Experiments

2.1 Experimental environment and data source

All experiments were conducted in a hardware environment based on the Intel i7 13700 processor, NVIDIA GTX4080 graphics card, and DDR4 64G memory, as well as a software environment with the Win10 system and Pytorch deep learning framework. During the training phase, we adopted a training strategy based on N-way K-shot meta-tasks. Specifically, in each training cycle, N categories are randomly selected from the training data, and then K labeled samples are selected from each category to construct the support set. Subsequently, a certain number of samples are randomly selected from the other samples of these N categories, and these samples constitute the query set. Finally, the model predicts the category labels of the query samples. In the validation and testing phases, we still use the aforementioned meta-task form for evaluation. It should be noted that the data in the validation set, test set, and training set all come from different categories, which means that $C_{\text{train}} \cap C_{\text{val}} \cap C_{\text{test}} = \varnothing$.

In order to validate the effectiveness of the model proposed in this study, we conducted experiments using two datasets: the miniImageNet[19] dataset and the self-constructed infrared target dataset miniInfra[2,4,20]. The miniImageNet dataset is a subset of the ImageNet dataset widely used for few-shot learning and has become the benchmark dataset for few-shot learning. The miniInfra dataset is a self-constructed infrared target dataset sourced from publicly available datasets and self-captured infrared images. Partial examples from both datasets are illustrated in Fig. 4. To ensure consistency in our experiments, we resized all images across the datasets to 84x84. Moreover, we employed data augmentation strategies such as random cropping and random flipping for both datasets.



(a)

(b)

Fig. 4　(a) Samples of miniImageNet dataset; (b) samples of miniInfra dataset
图4　(a)miniImageNet数据集示例;(b)miniInfra数据集示例

In this study, we employ ResNet12[21] as the backbone network, which accepts images of spatial dimensions 84×84 as input and generates the basic representation Z. All training and testing are conducted in the form of tasks. For each N-way K-shot classification task, we test 15 query samples per category within each task. For the miniImageNet dataset, 2000 meta-tasks were randomly selected from the test set, and for the miniInfra datas-

et, 500 meta-tasks were randomly selected from the test set. Subsequently, we calculate the average classification accuracy and set the confidence intervals to 95%.

2.2 Few-shot classification based on the miniImageNet dataset

The miniImageNet dataset is composed of 100 categories, each containing 600 images, totaling 60,000 visible light images. Following the partitioning standards of previous literature[22], we designate 64, 16 and 20 categories as the training, validation and test sets, respectively. We conduct 5-way 1-shot and 5-way 5-shot classification tasks.

During the experiments on the miniImageNet dataset, we use the Stochastic Gradient Descent (SGD) optimizer for 80 epochs of training, each epoch consisting of 300 meta-tasks. The initial learning rate is set to 0.1, and a learning rate decay strategy is adopted. At the 60th and 70th epochs, the learning rate is multiplied by a decay factor of 0.05. In the experiments, the temperature factor τ of the metric loss function is set to 0.2, and the hyperparameter λ for balancing the loss weight is set to 0.25.

Table 1 presents the classification results of CCNet on the miniImageNet dataset. We have selected several mainstream methods in the field of few-shot learning in recent years for comparison, including meta-learning, relation networks, adversarial networks, and self-supervised learning. Our method belongs to metric learning methods which can achieve efficient learning for few-shot image classification without any pre-training process and post-processing operations. Although the backbone network of our model is smaller than that of some other methods[25-26, 30], the experimental results on the miniImageNet dataset show that the performance of our model still surpasses these methods. Unlike existing metric-based methods, which independently extract features from the support set and query set, leading to features scattered on non-target objects, CCNet can highlight the target object area and obtain more distinctive features. Specifically, CCNet first focuses on the target object through SAM, which is then combined with the cross-attention module CA, so that it can learn similar features between images and find the differences between fine-grained images more easily. In terms of inference speed, we evaluated 2000 5-way 5-shot tasks on an NVIDIA RTX 4080 GPU, which took approximately 1.5 minutes.

Table 1 Classification results on the miniImageNet dataset (average accuracy with 95% confidence interval)

表1 miniImageNet数据集上的分类结果(95%置信区间的平均准确率)

Method	Backbone	5-way 1-shot	5-way 5-shot
MAML[23]	ConvNet	48.70±0.84	63.11±0.92
RelationNet[19]	ConvNet	50.44±0.82	65.32±0.70
CAN[24]	ResNet12	63.85±0.48	79.44±0.34
AFHN[25]	ResNet18	62.38±0.72	78.16±0.56
PSST[26]	WRN-28-10	64.05±0.49	80.24±0.45
NCA[27]	ResNet12	62.55±0.12	78.27±0.09
Mata-baseline[28]	ResNet12	63.17±0.23	79.26±0.17
MIAN[29]	ResNet12	64.27±0.35	81.24±0.26
TFH[30]	ResNet18	64.49±0.84	79.94±0.60
CCNet(ours)	ResNet12	66.20±0.43	81.82±0.31

2.3 Aircraft classification based on the miniInfra dataset

The miniInfra dataset comprises 33 classes of terrestrial targets and 8 classes of aircraft targets. Terrestrial targets encompass various categories such as buildings, bicycles, pedestrians, cars, animals, and boats, with each class containing 100 to 200 infrared images. The 8 classes of aircraft targets include trainer aircraft, civil aviation aircraft, three types of helicopters (Z-8, Z-9, Z-15), and three types of jet aircraft (J-7, J-8, J-11), with each class containing 40 to 80 images. The granularity of aircraft target classification is finer than that of terrestrial targets.

Given the severe shortage of infrared aircraft data and to validate the model's ability to recognize fine-grained targets, we selected 25 types of ground targets as the training set, 8 types of ground targets as the validation set, and finally select 8 types of aircraft targets as the test set. The experiments include two standard few shot classification tasks: 5-way 1-shot and 5-way 5-shot. Considering that there are 8 types of aircraft, we added two specific classification tasks: 8-way 1-shot and 8-way 5-shot to test the model's generalization ability for few shot infrared aircraft in a real environment. Consistent with the experimental setup in the miniImageNet dataset, the experiment still uses the SGD optimizer and adopts a learning rate decay strategy. Since the size of the miniInfra dataset is much smaller than the miniImageNet dataset, to prevent overfitting, in the infrared aircraft classification task, we adjusted the number of training epochs to 20 and set the initial learning rate to 0.01.

We compared the experimental results with the existing infrared aircraft classification methods[2,4], and the specific results are shown in Table 2. It is worth noting that our training method is end-to-end and does not require any additional data for pre-training. From the results in the table, it can be seen that under the same conditions without using additional data for pre-training, the accuracy of our model in the four classification tasks is significantly better than the existing two classification methods. Especially in the 8-way 1-shot classification task, the model achieved a performance improvement of more than 3% compared to the previous best method, and also achieved a performance improvement of more than 2% in the 8-way 5-shot classification task. It can also be seen from Table 2 that even when compared with the results of the two methods using miniImageNet for pre-training, the accuracy of our method without pre-training is still higher than that of the methods proposed in Ref. [2] and [4].

2.4 Ablation experiments

To delve deeper into the impact of the core modules in CCNet, we conducted a series of ablation experiments

Table 2　Classification results on the miniInfra dataset (average accuracy with 95% confidence interval)
表 2　miniInfra 数据集上的分类结果(95%置信区间的平均准确率)

Method	Pre-train	5-way 1-shot	5-way 5-shot	8-way 1-shot	8-way 5-shot
Improved	No	84. 37±1. 31	93. 66±0. 76	77. 56±1. 46	90. 58±0. 64
Relation Network[4]	Yes	82. 79±0. 75	94. 51±0. 82	78. 47±0. 94	89. 82±1. 02
MLFC[2]	No	—	—	78. 58±0. 97	91. 12±0. 37
	Yes	—	—	81. 27±0. 91	92. 74±0. 35
CCNet(ours)	No	85. 58±0. 97	95. 09±0. 46	81. 95±0. 62	93. 26±0. 38

on the miniImageNet and miniInfra datasets. These experiments included scenarios where two core modules were missing simultaneously, as well as cases where only one of the modules was used independently. We constructed a baseline model that only contains the backbone network and does not include any additional modules, to evaluate the effectiveness of the core modules in CCNet. We carried out 5-way 1-shot ablation experiments on the miniImageNet and miniInfra datasets. As can be seen from Fig. 5, although the training accuracy of CCNet (blue line) is slightly lower than that of the baseline model (red line), the validation accuracy of CCNet (green line) has significantly improved compared to the validation accuracy of the baseline model (purple line). This result indicates that compared to the baseline model, CCNet has stronger generalization ability when applying the trained model to unseen new categories.

In this study, further ablation experiments are conducted on the 5-way 1-shot tasks of two datasets to individually validate the effectiveness of the SAM module and the CA module. When only the CA module is used, the basic representation Z_q is taken as input; when only the SAM module is used, its output is directly utilized for classification. The results of the ablation experiments are presented in Fig. 6. The experimental results demonstrate that both the SAM and CA modules can significantly improve the accuracy of classification compared to the baseline model. The SAM module can generate reliable representation and provide robust support for the classification tasks, while the CA module can further enhance the representations generated by SAM and improve the cross-correlation between images, thus further improving the classification accuracy.

We also present the results of class activation mapping (CAM) feature visualization using our CCNet, encompassing both visible and infrared images, as illustrated in Fig. 7. In the CAM visualizations, the regions with warmer colors (e. g. , red, yellow) represent areas in the input image that contribute more significantly to the network's classification decision. In other words, cooler colors (e. g. , blue) indicate areas that the model does not focus on or that have lower importance. Figure 7(a) depicts an image containing both a cat and a dog; however, since our objective is to identify the cat, the CAM visualization in Fig. 7(b) clearly highlights the network's focus on the cat's region. In the infrared image shown in Fig. 7(c), despite the presence of a complex background with pedestrians and a dog, the network accurately concentrates on the dog, which is the target for classification. Furthermore, for the infrared image of a commercial aircraft against a cloudy background in Fig. 7(e), the CAM result in Fig. 7(f) demonstrates that the network effectively attends to and emphasizes the aircraft target region.

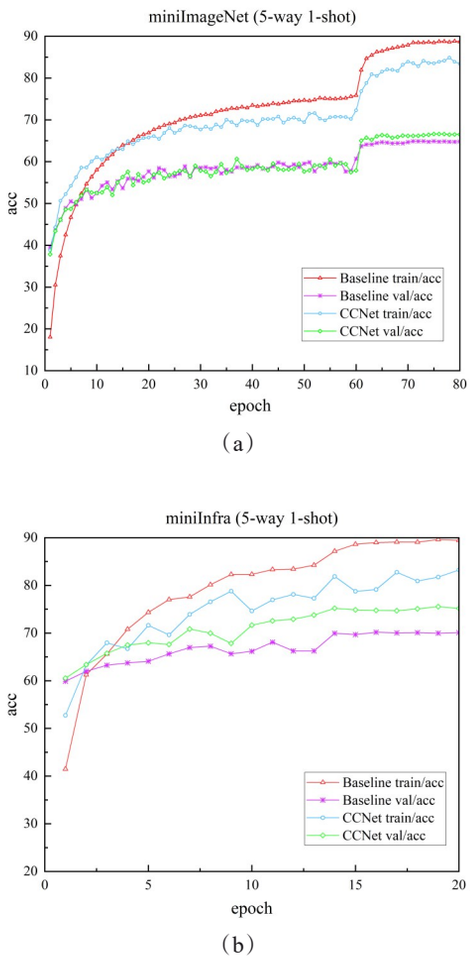


Fig. 5　(a) Training and validation accuracy curves of the baseline model and CCNet model on miniImageNet dataset; (b) training and validation accuracy curves of the baseline model and CCNet model on miniInfra dataset
图 5　(a)基线模型和 CCNet 模型在 miniImageNet 数据集训练和验证准确率曲线;(b)基线模型和 CCNet 模型在 miniInfra 数据集训练和验证准确率曲线

2. 5　Performance and parameter comparison of different attention modules

In this study, we replaced different attention mod-

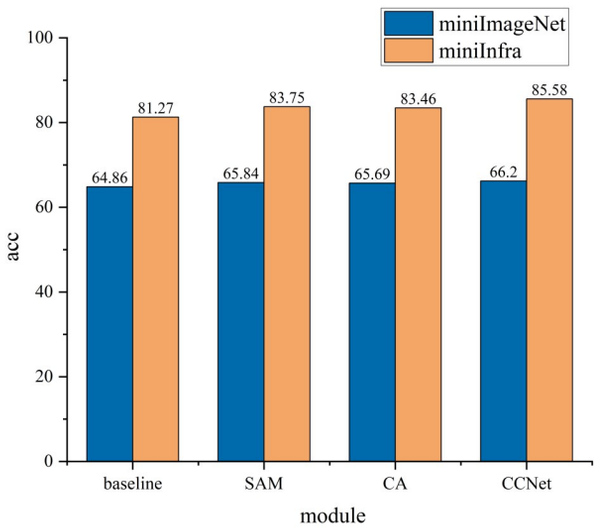


Fig. 6 Ablation experiment results on miniImageNet and miniInfra dataset
图6 在miniImageNet和miniInfra数据集上的消融实验结果

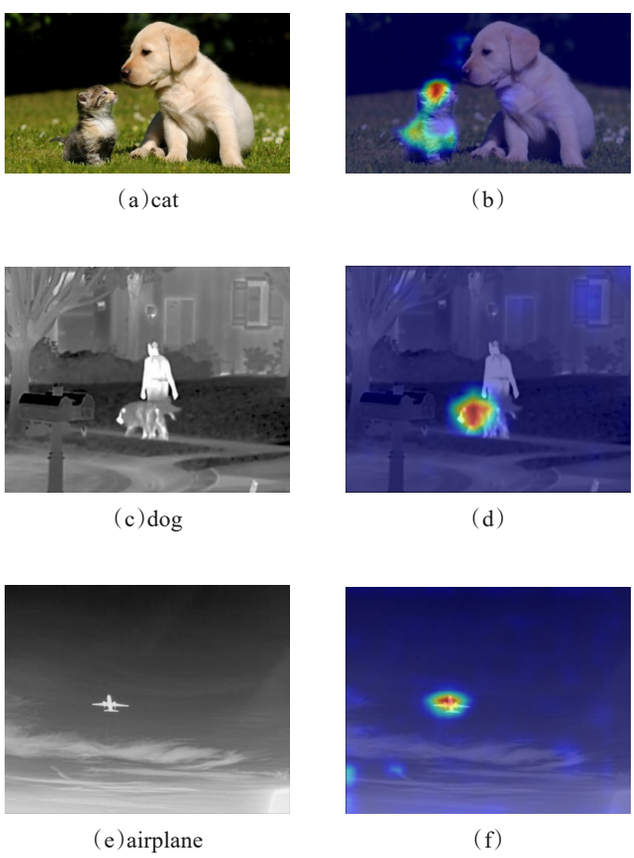


Fig. 7 The class activation mapping (CAM) feature visualization of CCNet
图7 CCNet的类激活映射(CAM)特征可视化

ules in the proposed CCNet network model to compare the accuracy and parameter scale of the proposed modules with existing attention modules. Firstly, we evaluated the self-attention and cross-attention methods based on feature similarity, which focus on the correlation of the image spatial structure features.

Table. 3 Comparison of accuracy and parameter quantities of different attention modules
表3 不同注意力模块的精度和参数量比较

Module	Self	Cross	miniImageNet	miniInfra	Add params
Baseline	√	×	64.86	81.27	0 k
SE[31]	√	×	66.37	81.99	102.4 k
SCE[32]	√	×	62.96	79.80	89.2 k
LSA[33]	√	×	64.77	80.62	1 644.16 k
NLSA[34]	√	×	65.67	82.34	822.1 k
CBAM[35]	√	×	64.77	82.79	102.5 k
SCR[9]	√	×	64.43	78.80	157.3 k
CCA[9]	×	√	66.00	84.26	45.8 k
SAM	√	×	65.84	83.31	0 k
CA	×	√	65.69	84.30	9.41 k

As shown in Table 3, the results of the SAM module and the CA module on the two datasets are superior to other attention methods. Specifically, on the miniImageNet dataset, the performance of SAM in the self-attention module is second only to the SE attention module, while on the miniInfra dataset, SAM achieved the best result among the self-attention methods. It is worth noting that compared with the baseline model, SAM does not introduce additional parameters. Among the cross-attention methods, the performance of the CA module and the CCA module is similar, but the CA module introduces fewer parameters.

3 Conclusion

In this study, we have proposed a few-shot infrared aircraft classification method based on the cross-correlation network, which can effectively solve the classification problem of infrared aircraft when the number of samples is severely insufficient. In the research process, in order to reduce model parameters and specifically target the structural features of infrared aircraft target images, we introduce a parameter-free self-attention mechanism to analyze the self-correlation within images. Meanwhile, we design a cross-attention mechanism to investigate the self-correlation between images, which effectively enhances the model's capability to extract features from infrared images. The experimental results show that our method significantly outperforms existing methods in aerial target classification accuracy on the infrared dataset, with an improvement of up to 3% in classification accuracy for specific tasks. Furthermore, the tests on the public miniImageNet dataset and the ablation experiments further verify the effectiveness and contributions of the proposed modules. The method proposed in this paper not only has broad application potential in aircraft detection, but also has great application value in civilian fields where data is scarce, such as medical. But at the same time in the research tasks of this paper it only involves the single task of aircraft classification. However,

in actual application scenarios of the infrared detection system it involves a series of complex tasks such as target detection target recognition and target tracking. Therefore, how to deploy the few-shot model to these actual application scenarios and maintain good performance under multiple tasks will be the focus of the next stage of work.

References

[1] Ning C, Liu W, Wang X. Infrared Object Recognition Based on Monogenic Features and Multiple Kernel Learning[C]//2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC). 2018: 204-208.

[2] Chen R M, Liu S J, Miao Z, *et al*. Infrared aircraft few-shot classification method based on meta learning[J]. *Journal of Infrared and Millimeter Waves*, 2021, **40**(4): 554-560.

[3] Li W, Chen Q, Gu G, *et al*. Visible-infrared image matching based on parameter-free attention mechanism and target-aware graph attention mechanism[J]. *Expert Systems with Applications*, 2024, **238**: 122038.

[4] Jin L, Liu S J, Wang X, *et al*. Infrared aircraft classification method with small samples based on improved relation network[J]. *Acta Optica Sinica*, 2020, **40**(8): 0811005.

[5] Luo X, Wu H, Zhang J, *et al*. A closer look at few-shot classification again[C]//Proceedings of the 40th International Conference on Machine Learning. 2023, **202**: 23103-23123.

[6] Li X, Yang X, Ma Z, *et al*. Deep metric learning for few-shot image classification: A Review of recent developments[J]. *Pattern Recognition*, 2023, **138**: 109381.

[7] Shi B, Li W, Huo J, *et al*. Global- and local-aware feature augmentation with semantic orthogonality for few-shot image classification[J]. *Pattern Recognition*, 2023, **142**: 109702.

[8] Hou R, Chang H, Ma B, *et al*. Cross Attention Network for Few-shot Classification[C]//Advances in Neural Information Processing Systems. 2019, 32.

[9] Kang D, Kwon H, Min J, *et al*. Relational Embedding for Few-Shot Classification[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 8802-8813.

[10] Li J. EACCNet: Enhanced Auto-Cross Correlation Network for Few-Shot Classification[C]//Knowledge Science, Engineering and Management. 2023, **14117**: 354-365.

[11] Kwon H, Kim M, Kwak S, *et al*. Learning Self-Similarity in Space and Time as Generalized Motion for Video Action Recognition[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 13045-13055.

[12] Lee S, Lee S, Seong H, *et al*. Revisiting Self-Similarity: Structural Embedding for Image Retrieval[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 23412-23421.

[13] Wang L, Lei S, He J, *et al*. Self-Correlation and Cross-Correlation Learning for Few-Shot Remote Sensing Image Semantic Segmentation [C]//Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems. 2023: 1-10.

[14] Zhong Y, Su Y, Zhao H. Self-similarity feature based few-shot learning via hierarchical relation network[J]. *International Journal of Machine Learning and Cybernetics*, 2023, **14**(12): 4237-4249.

[15] Yang L, Zhang R-Y, Li L, *et al*. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks[C]//Proceedings of the 38th International Conference on Machine Learning. 2021: 11863-11874.

[16] Wen X, Cao C, Li Y, *et al*. DRSN with Simple Parameter-Free Attention Module for Specific Emitter Identification[C]//2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 2022: 192-200.

[17] Tan S, Zhang L, Shu X, *et al*. A feature-wise attention module based on the difference with surrounding features for convolutional neural networks[J]. *Frontiers of Computer Science*, 2023, **17**(6): 176338.

[18] Webb BS, Dhruv NT, Solomon SG, *et al*. Early and Late Mechanisms of Surround Suppression in Striate Cortex of Macaque[J]. *Journal of Neuroscience*, 2005, **25**(50): 11666-11675.

[19] Vinyals O, Blundell C, Lillicrap T, *et al*. Matching Networks for One Shot Learning[C]//Advances in Neural Information Processing Systems. 2016, 29.

[20] Liu Q, Li X, Yuan D, *et al*. LSOTB-TIR: A Large-Scale High-Diversity Thermal Infrared Single Object Tracking Benchmark [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2023: 1-14.

[21] He K, Zhang X, Ren S, *et al*. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778.

[22] Ravi S, Larochelle H. Optimization as a model for few-shot learning [C]//International Conference on Learning Representations. 2017.

[23] Finn C, Abbeel P, Levine S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks[C]//Proceedings of the 34th International Conference on Machine Learning. 2017: 1126-1135.

[24] Hou R, Chang H, MA B, *et al*. Cross Attention Network for Few-shot Classification[C]//Advances in Neural Information Processing Systems. 2019, 32.

[25] Li K, Zhang Y, Li K, *et al*. Adversarial Feature Hallucination Networks for Few-Shot Learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 13467-13476.

[26] Chen Z, Ge J, Zhan H, *et al*. Pareto Self-Supervised Training for Few-Shot Learning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 13658-13667.

[27] Laenen S, Bertinetto L. On Episodes, Prototypical Networks, and Few-Shot Learning[C]//Advances in Neural Information Processing Systems. 2021, **34**: 24581-24592.

[28] Chen Y, Liu Z, Xu H, *et al*. Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning[C]//**2021**: 9062-9071.

[29] Qin Z, Wang H, Mawuli CB, *et al*. Multi-instance attention network for few-shot learning [J]. *Information Sciences*, 2022, **611**: 464-475.

[30] Lazarou M, Stathaki T, Avrithis Y. Tensor feature hallucination for few-shot learning[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2022: 2050-2060.

[31] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[32] Huang S, Wang Q, Zhang S, *et al*. Dynamic Context Correspondence Network for Semantic Alignment[C]//**2019**: 2010-2019.

[33] Ramachandran P, Parmar N, Vaswani A, *et al*. Stand-Alone Self-Attention in Vision Models[C]//Advances in Neural Information Processing Systems. 2019, 32.

[34] Wang X, Girshick R, Gupta A, *et al*. Non-local Neural Networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.

[35] Woo S, Park J, Lee J-Y, *et al*. CBAM: Convolutional Block Attention Module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.