

Article

A Counterfactual Fine-Grained Aircraft Classification Network for Remote Sensing Images Based on Normalized Coordinate Attention

Zeya Zhao¹, Wenyin Tuo ², Shuai Zhang¹ and Xinbo Zhao^{2,*}¹ Beijing Institute of Tracking and Communication Technology, Beijing 100094, China; zzy_kelly@163.com (Z.Z.); zhangshuaioec@163.com (S.Z.)² School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China; nwpu_twy@mail.nwpu.edu.cn

* Correspondence: xbozhao@nwpu.edu.cn

Abstract

Fine-grained aircraft classification in remote sensing is a critical task within the field of remote sensing image processing, aiming to precisely distinguish between different types of aircraft in aerial images. Due to the high visual similarity among aircraft targets in remote sensing images, accurately capturing subtle and discriminative features becomes a key technical challenge for fine-grained aircraft classification. In this context, we propose a Normalized Coordinate Attention-Based Counterfactual Classification Network (NCC-Net), which emphasizes the spatial positional information of aircraft targets and effectively captures long-range dependencies, thereby enabling precise localization of various aircraft components. Furthermore, we analyze the proposed network from a causal perspective, encouraging the model to focus on key discriminative features of the aircraft while minimizing distraction from the surrounding environment and background. Experimental results on three benchmark datasets demonstrate the superiority of our method. Specifically, NCC-Net achieves Top-1 classification accuracies of 97.7% on FAIR1M, 95.2% on MTARSI2, and 98.4% on ARSI120, outperforming several state-of-the-art methods. These results highlight the effectiveness and generalizability of our proposed method for fine-grained remote sensing target recognition.

Keywords: fine-grained aircraft classification; remote sensing image; causal inference; explainability analysis



Academic Editor: Yutaka Ishibashi

Received: 19 June 2025

Revised: 24 July 2025

Accepted: 29 July 2025

Published: 12 August 2025

Citation: Zhao, Z.; Tuo, W.; Zhang, S.; Zhao, X. A Counterfactual

Fine-Grained Aircraft Classification Network for Remote Sensing Images Based on Normalized Coordinate Attention. *Appl. Sci.* **2025**, *15*, 8903. <https://doi.org/10.3390/app15168903>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the continuous improvement of spatial resolution in remote sensing images, countries worldwide have been leveraging advanced remote sensing technologies and equipment to acquire large volumes of high-quality remote sensing data, laying a solid foundation for the interpretation and analysis of such imagery. Within this field, fine-grained aircraft classification in remote sensing images has attracted significant attention [1–3]. Through in-depth image analysis, it can be observed that aircraft exhibit distinctive and stable identifiable features in imagery, such as paired wings and symmetrical structures, which provide unique cues for fine-grained classification. However, the subtle details of aircraft targets in remote sensing images are often difficult to detect and are usually highly dependent on spatial positional information. Therefore, accurately localizing aircraft targets is crucial to ensuring classification accuracy. Early fine-grained image classification

algorithms based on handcrafted features primarily focused on local image details. Yet, the process of selecting local features was cumbersome, the representational capacity was limited, and such methods often overlooked the correlations between different local features, as well as the spatial relationships between local and global features. As a result, they struggled to capture subtle discriminative details and failed to achieve satisfactory outcomes. With the rise of deep learning, features automatically extracted from neural networks have demonstrated much stronger representational power compared to handcrafted features, which has significantly advanced the development of fine-grained image classification algorithms. Early deep learning-based approaches typically employed auxiliary annotations, such as object parts [4] and keypoints [5], as supervisory signals to train models. While these methods accurately conveyed critical positional information of objects, many approaches [6,7] relied on predefined bounding boxes set during the preprocessing stage. Such reliance on prior knowledge limited their flexibility and adaptability. In light of these limitations, our research focuses on fine-grained classification under image-level weak supervision, aiming to reduce dependency on strong annotations while maintaining high classification precision.

Currently, a large body of research focuses on weakly supervised models trained solely with category labels. Lin et al. [8] proposed a bilinear convolutional neural network (B-CNN) comprising two VGG networks, where bilinear pooling offers superior feature fusion compared to standard linear models; however, its dimensionality is excessively high, limiting its generalizability. Sun et al. [9] introduced the “squeeze-and-multiple-excitation” module, which concatenates outputs from multiple activation modules as the final image feature vector. In practice, they employed two excitation branches, allowing the model to identify two discriminative regions from the same image. Peng et al. [10] proposed a two-level attention strategy, learning multi-view and multi-scale features at both the object and local levels to enhance feature representations. Rodriguez et al. [11] adopted a multi-scale attention fusion approach, extracting hierarchical feature information at different network depths and fusing attention maps at the output layer. Zheng et al. [12] proposed a trilinear attention sampling network comprising an attention module for locating fine-grained discriminative information, an attention sampler, and a feature distiller, thereby extracting more effective details and strengthening the network’s capacity to capture fine-grained features. The superiority of these models lies in their ability to identify highly discriminative local regions for classification tasks. However, when applied to fine-grained aircraft classification in remote sensing images, conventional models often struggle to precisely localize critical aircraft components and capture long-range dependencies. To address these two key challenges, our research incorporates position encoding from a spatial perspective, integrating a coordinate normalization mechanism to enhance the weights of local discriminative features, thereby accurately localizing distinguishable regions within the image.

At present, most deep networks are regarded as black-box models, providing only classification predictions without offering insights into the factors guiding those predictions. Given the possibility of spurious correlations between features and outcomes, this study introduces a counterfactual causal attention approach to enhance the predictive reasoning of the model. In the proposed causal multi-head attention mechanism, multiple attention maps are extracted from high-level feature representations, and convolutional layers are employed to highlight the discriminative parts of the object, thereby directing the network’s focus toward salient features. Furthermore, we incorporate a counterfactual intervention method based on causal inference to explore the intrinsic relationships between the generated attention maps and the predicted classes.

Unlike conventional attention mechanisms, counterfactual attention offers not only interpretability but also significant advantages in terms of model robustness. By actively intervening in the prediction pathway, counterfactual reasoning identifies feature regions that are causally related to the classification outcome, thereby reducing reliance on spurious correlations and background co-occurrence information. This approach greatly enhances the model's resilience to noisy labels and domain shifts. In remote sensing scenarios, where data distributions often vary due to differences in geographic regions, imaging devices, or environmental conditions, and where label noise is common, counterfactual reasoning provides a principled mechanism to simulate “what-if” interventions, i.e., whether removing a specific region changes the prediction outcome. This guides the model to focus on truly discriminative features, resulting in improved generalization and robustness in complex or real-world environments.

In addition, fine-grained aircraft classification faces challenges such as high feature dimensionality, complex processing pipelines, and diverse feature distributions. It has thus become imperative to deeply investigate the inherent connections between the object's attended regions and the final predictions. Some existing fine-grained classification methods [13–15] have employed Grad-CAM [16] and t-SNE [17] algorithms for interpretability analysis, using visualizations to explain model behavior. To clarify the underlying reasons behind network decisions at the visual level, we utilize Grad-CAM and t-SNE to visualize and analyze feature maps at critical stages, thereby concentrating on discriminative regions. The main contributions of this paper can be summarized as follows:

1. A novel attention mechanism is proposed: this design performs spatial position encoding and integrates a coordinate normalization mechanism to enhance the weights of key local features in the image, suppress interference from non-essential subject features, and accurately locate discriminative regions within the image.
2. A counterfactual intervention method based on causal inference is introduced to investigate the intrinsic relationships between the generated attention maps and the predicted classes, which not only enhanced the model's interpretability but also significantly improved its robustness in complex scenarios such as annotation errors and domain changes.
3. The proposed network demonstrates strong classification performance and interpretability on three challenging fine-grained aircraft classification datasets, showing great potential for applications in both civil and military domains.

2. Related Works

2.1. Fine-Grained Image Classification

In recent years, weakly supervised fine-grained image classification methods have become the mainstream direction for addressing fine-grained image recognition problems. Lin et al. [8] first proposed the bilinear convolutional neural network (B-CNN) as a feature extractor, and many recent works have improved upon it. In addition, some studies have replaced CNNs with Transformer networks [18,19], for example, Dosovitskiy et al. [18] proposed the Vision Transformer (ViT) for image classification, and He et al. [19] further designed a Transformer network specifically for fine-grained image classification, named TransFG. Furthermore, Wang et al. [20] proposed the Feature Pyramid Transformer (FPT), which integrates cross-scale feature fusion and achieved 94.1% accuracy on the Stanford Dogs dataset. Currently, part-based methods are the most popular approach in fine-grained classification. These methods aim to identify the discriminative parts for classification. Among them, attention mechanisms play an irreplaceable role in fine-grained visual recognition tasks. For example, in fine-grained image classification, Sermanet et al. [21] applied attention mechanisms by proposing an RNN model that learns visual attention.

Zheng et al. [22] proposed a channel grouping network to generate attention for multiple parts. Chen et al. [23] introduced a Semantic Contrastive Learning (SCL) framework, which enables effective feature alignment without requiring part annotations. Existing studies mainly rely on designing complex attention mechanisms to locate discriminative regions but often ignore the relationships between these regions. Starting from the goal of capturing fine aircraft features and the long-range dependencies within aircraft structures, this paper proposes a more effective network model.

2.2. Attention Mechanism

In the field of image classification, attention mechanisms are special structures embedded within models that mimic the human visual system's bias toward dominant features. SENet [24,25] adaptively redistributes the weights across channel dimensions through squeeze-and-excitation operations, enabling the network to focus on more efficient channels. CBAM [26] further advanced this idea by introducing spatial information encoding through large-kernel convolutions. Li et al. [27] proposed the Mobile-Former architecture, which integrates lightweight convolution with Transformer structures, while Wu et al. [28] designed a dynamic coordinate learning mechanism that optimizes positional encoding through deformable convolution. Current mainstream self-attention algorithms [29,30] focus on constructing spatial or channel attention. Yet, self-attention modules tend to be computationally intensive and inefficient. The CA method [31] proposed a more efficient approach to capturing positional and channel relationships, but it still suffers from some feature redundancy issues. Inspired by the CA approach for obtaining object positional information, the method proposed in this paper introduces a more lightweight attention mechanism that effectively reduces redundant features while enhancing model performance.

2.3. Explainable Analysis

In building explainable models, causal inference is widely regarded as a powerful tool for establishing interpretable relationships between features and decisions. Early on, causal inference found broad applications in fields such as medicine and economics. In recent years, some studies have introduced causal inference into computer vision, aiming to mimic human reasoning. For example, Tang et al. [32] used counterfactual reasoning to address data bias in scene graph generation and long-tail classification. Rao et al. [33] applied counterfactual training to mitigate spatial attention bias in fine-grained image recognition, while Niu et al. [34] reduced language bias in visual question answering tasks by subtracting direct language effects from the total causal effect. In the remote sensing domain, researchers have successfully leveraged causal inference tools for tasks like data reconstruction and relationship extraction, offering deeper insights into remote sensing and Earth sciences. For instance, Xiong et al. [35] focused on specific parts of ship targets by combining counterfactual causal attention with convolutional filters and visualizing decision bases. Liu et al. [36] proposed a causal disentanglement learning framework for fine-grained remote sensing classification, which improves robustness by 18.3% in cross-domain testing. Xu et al. [37] developed a causal domain adaptation mechanism that significantly reduces sensitivity to background interference. Based on these previous studies, causal inference can help disentangle key factors in the data, allowing models to develop a deeper understanding. Building on this technique, we designed and developed a system to explore the impact of counterfactual causal reasoning based on the normalized coordinate attention of the model.

3. Materials and Methods

This section focuses on explaining the basic structure of the NCC-Net Counterfactual Coordinate Attention Network. This network precisely locates airplane positions to obtain fine-grained features and introduces a counterfactual reasoning process to enhance decision-making, enabling the network to focus on airplane targets. The proposed NCC-Net is mainly divided into three parts:

1. **Feature Extraction Module:** We employ ResNet and MobileNet as the backbone networks for feature extraction. The core idea of ResNet is to build deep networks using residual blocks, which effectively mitigates the vanishing gradient problem. MobileNet, on the other hand, is a lightweight convolutional neural network characterized by its efficient and compact design. Both networks can extract high-level feature representations from input images, capturing advanced semantic information.
2. **Normalized Coordinate Attention Module:** This module emphasizes positional information by incorporating it into feature judgment, allowing the model to recognize the relative positions of different regions within the input data. This helps the model better understand the structure of the image or sequence and models the global context through global similarity, thus accelerating convergence and enabling more accurate classification decisions.
3. **Causal Discrepancy Learning Module:** By comparing the regions the model focuses on with the actual classification results, the module learns which areas are critical for correct classification, helping the model identify truly useful features and reduce reliance on noisy ones. The causal discrepancy learning module improves both the performance and interpretability of the model, allowing it to better understand the reasons behind its decisions.

3.1. Network Architecture

The main structure of NCC-Net is shown in Figure 1, and the overall process can be divided into three main steps. First, a deep convolutional network is used to extract high-level features of the aircraft target. Taking the ResNet backbone classification network as an example, the input images are first resized to a uniform size before being fed into ResNet. Next, the proposed normalized coordinate attention mechanism module is introduced into the backbone, embedding positional information into the channel attention. Through normalization, irrelevant weight information is suppressed, while ensuring consistent input and output dimensions to prevent the loss of critical pixel features during information transmission. After a series of convolutional and pooling operations, high-level features fused with global information are obtained from the backbone integrated with the attention mechanism $f^{l-1} \in R_{W \times H \times C}$, where W , H , and C represent the width, height, and number of channels of the feature map, respectively. Finally, after obtaining the high-level features, M causal attention maps A are generated, $\{A = A_1, A_2, \dots, A_M\}$, where M denotes the number of attention maps and $A_m \in A$. Through element-wise multiplication, the feature maps and attention maps A are used to generate regional feature maps and produce regional feature representations. The entire process is guided by a total loss function that includes both attention loss and classification loss to supervise the training.

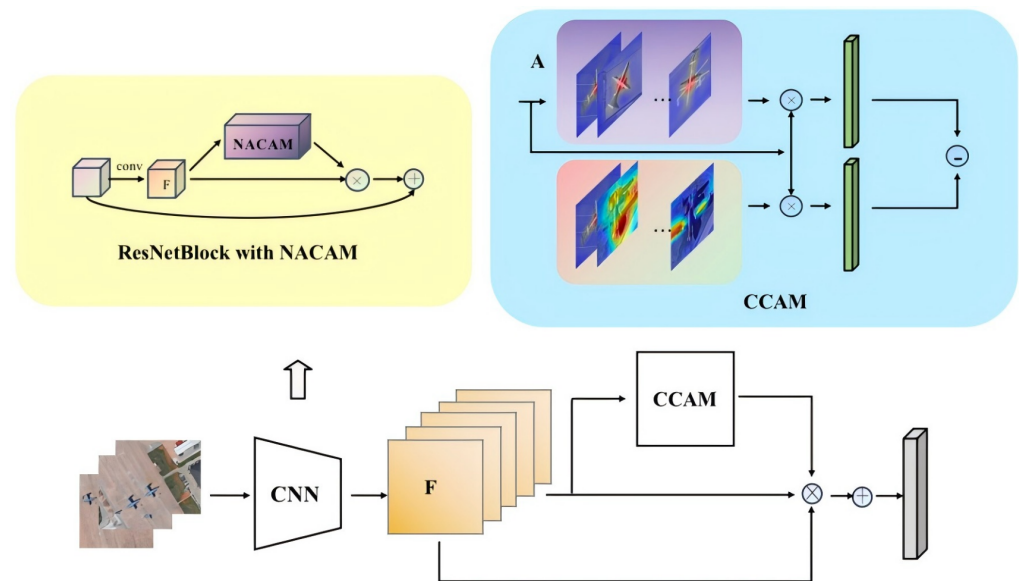


Figure 1. Network architecture.

3.2. Normalized Coordinate Attention Module

In fine-grained classification of remote sensing images, the diversity of aircraft positions, postures, and scales presents significant challenges, as these subtle variations are often difficult for traditional classification models to accurately capture. To address this issue, we emphasize the enhancement of positional information by integrating coordinate features. By introducing the coordinate information of the aircraft target, the model can more effectively focus on key features that are closely related to the aircraft's spatial location within the image. This strategy of reinforcing positional information significantly improves the accuracy of fine-grained classification tasks, enabling the model to better adapt to subtle variations in aircraft positioning and thereby achieve superior classification performance when faced with diverse aircraft variations.

Specifically, this mechanism simultaneously encodes both the channel relationships in the convolutional neural network and the long-range dependencies containing precise coordinate information, generating coordinate attention weights, which include two distinct attention maps along the horizontal and vertical directions. We decompose global pooling into a pair of one-dimensional feature encoding operations. For the input feature map f of size $H \times W \times C$, in order for the attention mechanism to preserve features containing precise coordinate information, we use two spatial fusion kernels for each channel encoding, namely feature encoders with kernel sizes of $(H, 1)$ and $(1, W)$, which independently encode the channel information along the horizontal and vertical directions:

$$z_c^h = \frac{1}{W} \sum_{0 \leq i \leq W} x_c(h, i) \quad (1)$$

$$z_c^w = \frac{1}{H} \sum_{0 \leq j \leq H} x_c(j, w) \quad (2)$$

In the equation, z_h^c represents the attention vector at height h in the c -th channel, and z_w^c represents the attention vector at width w in the c -th channel. $x_c \in \mathbb{R}^{H \times W}$, $z_h^c \in \mathbb{R}^{H \times 1}$, $z_w^c \in \mathbb{R}^{1 \times W}$. The transformations in (1) and (2) aggregate features along the two respective directions, producing a pair of direction-aware feature maps, which help the network more accurately locate the objects of interest. After obtaining the feature representations containing precise coordinate information, the two feature vectors are concatenated and

passed through a 1×1 convolution function F_1 and a nonlinear activation function (σ) to encode the spatial information along the horizontal and vertical directions, resulting in the final feature map:

$$f = \sigma(F_1[z_h, z_w]) \quad (3)$$

The feature map f is decomposed along two spatial dimensions into $f^h \in \mathbb{R}^{C/r \times H}$ and $f^w \in \mathbb{R}^{C/r \times W}$, where r (reduction ratio) serves to reduce both the dimensionality of feature parameters and the total parameter count. Then, f^h and f^w are fed into two 1D convolutions F_h and F_w , along with the sigmoid activation function (σ), obtaining the attention weight matrices g^h and g^w . The 1D convolutions transform f^h and f^w to match the channel dimension of input x , followed by attention weight allocation.

$$g^h = \sigma(F_h(f^h)) \quad (4)$$

$$g^w = \sigma(F_w(f^w)) \quad (5)$$

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (6)$$

where $x_c(i, j)$ and $y_c(i, j)$ are the values at coordinate (i, j) and channel c in the input and output feature maps, respectively; $g_c^h(i)$ and $g_c^w(j)$ are the values of g^h and g^w at channel c with coordinates $(i, 1)$ and $(1, j)$, respectively.

The coordinate attention module not only strengthens the capture of positional information and enhances the classification model's adaptability to diverse airplane samples, but also effectively reduces the influence of the background on the classification process, increasing the model's focus on airplane objects. By guiding the model to pay more attention to the airplane itself, coordinate attention helps highlight the salient features of airplanes, thereby improving the model's discriminative ability. It can partially suppress background interference in remote sensing images, allowing the model to more precisely process airplane-related features and thus improve robustness and accuracy in the classification process.

While coordinate attention indeed offers advantages in emphasizing positional information and feature weighting, it also has some potential drawbacks, such as limited robustness to multi-scale problems and possible bias in feature weight calculations. The main purpose of introducing a normalization mechanism is to address and optimize these weaknesses. This mechanism can reduce model instability caused by biased feature weight calculations. In some cases, coordinate attention may overemphasize specific positional information, leading to overfitting when the model encounters data variations. Normalization ensures a more balanced weight distribution, reducing unnecessary biases and improving the model's generalization ability and stability.

The normalization module NAM [38] redesigns the spatial attention module and the channel attention module. A scaling factor γ is introduced into the batch normalization process, as shown in Equation (7).

$$B_{out} = BN(B_{in}) = \gamma \frac{B_{in} - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} + \beta \quad (7)$$

In the formula, μ_B and σ_B are the mean and standard deviation of batch B , respectively, γ and β are learnable parameters, where γ is the scaling factor for each channel after compression. The introduced scale factor γ_i measures the significance of weights by evaluating the variance of each channel in the input features. A larger value indicates more informative channels, allowing the network to focus less on unimportant regions. $W_i = \frac{\gamma_i}{\sum_{j=0} \gamma_j}$ represents the obtained weights. To emphasize important features, NAM

employs batch normalization scale factors to indicate weight importance and utilizes weight contribution factors to enhance attention mechanisms, thereby obtaining more representative feature representations.

Therefore, the introduction of the normalization mechanism aims to overcome the shortcomings of coordinate attention, improving the model's robustness, generalization ability, and stability. By normalizing feature weights, the model can better handle issues at different scales and avoid problems caused by bias, thereby enabling coordinate attention to perform more effectively in tasks such as remote sensing aircraft recognition. The detailed structure of the NACAM module is shown in Figure 2. The implementation steps are as follows: First, global average pooling is applied in the horizontal and vertical directions, downsampling the input $C \times H \times W$ feature map to $H \times 1 \times C$ and $1 \times W \times C$ dimensions, respectively, to capture the spatial positional information of the input feature map. Second, information from different dimensions is fused and fed into the NAM module. Finally, the obtained feature components are split and upsampled, and the resulting feature maps are nonlinearly transformed using an activation function, allowing the horizontal and vertical components to be multiplied with the original features of the input module. The detailed structure of NACAM is shown in Figure 2.

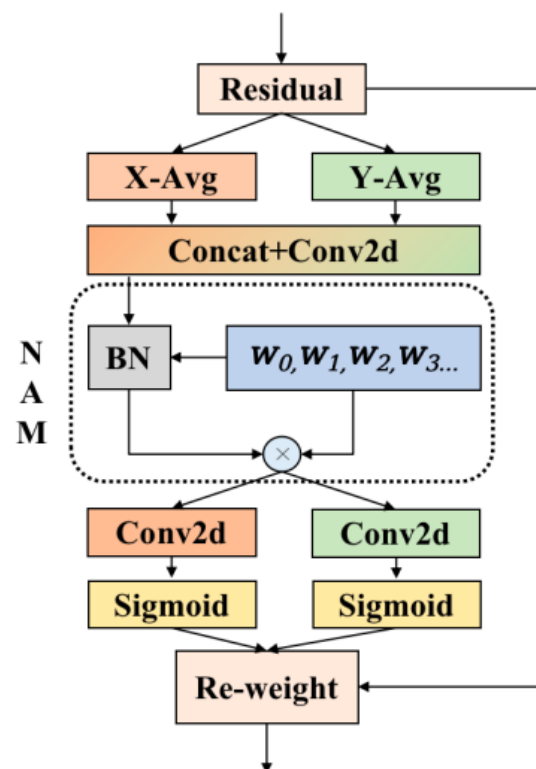


Figure 2. The detailed structure of the NACAM module.

For the application of NACAM, we adopt several mainstream classification networks and insert the NACAM module into the backbone to train and recognize fine-grained aircraft image datasets. Here, we use the ResNet series and the MobileNet series as examples to illustrate the insertion position of the NACAM module, as shown in Figure 3.

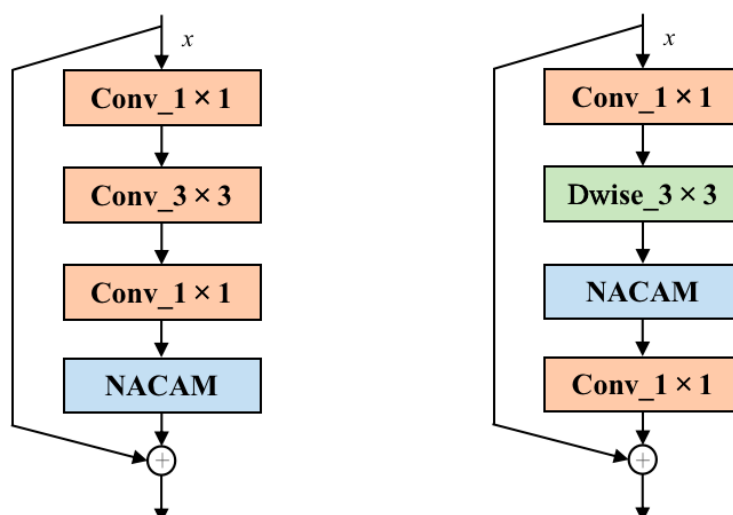


Figure 3. The detailed structure of the NACAM module.

3.3. Counterfactual Learning CCAM Module

The counterfactual attention model based on causal inference establishes causal relationships between features and prediction outcomes, guiding the learning process of attention maps. This makes the model's prediction logic more transparent, avoiding the black-box nature and lack of interpretability common in deep learning models. In this paper, we highlight the role of the attention module throughout the learning process by introducing causal inference methods. The causal model we construct is shown in Figure 4.

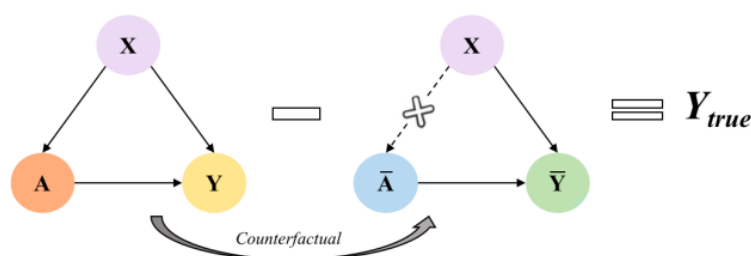


Figure 4. The detailed structure of the CCAM module.

The link $X \rightarrow A$ indicates that the attention model takes CNN feature maps as input and produces the corresponding attention maps.

The link $X \rightarrow A \rightarrow Y$ indicates that both the feature maps and attention maps jointly determine the final prediction.

We refer to node X as the causal parent of A , and Y as the causal child of X and A . Ideally, A predicts Y by capturing the effective properties of X . However, in real-world situations, there are some confusing attention regions in X that affect the learning process, causing the network to fall into suboptimal attention regions. Therefore, when we want to observe the effect of a particular variable, we perform an intervention by changing the state of different variables. For example, in our causal graph, $do(A = \bar{A})$ means we set the variable A to take the value $\bar{A}$ and cut the link $X \rightarrow A$, replacing the learned attention map with an imagined counterfactual attention map, in order to force the outcome to no longer be influenced by X .

Next, we will introduce the CCAM module in detail. Given an input image I , first use the backbone in section B to extract high-level features f^{l-1} .

In the first step, this module first uses an attention block to generate multi-head attention maps. As shown in Figure 5, the attention block is a lightweight model composed

of a 1×1 convolutional layer, a batch normalization layer, and a nonlinear activation layer ReLU. The features extracted from the specific layer f^{l-1} are input into this attention block to obtain M attention maps $A \in R^{M \times W \times H}$, where $A^k \in R^{W \times H}$ represents the k -th attention map. We use Bilinear Attention Pooling (BAP) to obtain the feature maps. First, we use bilinear interpolation to adjust the attention maps to the same scale as the feature maps, and then use the Attention Maps as a guide to re-assign each element in the Feature Maps. Then pooling and vectorization operations are performed to obtain the feature maps f_e . To generate the attention maps of the same channel in the same category, we focus on specific regions and assume that the attention map of the first channel of category one only focuses on the wing. A feature center c^t is set based on the number of categories, the number of feature maps, and the number of attention channels, and c^t is updated in the sliding-average manner in the following formula, with α set to 0.05 and f_k as the feature matrix. Through continuous update iterations, c^t tends to be stable.

$$c^t = c^{t-1} - \alpha(c^{t-1} - f_k) \quad (8)$$

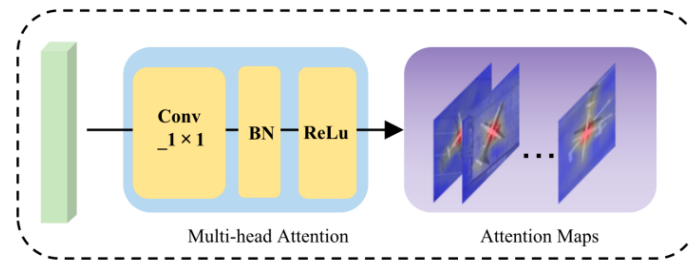


Figure 5. Multi-head attention mapping.

In the second step, we perform interpretability analysis on the prediction results by generating random counterfactual attention maps. The values of attention and counterfactual attention, A and $\bar{A}$, are generated by the processes $A(\cdot)$ and $\bar{A}(\cdot)$, respectively.

$$A = A(x) = \{A_0, A_1, \dots, A_{M-1}\} \quad (9)$$

$$\bar{A} = \bar{A}(x) = \{\bar{A}_0, \bar{A}_1, \dots, \bar{A}_{M-1}\} \quad (10)$$

The counterfactual attention maps can be achieved by randomly generating some attention maps. For a specific discriminative region, we obtain the feature representations of the attention map f_{matrix} and the counterfactual attention map $\bar{f}_{matrix}$ through the dot product, as shown below.

$$f_{matrix} = \sum_{m=0}^{M-1} (f^{l-1} \odot A_m) \quad (11)$$

$$\bar{f}_{matrix} = \sum_{m=0}^{M-1} (f^{l-1} \odot \bar{A}_m) \quad (12)$$

where $\odot$ denotes the element-wise multiplication of two tensors. In this model, we designed a classifier C composed of fully convolutional layers and batch normalization layers. Using attention maps for classification, the output $Y_{(local)}$ is expressed as:

$$Y_A = C[f_{matrix}] = C \left[\sum_{m=0}^{M-1} (f^{l-1} \odot A_m) \right] \quad (13)$$

Using the possibility of counterfactual intervention to analyze the direct causal relationship between A and Y after excluding confounding factors, we obtain the final prediction $\hat{Y}(\hat{f}_{local})$ after the intervention $A = \bar{A}$:

$$Y_{\bar{A}} = C[\hat{f}_{matrix}] = C\left[\sum_{m=0}^{M-1} (f^{l-1} \odot \bar{A}_m)\right] \quad (14)$$

The output after counterfactual intervention attention can be represented as $Y_{\bar{A}}$. By calculating the difference between Y_A and $Y_{\bar{A}}$, we can quantify the impact of multi-head attention features on the prediction results:

$$Y_{true} = Y_A - Y_{\bar{A}} \quad (15)$$

Y_{true} can be understood as the learning objective of the local attention mechanism, representing the intrinsic correlation between the attention map and prediction results after excluding false attention interference. Maximizing the likelihood difference forces the network to focus on fact-based attention learning rather than falling into the chaos represented by counterfactuals. Therefore, Y_{true} can be used to evaluate the learning quality of the attention module and serve as a supervision signal for the attention learning process.

We adopt Bilinear Attention Pooling (BAP) as the feature aggregation mechanism in the CCAM module due to its strong capacity for modeling region-level interactions and enhancing fine-grained feature representations. Unlike global average or max pooling, BAP explicitly captures the pairwise multiplicative relationships between attention maps and feature maps, allowing for a richer and more discriminative encoding of spatial dependencies. This capability is particularly well-aligned with causal reasoning, where the goal is to assess whether certain regions are necessary for a prediction outcome. In counterfactual analysis, evaluating the joint effect of localized features is critical for identifying truly causal areas. By modeling these interactions bilinearly, BAP offers an effective and interpretable way to support the intervention-based framework of counterfactual reasoning, thereby enhancing both causal discriminability and model robustness.

3.4. Loss Function

In order to ensure that each Attention Map consistently focuses on the same object part across different images, we additionally introduce the sum of squared differences between the feature map and the part center as a penalty term. For example, attention map A_1 focuses on the tail fins in different images, and A_2 focuses on the side wings. This reduces the randomness of the information captured by each attention map. The following formula encourages each feature map to be anchored to the center of each part, where the part centers are also updated during learning based on the feature maps, ensuring that the features of the same part on the same object are as similar as possible. Therefore, we propose a region independence loss to reduce the overlap between feature maps and maintain consistency across different inputs. With the help of center loss, the region independence loss is defined as follows:

$$L_{att} = \sum_{i=1}^B \sum_{j=1}^M \max\left(\|V_j^i - c_j^i\|_2^2 - m_{in}(y_i), 0\right) + \sum_{i,j \in (M,M), i \neq j} \max\left(m_{out} - \|c_i^t - c_j^t\|_2^2, 0\right) \quad (16)$$

where B is the batch size, M is the number of attention heads, m_{in} represents the margin between features and their corresponding centers, set to 0.1 and 0.05 depending on whether y_i is 0 or 1. m_{out} is the distance between each feature center. $c \in \mathbb{R}^{M \times N}$ denotes the feature centers of V , updated in each iteration as defined in Equation (8). The first part of L_{att} is

the intra-class loss that pulls V closer to the center c , while the second part is the inter-class loss that pushes centers apart.

In this study, the total loss function of the interpretable attention network consists of attention loss, classification loss, etc., as follows:

$$L_{all} = L_{CE}(Y_{true}, Y) + L_{CE}(Y_A, Y) + L_{att} + L_{RIL} \quad (17)$$

where L_{CE} denotes cross-entropy loss, Y represents ground truth labels, and L_{att} measures the difference between feature maps and center locations. Due to the lack of fine-grained labels, the network is prone to performance degradation.

4. Results

4.1. Experimental Setup

We adopt two public datasets, FAIR1M and MTARSI 2, as well as a self-constructed dataset, ARSI120, to train and evaluate the proposed network. The dataset splits are shown in Table 1. All experiments are conducted based on the PyTorch(2.3.0) deep learning framework, using Ubuntu 20.04, 32 GB RAM, an Intel(R) Core(TM) i7-6770K CPU, and four NVIDIA RTX 4090 GPUs.

Table 1. Datasets used in paper.

Name	Description		
	Class	Train	Test
FAIR1M	10	32,763	8191
MTARSI2	40	8164	2041
ARSJ120	101	11,553	2889

4.2. Dataset Descriptions

FAIR1M [39] is a large-scale dataset designed for fine-grained object detection and recognition in remote sensing imagery. It contains over one million instances across more than 15,000 images. Given the objectives of this study, we selected only the images that include complete airplane targets. After processing and filtering, we obtained 40,954 airplane image patches covering 10 types of airplanes.

MTARSI2 [40] is a reclassified and expanded version of the MTARSI dataset. The data were reorganized into 40 categories with additional data augmentation for each class. After filtering and organizing, the resulting dataset includes 10,205 images.

Proprietary Dataset ARSI120: To advance the state-of-the-art method in aircraft classification from remote sensing imagery, we constructed a large-scale dataset named ARSI120, consisting of 120 airplane categories. Addressing the issue of limited airplane categories in existing fine-grained aircraft recognition datasets, we manually collected 22,040 images representing 120 different airplane types. All sample images are carefully annotated, each containing exactly one complete airplane. The number of images per category ranges from 50 to 900, depending on the aircraft type. During dataset construction, we carefully selected representative samples for each category from airports worldwide, including those in the United States, Russia, Japan, France, and China. These images were captured under diverse conditions, such as varying times of day, seasons, and imaging settings, which enhances the intra-class variability of the dataset. We specifically ensured that the same airplane type was shown with different backgrounds and orientations to further enrich the diversity within each class. Due to factors such as lighting conditions, cloud cover, and variable spatial resolutions, remote sensing images may exhibit a range of visual characteristics. Furthermore, because some aircraft models are rare or sensitive due to their cost, function,

or restricted availability, sample imbalance is a significant challenge. To address this, we applied data augmentation techniques for underrepresented classes. Augmentation operations included simulation of various lighting conditions, image rotation at different angles, and cropping from the top, bottom, left, and right at different ratios. To meet the practical requirements of remote sensing aircraft recognition and deep learning, we ensured that each class in the ARSI120 dataset includes at least 60 images after augmentation. Figure 6 is a statistical display chart of the number of some categories in the dataset.

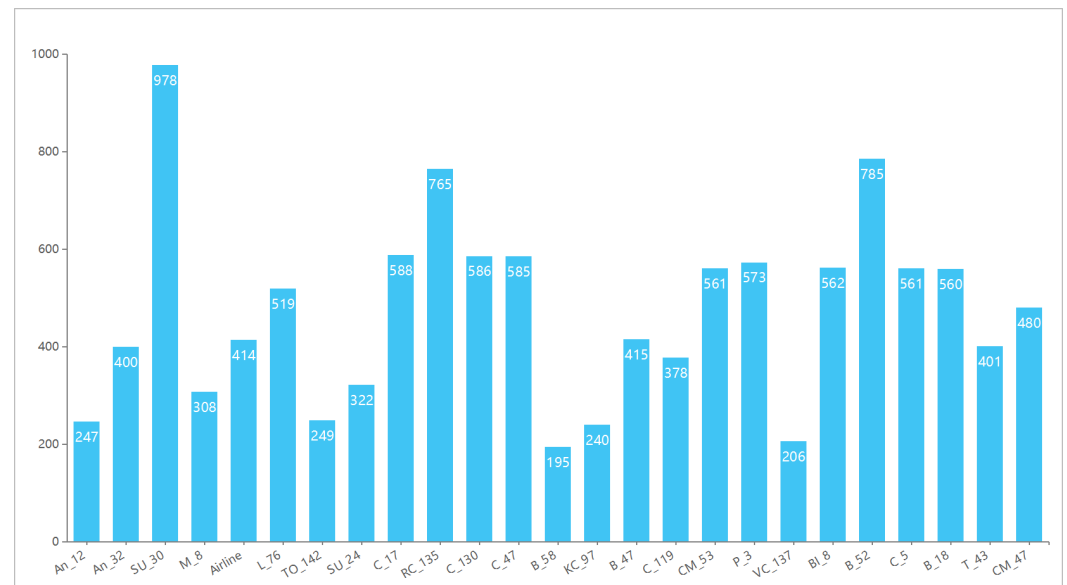


Figure 6. The number of certain categories in the dataset.

As illustrated in Figure 7, a subset of the dataset images reveals that the wings and tail structures of different aircraft consistently exhibit highly symmetrical configurations, while the spatial relationships between these components differ significantly across aircraft types. By systematically analyzing the variations in their relative positions, it becomes feasible to effectively distinguish among different categories of aircraft. Accordingly, the proposed model adopts multiple targeted strategies to attend to these critical regions, thereby enhancing the accuracy and robustness of aircraft type recognition and classification.



Figure 7. Some images of the dataset are displayed.

4.3. Ablation Study

To validate the effectiveness of each component of the proposed network, we conducted ablation experiments on three public datasets, employing ResNet101 [41] and MobileNetV2 [42] as the feature extraction backbones. The detailed results are reported in Tables 2 and 3.

Table 2. Ablation study in FAIR1M.

Backbone	Model			Params (M)	FLOPs (G)	FAIR1M		MTARSI2		ARSI120	
	CAM	NACAM	CCAM			Top-1 acc (%)	Top-5 acc (%)	Top-1 acc (%)	Top-5 acc (%)	Top-1 acc (%)	Top-5 acc (%)
ResNet101				44.5	7.82	96.1	99.7	93.1	99.1	95.6	98.6
	✓			46.1	7.91	96.7	99.8	94.1	99.8	97.5	98.9
		✓		48.9	7.92	96.8	99.8	94.4	99.8	97.6	99.1
			✓	45.2	7.92	96.9	99.9	94.8	100	97.7	99.2
	✓		✓	48.8	7.93	96.4	99.8	94.6	100	97.8	99.2
		✓	✓	48.9	7.93	97.7	99.9	95.2	100	98.4	99.5
MobileNet V2				3.5	0.32	93.2	99.6	90.8	98.6	93.2	96.7
	✓			3.9	0.34	94.1	99.7	91.2	99.4	93.3	96.8
		✓		4.2	0.34	94.2	99.7	91.3	99.5	93.5	96.8
			✓	4.1	0.34	94.6	99.8	92.1	99.5	93.8	97.1
	✓		✓	4.3	0.34	95.2	99.9	92.0	99.4	94.3	97.2
		✓	✓	4.4	0.34	95.8	99.9	92.7	99.8	94.9	97.6

Table 3. Compared with other method.

Models	Backbone	Acc(%)		
		Fair1M	MTARSI2	ARSI120
Resnet101		96.1	93.1	95.6
MobileNetV2		93.2	90.8	93.2
Vision Transformer [18]		96.8	93.2	95.9
WS-DAN [43]	ResNet50	96.9	93.4	96.1
API-Net [44]	DenseNet161	96.4	94.7	97.3
FFVT [45]	ViT	97.1	95.1	98.2
HERBS [46]	ViT	97.1	95.1	98.1
(Ours)	Resnet101	97.7	95.2	98.4
(Ours)	MobileNetV2	95.8	92.7	94.9

Table 2 presents the Top-1 and Top-5 accuracy of the networks employing different modules in the FAIR1M, MTARSI2, and ARSI120 datasets.

Effectiveness of NACAM: To evaluate the effectiveness of the proposed Normalized Coordinate Attention Module (NACAM) in fine-grained aircraft classification tasks, we conducted a series of comparative experiments. Under identical convolutional backbones, the introduction of NACAM led to consistent improvements in classification accuracy across different datasets. Specifically, using ResNet101[19] as the backbone, the classification accuracy in the FAIR1M, MTARSI2, and ARSI120 datasets increased by 1.6%, 2.1%, and 2.8%, respectively. Compared to models that used only the CAM module, the accuracy of Top-1 improved by 0.1%, 0.3%, and 0.1% across these dataset; compared to models that used the CAM and CCAM modules, the accuracy of Top-1 improved by 1.3%, 0.6%, and 0.6%, respectively. Similarly, when using MobileNetV2 as the backbone, classification accuracy improved by 2.6%, 1.9%, and 1.7% on these datasets, with Top-1 accuracy gains of 0.1%, 0.1%, and 0.2% over the CAM-only models, and 0.6%, 0.7%, and 0.6% over the CAM+CCAM models. These results demonstrate that integrating the NACAM module enables the network to more effectively focus on key features tightly associated with the spatial structure of aircraft images, thereby improving fine-grained classification accuracy. Moreover, we observed that incorporating NACAM does not significantly increase the model's floating-point operations or parameter count, indicating that the module

achieves an optimal balance between accuracy and computational efficiency, and offers advantages in terms of lightweight performance. Overall, the consistent improvements observed with both ResNet101 and MobileNetV2 backbones confirm the effectiveness of NACAM in capturing critical features, enhancing classification accuracy, and maintaining lightweight performance.

Effectiveness of CCAM: Compared to configurations using only NACAM, we further observed that integrating the Counterfactual Attention Module (CCAM) under the same convolutional backbone led to additional gains in classification accuracy. Specifically, with ResNet101 as the backbone, the classification accuracy on the FAIR1M, MTARSI, and MTARSI2 datasets increased by 1.4%, 1.3%, and 0.6%, respectively, corresponding to Top-1 accuracy improvements of 1.6%, 1.4%, and 1.4% over models using only NACAM. Likewise, using MobileNetV2 as the backbone, classification accuracy improved by 1.4%, 1.3%, and 0.6% on these datasets, with corresponding Top-1 accuracy gains of 1.6%, 1.4%, and 1.4% over NACAM-only models. These experimental results indicate that the randomly generated Counterfactual Attention Module (CCAM) significantly enhances classification accuracy under identical convolutional backbones. This improvement can be attributed to the CCAM module's ability to make the model's prediction logic more transparent by guiding the attention mechanism to explore the true causal relationships between attended regions and classification outcomes. These findings further validate the effectiveness of CCAM and its capability to improve performance in fine-grained aircraft classification tasks.

Figures 8 and 9 further demonstrate the effectiveness of the NACAM and CCAM modules on the two feature extraction networks, ResNet101 [39] and MobileNetV2 [40], through visual charts.

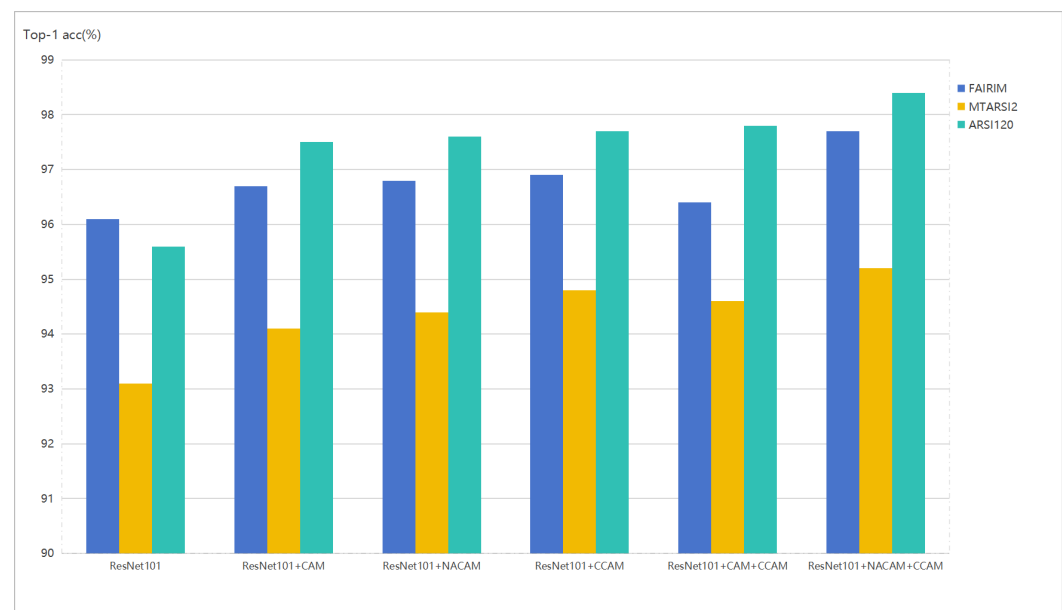


Figure 8. The accuracy of the ResNet101 extractor with different network structure configurations.

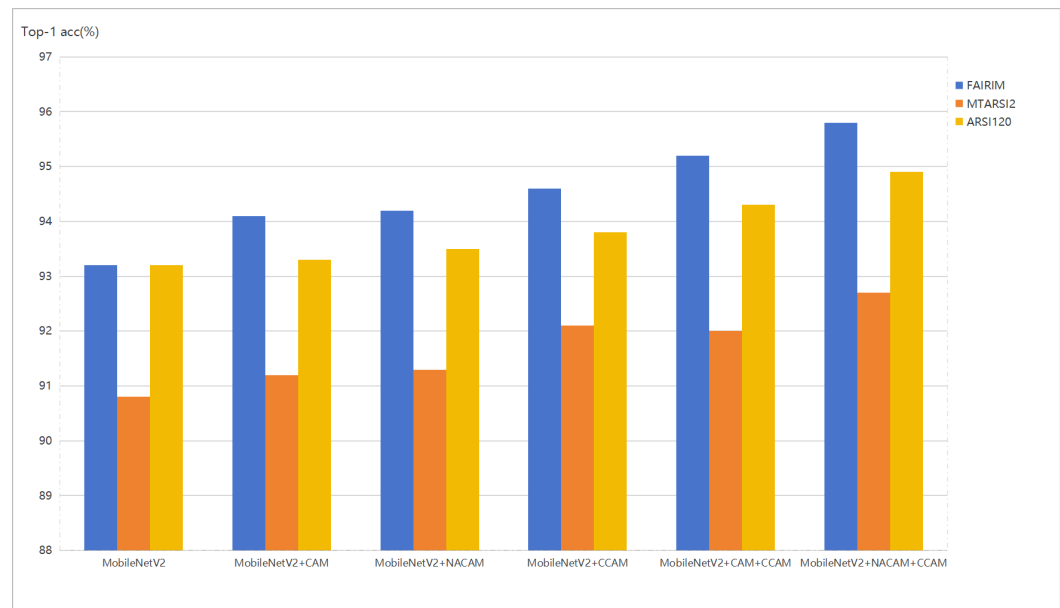


Figure 9. The accuracy of the MobileNetV2 extractor with different network structure configurations.

4.4. Comparison with Other Methods

To demonstrate the effectiveness of our proposed network, we conducted a comprehensive comparison against several existing fine-grained classification models. Specifically, we compared our approach with traditional deep learning models, including ResNet101, MobileNetV2, and Vision Transformer (ViT); attention-based fine-grained classification models, such as WS-DAN and API-Net; as well as state-of-the-art ViT-based fine-grained classification models, including FFVT and HERBS. The detailed comparative results are summarized in Table 3.

- **ResNet101** : This model introduces residual connections to solve the gradient problem, making the network training more stable.
- **MobileNetV2**: This lightweight convolutional neural network architecture employs strategies like depthwise separable convolutions to reduce model parameters and computational costs while maintaining good performance.
- **Vision Transformer**: As a currently popular backbone model, it utilizes self-attention to capture both global and local relationships.
- **WS-DAN**: This model adopts a weakly-supervised learning-based image enhancement method combined with attention mechanisms for fine-grained classification.
- **API-Net**: The model proposes a simple yet effective attention-pairing interaction network that progressively recognizes pairs of fine-grained images through interaction, achieving 90% accuracy on the CUB200-2011 dataset.
- **FFVT**: Applying the newer Transformer framework for feature fusion, this visual transformer compensates for local, low-level, and intermediate information by aggregating important tokens from each transformer layer, reaching 91.6% accuracy on the CUB200-2011 dataset.
- **HERBS**: Comprising two modules—a high-temperature refinement module and a background suppression module—this model extracts discriminative features while suppressing background noise, achieving 93% accuracy on the CUB200-2011 dataset.

In this subsection, we provide a detailed comparison between our method and all baseline models, summarizing the experimental results on three datasets, as shown in Table 3. The following is a detailed explanation of these observations: Overall, our method demonstrates superior performance across all three datasets. From the experimental results,

we observe that models using the lightweight backbone MobileNetV2 achieve lower accuracy compared to models using ResNet101, likely due to the relatively shallow depth of MobileNetV2, while ResNet101 can more effectively capture deep features from the data.

For the FAIR1M dataset, taking the models with ResNet101 as the backbone as an example, our method improves accuracy compared to the best-performing baseline methods by 0.8%, 1.3%, 0.6%, and 0.6%, respectively. For the MTARSI2 dataset, our method achieves improvements of 1.8%, 0.5%, 0.1%, and 0.1% compared to the best baseline methods. For the ARSI120 dataset, our method shows improvements of 2.3%, 1.1%, 0.2%, and 0.3%, respectively.

When compared with current attention-based models such as WS-DAN and API-Net, our proposed Normalized Coordinate Attention Module (NACAM) achieves higher Top-1 accuracy on all three datasets. This demonstrates that our model can more accurately locate key parts and more precisely leverage the positional information associated with target features, thereby mitigating the negative impact of irrelevant features and improving its applicability to fine-grained aircraft classification tasks. Moreover, our proposed model can capture long-range dependencies and adjust the weights between features at different positions through positional encoding. Compared to the Vision Transformer baseline model and self-attention-based models such as FFVT and HERBS, our method achieves higher Top-1 accuracy across all three datasets. In addition, the proposed NCC-Net offers high computational efficiency and is easy to implement.

These results further emphasize the superiority of our approach in extracting positional information, which enhances the model's focus on local features of the target and effectively improves accuracy in fine-grained aircraft classification tasks. This indicates that our counterfactual attention network is a powerful tool for enhancing model performance.

4.5. Interpretability Analysis

We adopted the Gradient-Weighted Class Activation Mapping (Grad-CAM) method for feature analysis and applied it to three different datasets. The visualization results in Figure 10 illustrate the impact of various convolutional neural networks in learning category-specific features, as well as the discriminative power and reliability of the attention features used during training. These visualizations provide an intuitive explanation for the performance of our method across different datasets, reinforcing its superior performance in fine-grained aircraft classification tasks.

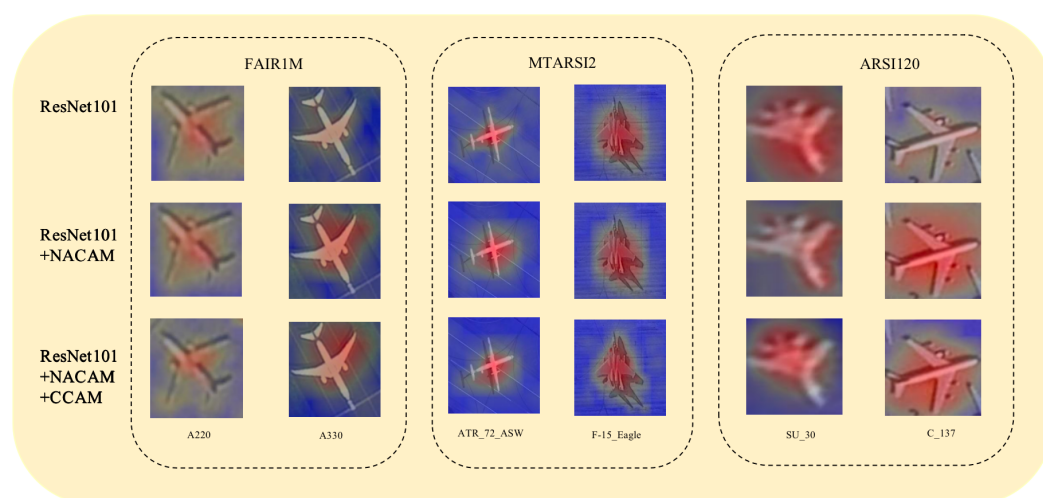


Figure 10. Gradient-weighted class activation mapping.

By closely examining the key points in the images, we can clearly observe that our method aligns more accurately with the aircraft contours compared to the conventional

convolutional backbone network ResNet101. Our approach focuses more precisely on the aircraft targets, delineating the boundaries between the targets and the background. For example, on the FAIR1M dataset, our method can more accurately outline the contour of the A330 aircraft, reducing overlap with the background. Similarly, on the MTARSI2 dataset, our approach can more precisely depict the outline of the F-15 aircraft and reduce background interference, thereby improving classification accuracy.

The NACAM module enables the network to more accurately capture the key parts of the aircraft while ignoring other irrelevant regions in the image. The CCAM module further directs the network's focus toward the aircraft target, reducing attention to background information and showing a lower interest in irrelevant areas. These visualization results further highlight the superiority of the proposed NACAM and CCAM modules in fine-grained aircraft classification tasks. They help the network better understand and focus on the critical features of the aircraft, thereby enhancing classification performance. This suggests that our approach holds promising potential in fine-grained classification, particularly in distinguishing aircraft from the background, demonstrating excellent performance and clearer feature representations.

In order to more directly evaluate the model's performance, we used the t-SNE algorithm to generate two-dimensional feature distribution maps for visual analysis of the model. t-SNE is an unsupervised nonlinear dimensionality reduction technique that can capture the complex manifold structure of high-dimensional data and is widely used for exploring and visualizing such data.

By comparing the feature distributions in Figures 11–13, we can clearly observe that the feature distribution of the NCC-Net model is more compact compared to that of using only ResNet101. This means that our proposed model can more effectively separate different semantic information, making the feature kernels with similar semantic information closer together in the feature space. This further verifies that our network can reduce the mutual interference between discriminative features of different target parts, provide higher-quality local features, and thus significantly improve classification performance. By observing the feature distribution on ARSI20, we find that even when facing more than 100 categories, our classification algorithm still shows excellent inter-class clustering performance. This further confirms the effectiveness of our proposed ResNet101 + NACAM + CCAM model in the fine-grained aircraft classification task. This kind of visual analysis provides intuitive and strong support for our research, further emphasizing the superiority of the model in capturing category-specific features.

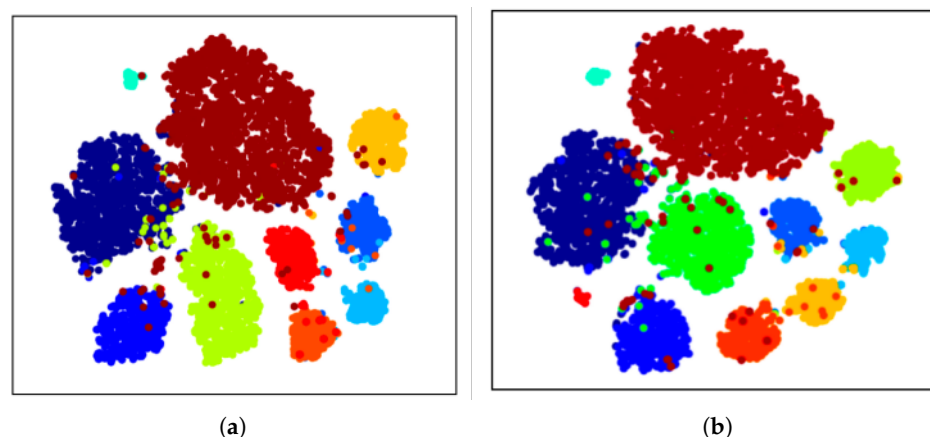


Figure 11. Feature distribution of some filter groups learned on the FAIR1M dataset. (a) Feature visualization of ResNet101. (b) Feature visualization of ResNet101 + NACAM + CCAM.

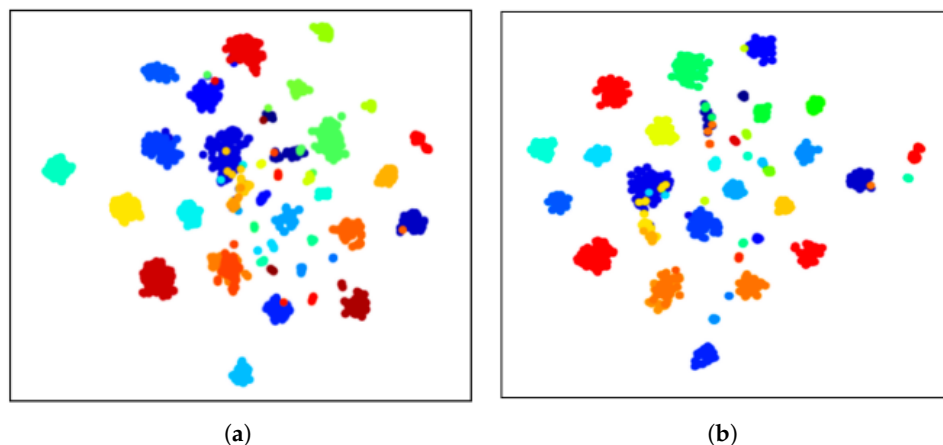


Figure 12. Feature distribution of some filter groups learned on the MTARSI2 dataset. (a) Feature visualization of ResNet101. (b) Feature visualization of ResNet101 + NACAM + CCAM.

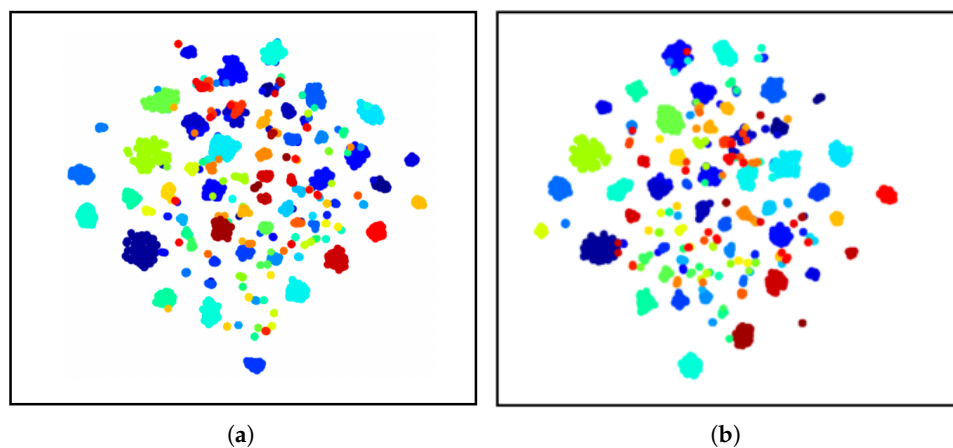


Figure 13. Feature distribution of some filter groups learned on the ARSI20 dataset. (a) Feature visualization of ResNet101. (b) Feature visualization of ResNet101 + NACAM + CCAM.

5. Conclusions

In our study, we proposed a novel normalized coordinate attention-based counterfactual fine-grained classification network for remote sensing aircraft. This network integrates the basic fine-grained classification framework while leveraging normalized coordinate attention to emphasize the positional information of aircraft parts. It can better capture subtle differences between aircraft and obtain long-range dependencies between different positions, enabling more precise classification. By using causal guidance to adjust the predictions between the original image representation and the counterfactual representation, the classification model becomes more focused on invariant foreground information while ignoring varying background information, effectively eliminating the influence of co-occurring backgrounds and further addressing the problem of the fine localization of aircraft targets. We conducted extensive experiments on multiple remote sensing aircraft classification datasets to demonstrate the effectiveness of this model.

In future work, we aim to further optimize both NACAM and CCAM modules to better handle scale variations and background interference. We also plan to explore the generalization of NCC-Net to multi-object and multi-category scenarios in remote sensing, incorporating few-shot learning and cross-domain causal reasoning to enhance performance under data-scarce or dynamic conditions. In addition, we will also focus on the reasoning efficiency of the proposed network, including optimizing the model complexity, reducing

memory usage, and accelerating the prediction speed without affecting the accuracy rate. Meanwhile, we will explore the promotion capability of this network on other remote sensing targets such as ships and vehicles and verify its cross-task migration performance.

Author Contributions: Conceptualization, X.Z.; Methodology, Z.Z.; Formal analysis, S.Z.; Writing—original draft, W.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NAME OF National Natural Science Foundation of China grant number 61871326.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hsieh, J.W.; Chen, J.M.; Chuang, C.H.; Fan, K.C.. Aircraft type recognition in satellite images. *IEE Proc. Vis. Image Signal Process.* **2005**, *152*, 307–315. [[CrossRef](#)]
2. Liu, G.; Sun, X.; Fu, K.; Wang, H. Aircraft recognition in high-resolution satellite images using coarse-to-fine shape prior. *IEEE Geosci. Remote Sens. Lett.* **2012**, *10*, 573–577. [[CrossRef](#)]
3. Zhao, A.; Fu, K.; Wang, S.; Zuo, J.; Zhang, Y.; Hu, Y.; Wang, H. Aircraft recognition based on landmark detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1413–1417. [[CrossRef](#)]
4. Goring, C.; Rodner, E.; Freytag, A.; Denzler, J. Nonparametric part transfer for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2489–2496.
5. Branson, S.; Van Horn, G.; Belongie, S.; Perona, P. Bird species categorization using pose normalized deep convolutional nets. *arXiv* **2014**, arXiv:1406.2952. [[CrossRef](#)]
6. Lin, D.; Shen, X.; Lu, C.; Jia, J. Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1487–1500. [[CrossRef](#)]
7. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 842–850.
8. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear Convolutional Neural Networks for Fine-grained Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1309–1322. [[CrossRef](#)]
9. Sun, M.; Yuan, Y.; Zhou, F.; Ding, E. Multi-attention multi-class constraint for fine-grained image recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 805–821.
10. Peng, Y.; He, X.; Zhao, J. Object-part attention model for fine-grained image classification. *IEEE Trans. Image Process.* **2017**, *27*, 1487–1500. [[CrossRef](#)]
11. Rodriguez, P.; Velazquez, D.; Cucurull, G.; Gonfaus, J.M.; Roca, F.X.; Gonzalez, J. Pay attention to the activations: A modular attention mechanism for fine-grained image recognition. *IEEE Trans. Multimed.* **2019**, *22*, 502–514. [[CrossRef](#)]
12. Zheng, H.; Fu, J.; Zha, Z.J.; Luo, J. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5012–5021.
13. Bera, A.; Wharton, Z.; Liu, Y.; Bessis, N.; Behera, A. SR-GNN: Spatial relation-aware graph neural network for fine-grained image categorization. *IEEE Trans. Image Process.* **2022**, *31*, 6017–6031. [[CrossRef](#)]
14. Touvron, H.; Sablayrolles, A.; Douze, M.; Cord, M.; Jégou, H. Graft: Learning fine-grained image representations with coarse labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 874–884.
15. Sun, X.; Xu, H.; Dong, J.; Zhou, H.; Chen, C.; Li, Q. Few-shot learning for domain-specific fine-grained image classification. *IEEE Trans. Ind. Electron.* **2020**, *68*, 3588–3598. [[CrossRef](#)]
16. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

17. Van der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **2014**, *15*, 3221–3245.
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
19. He, J.; Chen, J.N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C. TransFG: A transformer architecture for fine-grained recognition. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 852–860. [[CrossRef](#)]
20. Raistrick, A.; Lipson, L.; Ma, Z.; Mei, L.; Wang, M.; Zuo, Y.; Kayan, K.; Wen, H.; Han, B.; Wang, Y.; Newell, A. Infinite Photorealistic Worlds Using Procedural Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 12630–12641. [[CrossRef](#)]
21. Sermanet, P.; Frome, A.; Real, E. Attention for fine-grained categorization. *arXiv* **2014**, arXiv:1412.7054.
22. Zheng, H.; Fu, J.; Mei, T.; Luo, J. Learning multi-attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5209–5217.
23. Chen, J.; Gao, Z.; Wu, X.; Luo, J. Meta-Causal Learning for Single Domain Generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7683–7692. [[CrossRef](#)]
24. Ding, Y.; Ma, Z.; Wen, S.; Xie, J.; Chang, D.; Si, Z.; Wu, M.; Ling, H. AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Trans. Image Process.* **2021**, *30*, 2826–2836. [[CrossRef](#)]
25. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
27. Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; Liu, Z. Mobile-Former: Bridging MobileNet and Transformer for Efficient Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–21 June 2024; pp. 6541–6550.
28. Price, I.; Tanner, J. Improved projection learning for lower dimensional feature maps. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
29. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global second-order pooling convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3024–3033.
30. Liu, J.J.; Hou, Q.; Cheng, M.M.; Wang, C.; Feng, J. Improving convolutional networks with self-calibrated convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10096–10105.
31. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
32. Tang, K.; Niu, Y.; Huang, J.; Shi, J.; Zhang, H. Unbiased scene graph generation from biased training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3716–3725.
33. Rao, Y.; Chen, G.; Lu, J.; Zhou, J. Counterfactual attention learning for fine-grained visual categorization and re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1025–1034.
34. Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.S.; Wen, J.R. Counterfactual VQA: A cause-effect look at language bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12700–12710.
35. Xiong, W.; Xiong, Z.; Cui, Y. An explainable attention network for fine-grained ship classification using remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620314. [[CrossRef](#)]
36. Pan, C.; Li, R.; Hu, Q.; Niu, C.; Liu, W.; Lu, W. Contrastive Learning Network Based on Causal Attention for Fine-Grained Ship Classification in Remote Sensing Scenarios. *Remote Sens.* **2023**, *15*, 3393. [[CrossRef](#)]
37. Xu, K.; Wang, L.; Zhang, Q. Causal Domain Adaptation via Contrastive Conditional Transfer for Fine-Grained Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–21 June 2024; pp. 412–419.
38. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based attention module. *arXiv* **2021**, arXiv:2111.12419. [[CrossRef](#)]
39. Rudd-Orthner, R.; Mihaylova, L. Multi-type Aircraft of Remote Sensing Images: MTARSI 2. Zenodo. 2021. Available online: <https://oa.mg/work/10.5281/zenodo.5044950> (accessed on 28 July 2025).
40. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; Weinmann, M. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [[CrossRef](#)]
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

42. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
43. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; Weinmann, M. SWS-DAN: Subtler WS-DAN for fine-grained image classification. *J. Vis. Commun. Image Represent.* **2021**, *79*, 103245. [[CrossRef](#)]
44. Dong, X.; Liu, H.; Ji, R.; Cao, L.; Ye, Q.; Liu, J.; Tian, Q. Api-net: Robust generative classifier via a single discriminator. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 379–394. [[CrossRef](#)]
45. Wang, J.; Yu, X.; Gao, Y. Feature fusion vision transformer for fine-grained visual categorization. *arXiv* **2021**, arXiv:2107.02341.
46. Chou, P.Y.; Kao, Y.Y.; Lin, C.H. Fine-grained visual classification with high-temperature refinement and background suppression. *arXiv* **2023**, arXiv:2303.06442.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.