AI Engineering
https://www.sciltp.com/journals/aieng

Check for updates

Article

Bird-Drone Recognition under Closed-Set and Open-Set Scenarios: A Comparative Analysis of Deep Learning Models

Zehua Tang *, Victor Lawrence and Hong Man

Department of Electrical & Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA
* Correspondence: ztang24@stevens.edu or tangzh656995@gmail.com

How To Cite: Tang, Z.; Lawrence, V.; Man, H. Bird-Drone Recognition under Closed-Set and Open-Set Scenarios: A Comparative Analysis of Deep Learning Models. *AI Engineering* **2026**, *2*(1), 3. https://doi.org/10.53941/aieng.2026.100003

Received: 12 February 2026
Revised: 13 April 2026
Accepted: 20 May 2026
Published: 11 June 2026

Abstract: In practical applications, image recognition of birds and drones faces certain challenges, particularly in real-world scenarios where the system may encounter unknown aerial targets not included in the training data. Nevertheless, most current research still focuses primarily on classification performance under controlled conditions, assuming that test samples belong to known categories. To address this issue, this study conducted a systematic comparison of eight deep learning models (including four convolutional neural network (CNN) models and four Transformer models) under unified training and evaluation conditions. In addition to testing under closed conditions, an open-domain detection scenario was constructed by simulating real-world environments and introducing unknown categories during the testing phase. Experimental results indicate that while models achieve high accuracy under closed-domain conditions, their performance under open-domain conditions varies significantly. In particular, unknown targets that closely resemble drones are more difficult to correctly reject, while some unknown samples are easily misclassified as drones. Therefore, accuracy under closed-domain conditions does not fully reflect the reliability of model detection in real-world operational environments, and evaluation under open-domain conditions is of great significance for analyzing model performance and practical applications.

Keywords: deep learning; bird–drone recognition; open-set evaluation

1. Introduction

In today's world, research on the classification and recognition of flying objects still faces many challenges [1,2]. For example, correctly identifying birds in flight and distant aircraft at airports, or correctly identifying and distinguishing between birds in flight and potentially hazardous drones in restricted areas. However, most existing bird-drone studies report model detection performance under the closed-set assumption, where all species are known [1–6]. In this setting, modern deep learning models may achieve near-saturation accuracy due to their efficient learning capabilities, which could mask reliability issues that only become apparent when the test distribution deviates from the training data [7–9]. To address this issue, this paper compares the performance of CNN and Transformer models in the bird–drone recognition task under a unified experimental setup, with a particular focus on examining their behavioral differences under closed-set and open-set conditions.

Convolutional neural networks (CNNs) and vision Transformer-based models have become the backbone architectures commonly used in current image classification tasks. Among them, representative CNN architectures such as ResNet, VGG, DenseNet, and EfficientNet have demonstrated strong generalization and feature extraction capabilities [10–13], while Transformer-based models such as ViT and Swin Transformer offer a modeling approach distinct from traditional CNNs through global attention mechanisms and hierarchical representations [14,15].

Building upon these two types of architectures and based on closed-set experiments, we will further investigate the performance of bird-drone deep learning models in open-set scenarios.

In practical deployment, the system will receive images of aerial targets that did not appear during the training phase, such as airplanes, balloons, or other flying objects. Such scenarios correspond to open-set recognition problems, in which the model must not only recognize known classes but also minimize the likelihood of classifying unknown inputs with high confidence into known labels, thereby reducing open-set risk [7]. Addressing this requirement poses a critical challenge in selecting a scoring algorithm: under distribution shifts, classical classifiers based on the standard softmax function often exhibit excessively high confidence scores. Moreover, Maximum Softmax Probability (MSP)—a common and intuitive baseline method optimized on the softmax—is typically insufficient to reliably reject unknown samples [8]. To address this issue, existing research has proposed various alternatives, including OpenMax [16], ODIN [17], methods based on Mahalanobis distance [18], energy-based scoring methods [19], and hybrid methods that combine the logit layer with feature representation information, such as ViM [20]. Furthermore, some studies have employed calibration methods such as temperature scaling to improve the consistency between model confidence and prediction accuracy; research on these methods is of significant importance for security decision-making in open-world scenarios [21].

Against this research backdrop, this paper will compare the performance of representative Convolutional Neural Networks (CNNs) and Transformer models for bird and drone classification within a unified evaluation framework. All models employ the same data partitioning, pre-training initialization, training protocol, input dimensions, and open-set decision rules, thereby ensuring that the comparison focuses primarily on the model architectures themselves rather than differences in experimental settings. Furthermore, this paper defines open-set testing across three scenarios—"aircraft", "balloon", and "mixed unknown"—to correspond to real-world scenarios involving unknown object recognition of varying difficulty. In addition to traditional classification metrics, this paper conducts a more detailed examination of model behavior by introducing threshold sensitivity analysis, score-based open-set metrics, and confusion matrix analysis. The objective of this paper is not only to compare the recognition performance of different models under standard closed-set conditions but also to observe the behavioral differences they exhibit under open-set conditions that more closely resemble real-world deployment scenarios, thereby providing more practical guidance for model selection in bird-drone recognition systems.

The work presented in this paper focuses on the following aspects. First, under a unified framework encompassing data partitioning, pre-training initialization, training settings, and evaluation procedures, we conduct a systematic comparison of four convolutional neural network (CNN) models and four Transformer models, thereby minimizing the impact of experimental variations on the interpretation of results. Second, this study established an open-set evaluation framework encompassing near out-of-distribution (near-OOD), far out-of-distribution (far-OOD), and mixed unknown cases to analyze the identification and rejection difficulties of different types of unknown aerial targets in a more hierarchical manner. Furthermore, in addition to classification results at a fixed threshold, this study conducted a more comprehensive evaluation of model performance by incorporating threshold sensitivity analysis and score-based open-set metrics such as AUROC, AUPR, and FPR95. Finally, this paper analyzes the model's prediction behavior using confusion matrices and representative success and failure samples, focusing on the phenomenon where unknown samples are clustered into the "drone" class when they are not successfully rejected, as well as the typical performance of different architectures in known-class recognition and unknown-class handling. Research findings indicate that classification models that perform well on closed datasets do not necessarily perform reliably on open datasets. These findings provide guidance for selecting "flying object classification and recognition" models, ensuring that the chosen models are better suited to real-world environments.

2. Related Work

2.1. Bird–Drone Visual Recognition and Benchmarks

In applications related to air safety monitoring and drone surveillance, the identification of birds and drones has become a topic of significant interest. Due to the typically small size of the targets, complex backgrounds, and limited available pixel information, this task poses major challenges in terms of visual discrimination. Many existing studies have explored bird-drone classification methods based on convolutional neural networks (CNNs) and achieved good classification performance under conditions where the training and test sets share the same distribution. For example, Rahman et al. proposed a vision-based convolutional neural network recognition method and constructed a corresponding dataset, demonstrating that deep models can achieve competitive classification accuracy under controlled conditions [5]. Furthermore, in recent years, a series of publicly available datasets specifically designed for bird and drone detection or classification have been released to support reproducible

experimental comparisons. For instance, Shandilya et al. released a drone-bird segmentation dataset suitable for YOLO [3]; Akyon et al. proposed sequence-based model baselines and constructed corresponding trajectory/ sequence classification datasets to reduce false positives in temporally continuous scenarios [6]. Subsequent research has continued to expand the scale of datasets and model architectures; for example, the BirDrone dataset and the YOLOBirDrone framework both target scenarios involving small and complex targets [4].

Although these studies have advanced drone and bird recognition technology, most existing evaluations remain confined to closed-set scenarios or primarily treat false positives as detection errors, with few conducting specific analyses of model performance when encountering unknown aerial objects during testing. In this research context, the "Drones vs. Birds" Grand Challenge and related comparative studies have brought together various deep learning methods and highlighted practical challenges in real-world scenarios, such as scale variations, motion blur, and high false positive rates—conditions commonly encountered in real-world environments [1,2]. This is precisely the problem that the open-set evaluation framework proposed in this paper aims to address.

2.2. Open-Set Recognition and OOD Detection

The concept of Open Set Recognition (OSR) provides a formal description of scenarios where unknown categories may arise during the inference phase, emphasizing that models should not force all inputs into a known set of labels, but rather strive to mitigate open set risk [7]. For example, in real-world object recognition scenarios, a model should not classify objects other than drones and birds as one of the former two. Building on this, Out-of-Distribution (OOD) recognition proposes that there are degrees of similarity between different types of objects [9]; for instance, airplanes are more similar to drones and birds in terms of shape and surface texture, whereas balloons are not. Meanwhile, other studies have summarized the task classification and benchmarking of OSR from the perspective of image recognition [22]. To quantify this metric, existing researchers have designed various algorithms, such as the classic softmax algorithm and its optimized variant, Maximum Softmax Probability. However, these two algorithms still cannot reliably reject unknown samples [8]. Consequently, various advanced algorithms have emerged to further improve model performance, such as OpenMax [16], ODIN [17], methods based on Mahalanobis distance [18], energy-based scoring methods [19], and hybrid methods that combine the logit layer with feature representation information, such as ViM [20]. Meanwhile, other studies have summarized the task classification and benchmarking of OSR from the perspective of image recognition [22].

In this study, we adopt a threshold-based open-set evaluation method based on Maximum Softmax Probability (MSP) as a baseline to lay the groundwork for subsequent research. The set scenarios—near-OOD airplanes, far-OOD balloons, and mixed unknown samples—also correspond to a core issue emphasized in existing OSR/OOD literature: when unknown samples are semantically closer to known categories, their identification and rejection are typically more difficult [8,9].

2.3. CNNs vs. Vision Transformers under Distribution Shift

Convolutional neural networks (CNNs) and visual Transformers exhibit significant differences in inductive bias and feature representation; consequently, they often demonstrate markedly different generalization characteristics in classification tasks under out-of-distribution (OOD) conditions. In recent years, research on OOD generalization has increasingly examined both CNN and Transformer backbone networks, highlighting that a model's OOD performance is determined not only by in-distribution accuracy but also by the fact that different architectures typically require corresponding scoring or calibration strategies [9]. For example, the recently proposed ViM method integrates information from both the feature space and the log-likelihood space, achieving performance improvements across various backbone networks, including both CNNs and Transformers. This suggests that designing customized scoring mechanisms tailored to architectural features can help mitigate the overconfidence problem in open-set task scenarios [20].

Based on these insights, this paper conducts an application-oriented comparative analysis of representative Convolutional Neural Network (CNN) and Transformer models for the bird-drone recognition task. In addition to overall performance metrics, this paper further examines the behavioral differences between architectures in known-class recognition and unknown-class handling by combining confusion matrices and representative samples under various open-set conditions.

3. Materials and Methods

3.1. Problem Formulation

This paper investigates the bird–drone visual recognition task under two settings: closed-set and open-set.

In the closed-set setting, all test samples belong to the set of known classes observed during the training phase. In this study, the known classes include Bird and Drone. Both model training and evaluation are based on the assumption that no objects from other classes will appear during the testing phase.

In contrast, the open-set setting explicitly accounts for the possibility of unknown classes appearing during the testing phase. In this setting, the model is trained using only known classes (Bird and Drone), while the test data also includes samples from unseen classes that do not belong to any known class in the training label space. Consequently, the goal of open-set evaluation is not only to correctly identify known samples but also to ensure that the model can appropriately reject unknown samples rather than forcing them into known classes [7].

3.2. Dataset Construction and Open-Set Settings

The dataset used in this study includes two known classes (Bird and Drone), as well as multiple unknown classes introduced only during the open-set evaluation phase. Images for the four classes—Bird, Drone, Plane, and Balloon—were sourced from publicly available online resources, primarily the Kaggle dataset and Google Image Search. After the initial collection was completed, all images underwent manual screening and cleaning to improve data quality. During this process, duplicate or near-duplicate samples, low-resolution images, samples with unclear objects, and images with overly complex backgrounds or obvious distractions were removed. Additionally, images with ambiguous class labels or poor overall visual quality were excluded from the final dataset.

After this manual screening and cleaning process, the remaining images were reorganized by class and compiled into the final dataset used in this study. Representative image samples from both known and unknown classes are illustrated in Figure 1.

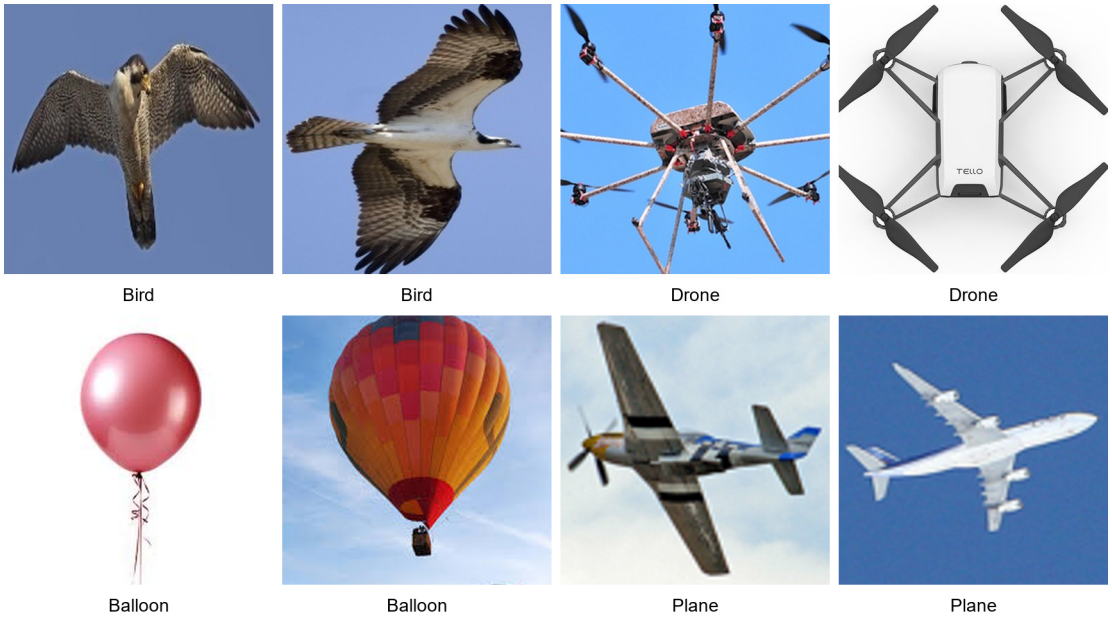


Figure 1. Dataset example of known and unknown objects.

All models are trained exclusively on the known classes using identical training and validation splits to ensure a fair and consistent comparison across architectures. To simulate realistic deployment scenarios with increasing levels of difficulty, we design three evaluation settings:

(1) Closed-set setting: only samples from the known classes (Bird and Drone) are included in the test set.
(2) Open-set setting with a single unknown class: one unknown category is introduced at test time, with two configurations where the unknown class corresponds to either Plane or Balloon.
(3) Open-set setting with multiple unknown classes: both Plane and Balloon are included in the test set simultaneously.

In all open-set settings, unknown samples are strictly excluded from both training and validation and are used exclusively for testing, consistent with the open-set evaluation protocol [7]. To enable a fair and balanced evaluation of unknown rejection performance, the test set is augmented with a controlled number of unknown samples. Specifically, in the single-unknown-class setting, 200 images were selected from each unknown class (Plane or Balloon) for evaluation. In the dual-unknown-class setting, a balanced unknown test set consisting of

200 images was constructed by randomly sampling 100 images each from the Plane and Balloon datasets and combining them. The data distribution for the known classes is summarized in Table 1.

Table 1. Dataset split for known objects (bird and drone).

Class	Training	Validation	Testing
Bird	685	200	200
Drone	685	200	200
Total	1370	400	400

In this study, we adopt the standard terminology used in open-domain recognition. In-domain (ID) categories refer to object categories that have already appeared during the training phase and belong to the model's known label space, specifically including "Bird" and "Drone". Out-of-distribution (OOD) samples, also known as unknown categories, refer to inputs that appear during the testing phase but do not belong to any known category in the training label space; these can be regarded as objects the model has not encountered before, such as new types of aerial targets that may appear in real-world deployments. Unlike the closed-set setting, where all test samples are assumed to belong to the set of known classes by default, in open-set scenarios, the model must also determine whether a test input should be classified into a known class or rejected as an unknown sample. To investigate the impact of different types of unknown targets, this paper defines three OOD conditions: (i) Near-OOD unknown class "Plane", which, due to its shared characteristic of being a "flying object", is semantically and structurally closer to "Drone" and "Bird"; (ii) Far-OOD unknown class "Balloon", whose appearance and structure differ more markedly from known categories; and (iii) a mixed unknown setting, which includes samples from both the 'Plane' and "Balloon" classes. For all open-set evaluations, this paper employs a threshold decision mechanism based on the maximum softmax probability (MSP): when the maximum category probability of a test sample falls below a set threshold, it is classified as Unknown; otherwise, it is assigned to the Bird or Drone category.

3.3. Model Architectures

This paper evaluates several representative convolutional neural networks (CNNs) and Transformer models for the bird–drone recognition task. All models were initialized using pre-trained weights from ImageNet [23], and their final classification heads were replaced with output layers corresponding to the target categories, while the backbone network structures remained unchanged. To conduct a comprehensive comparison of backbone networks commonly used in image processing, this study selected four convolutional neural network (CNN) models (ResNet50, VGG19, DenseNet121, and EfficientNet_B0) and four Transformer models (ViT_B16, ViT_B32, Swin_Tiny, and Swin_Small). These models represent various mainstream architectural approaches in modern image recognition, including residual learning, stacked deep convolutions, dense connections, combined scaling, global self-attention, and hierarchical window attention.

Among Convolutional Neural Networks (CNNs), ResNet50 is a widely used residual network that demonstrates good optimization stability in deeper architectures [10]. As a classic deep CNN foundation model, VGG19 adopts a relatively simple stacked convolutional structure [11]. DenseNet121 improves feature reuse through dense connections and facilitates inter-layer information transfer [12]. EfficientNet_B0, based on a three-dimensional combined scaling strategy involving depth, width, and resolution, has been selected as a lightweight yet competitive CNN baseline model [13].

Among Transformer models, ViT_B16 and ViT_B32 represent standard visual Transformer architectures with different patch sizes, which can be used to investigate the impact of token granularity on the fine-grained bird-drone recognition task [14]. Swin_Tiny and Swin_Small belong to the category of hierarchical visual transformers, which better model local information and multi-scale structures through a sliding-window attention mechanism [15]. Overall, these eight models constitute a diverse and relatively balanced collection of architectures, suitable for analyzing performance differences among various models when handling unknown aerial targets under closed-class and open-class classification conditions. Table 2 summarizes the parameter scale, depth definition, and corresponding references for each model.

Table 2. An overview of the evaluated model architectures.

Model	Family	Parameters (M)	Network Depth	Ref.
ResNet50	CNN	~25.6	50 layers	[10]
VGG19	CNN	~143.7	19 layers	[11]
DenseNet121	CNN	~8.0	121 layers	[12]
EfficientNet_B0	CNN	~5.0	B0	[13]
ViT_B16	Transformer	~86.6	12 blocks	[14]
ViT_B32	Transformer	~88.2	12 blocks	[14]
Swin_Tiny	Transformer	~28.3	2, 2, 6, 2	[15]
Swin_Small	Transformer	~49.6	2, 2, 18, 2	[15]

Network depth is reported based on the original definitions used in the relevant reference papers. For EfficientNet_B0, the model configuration name is reported rather than the exact number of layers; for variants of Swin Transformer, the block configurations for each stage are provided.

3.4. Training and Evaluation

All models were evaluated under a unified training and evaluation protocol to ensure that comparisons between different architectures were based on as consistent conditions as possible. Specifically, all models used the same training, validation, and test data splits, the same ImageNet pre-training initialization strategy [23], the same input resolution and data augmentation process, as well as consistent optimizer settings, batch size, number of training epochs, and model selection criteria. Under these conditions, performance differences between models can be attributed more to the architecture itself rather than variations in training conditions.

All input images are in RGB format and uniformly resized to 224 × 224. During the training phase, standard data augmentation methods such as random scaling, cropping, and horizontal flipping are applied, followed by normalization based on the mean and standard deviation of ImageNet. During the validation and testing phases, images are scaled and center-cropped to ensure that all models receive inputs of consistent dimensions.

Optimization is performed using stochastic gradient descent (SGD) with a learning rate of 0.001 and momentum of 0.9. All models are trained for 15 epochs with a batch size of 16 using the cross-entropy loss function. Model selection is based solely on validation accuracy, and the parameters that achieve the best validation performance are retained for subsequent evaluation. No samples from the test set are used during training or model selection.

For closed-set inference, each test sample is assigned to the known class with the highest posterior probability from the softmax output. In this setting, all test samples are assumed to belong to one of the known categories; therefore, no rejection mechanism is applied.

For open-set inference, we adopt a confidence-based rejection strategy based on the maximum softmax probability (MSP) [8]. Given an input image x, the model first outputs logits for the K known classes, which are converted into posterior probabilities through the softmax function, as shown in Equation (1).

$$p(y = i \mid x) = \frac{\exp(z_i(x))}{\sum_{j=1}^{K} \exp(z_j(x))}, \quad i \in \{1,2,\dots,K\} \tag{1}$$

where $z_i(x)$ denotes the logit of class i for input x, and K is the number of known classes. In this work, $K = 2$, corresponding to Bird and Drone.

The maximum softmax probability is then defined as Equation (2).

$$\mathrm{MSP}(x) = \max_i p(y = i \mid x) \tag{2}$$

Based on this score, the final open-set decision is made according to Equation (3).

$$\hat{y}(x) = \begin{cases} \text{Unknown}, & \text{if } \mathrm{MSP}(x) < \tau \\ \arg\max_y p(y \mid x), & \text{if } \mathrm{MSP}(x) \geq \tau \end{cases} \tag{3}$$

where τ is the decision threshold. In this study, the threshold is fixed at $\tau = 0.8$ and applied consistently across all models and open-set settings to preserve comparability under a shared decision rule. The rationale for this choice is further examined in Section 4.2 through threshold-sensitivity analysis. Model performance is evaluated using standard classification metrics, as described in Section 4.2. Confusion matrices are also used in the results section to provide additional insight into model prediction behavior, especially in open-set scenarios.

4. Experiments and Analysis

4.1. Experimental Setup

All experiments were conducted on a local workstation running Windows 11. The models were implemented in Python using the PyTorch framework and the torchvision library, and were trained and evaluated on a CUDA-enabled NVIDIA RTX 4070 Ti Super GPU (NVIDIA Corporation, Santa Clara, CA, USA). This hardware and software environment provided the necessary computational support for the deep neural network experiments.

To ensure the reproducibility of results and consistency in comparisons between different models, all training and evaluation were conducted in a standardized experimental environment using the same codebase and software configuration. Throughout the training and validation process, the test data was not used for model training or selection.

To further support the validation of the results, this paper plans to release the relevant code and experimental materials, including data preprocessing scripts, training and evaluation scripts, data partitioning lists for each experiment, and the training weights of the models reported in the paper. These materials will also assist future research in conducting comparisons and reproducing results under the same experimental setup.

4.2. Evaluation Metrics

Model performance is evaluated using both standard classification metrics and score-based open-set detection metrics. Following common practices in classification research [24,25], this paper employs precision, recall, F1 score, and overall accuracy as basic evaluation metrics. Precision represents the proportion of samples predicted to belong to a certain class that are actually correct, recall represents the proportion of true samples in that class that are successfully identified, the F1 score is used to comprehensively reflect the balance between precision and recall, and overall accuracy represents the proportion of samples correctly classified in the entire test set.

In a closed-set scenario, accuracy serves as the primary metric for evaluating overall recognition performance, while precision, recall, and the F1 score provide supplementary category-level information. In an open-set scenario, these metrics must be analyzed in conjunction with the confusion matrix to determine whether unknown samples are correctly rejected or incorrectly classified into known categories. Since open-set evaluation involves both maintaining recognition capabilities for known classes and effectively rejecting unknown samples, a single metric cannot fully reflect the model's behavior.

Based on this consideration, in the open-set threshold sensitivity analysis, this paper further focuses on known-class accuracy, unknown-class recall, overall accuracy, and Macro F1. The known-class accuracy is used to measure whether the model's ability to correctly identify samples from the two known classes, Bird and Drone, is maintained after the introduction of threshold-based decision-making; the unknown-class recall is used to measure the proportion of all true unknown samples that are correctly classified as Unknown by the model. The overall accuracy summarizes the model's performance across the entire test set, while Macro F1 provides a more balanced evaluation perspective across different classes.

To verify the validity of the selected threshold under different data set conditions, this paper conducted threshold scans under three settings: Plane-only, Balloon-only, and mixed unknown. As shown in Figure 2, all three scenarios exhibit a consistent trade-off: as the MSP threshold increases, the recall rate for the unknown class generally improves, but the accuracy for known classes correspondingly decreases. Among these, the Plane scenario is the most challenging, the Balloon scenario is the easiest to distinguish, and the mixed unknown scenario exhibits a trend intermediate between the two. Based on this observation, this paper adopts $\tau = 0.8$ as the unified operating point, as this threshold achieves a relatively reasonable balance between maintaining known-class performance and enhancing the ability to reject unknown samples, while avoiding the need to adjust thresholds separately for different models.

To complement the fixed-threshold evaluation, score-based detection metrics are also reported, including AUROC, AUPR, and FPR95. These metrics are computed from the confidence score used to distinguish known and unknown samples and provide a threshold-independent view of detection quality. AUROC summarizes the separability between known and unknown samples across all possible thresholds [26]. AUPR complements AUROC by emphasizing precision-recall behavior and is especially informative when retrieval quality for the positive class is of primary interest [27]. FPR95 reports the false positive rate when the true positive rate is fixed at 95%, and is widely used in out-of-distribution detection benchmarks to reflect how often unwanted samples are still accepted under a high-recall operating condition [17]. In this way, the fixed-threshold results at $\tau = 0.8$ reflect model behavior under a shared operating point, whereas AUROC, AUPR, and FPR95 evaluate the intrinsic score-based separability between known and unknown samples without relying on a single threshold.

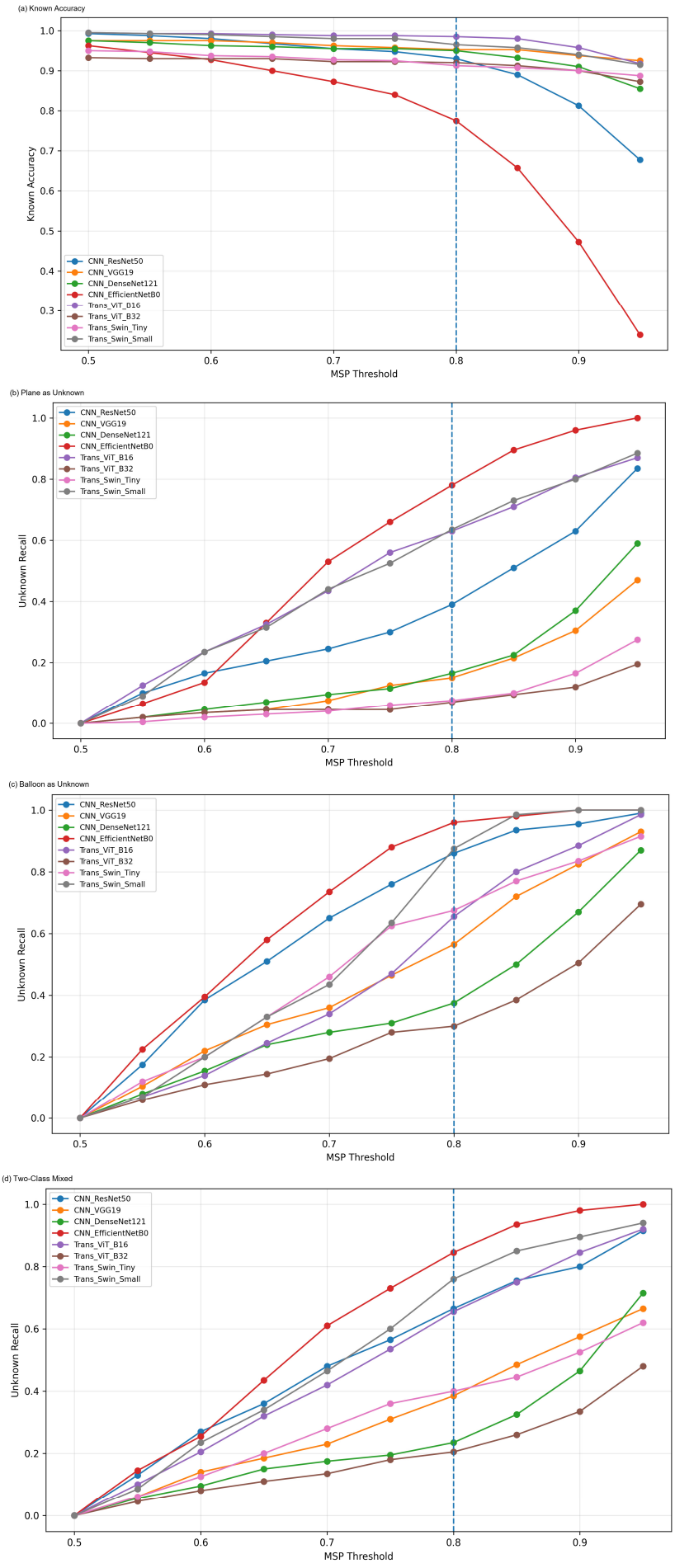


Figure 2. Threshold-sensitivity analysis under open-set evaluation.

The mathematical definitions of the standard classification metrics are summarized in Table 3.

Table 3. Evaluation metrics.

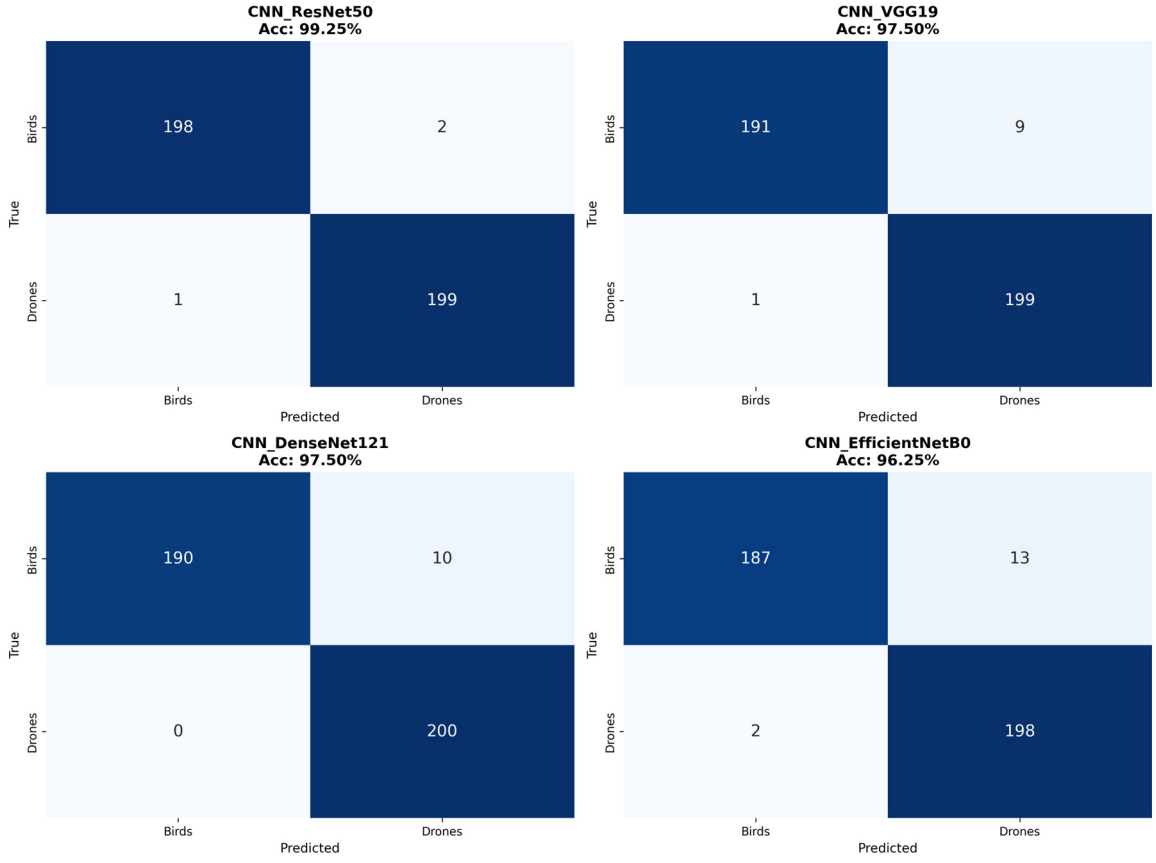
Assessments	Formula	Assessments	Formula
Precision(P)	$\frac{TP}{TP + FP}$	Recall(R)	$\frac{TP}{TP + FN}$
F1 Score	$2 \times \frac{P \times R}{P + R}$	Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$

4.3. Closed-Set Results

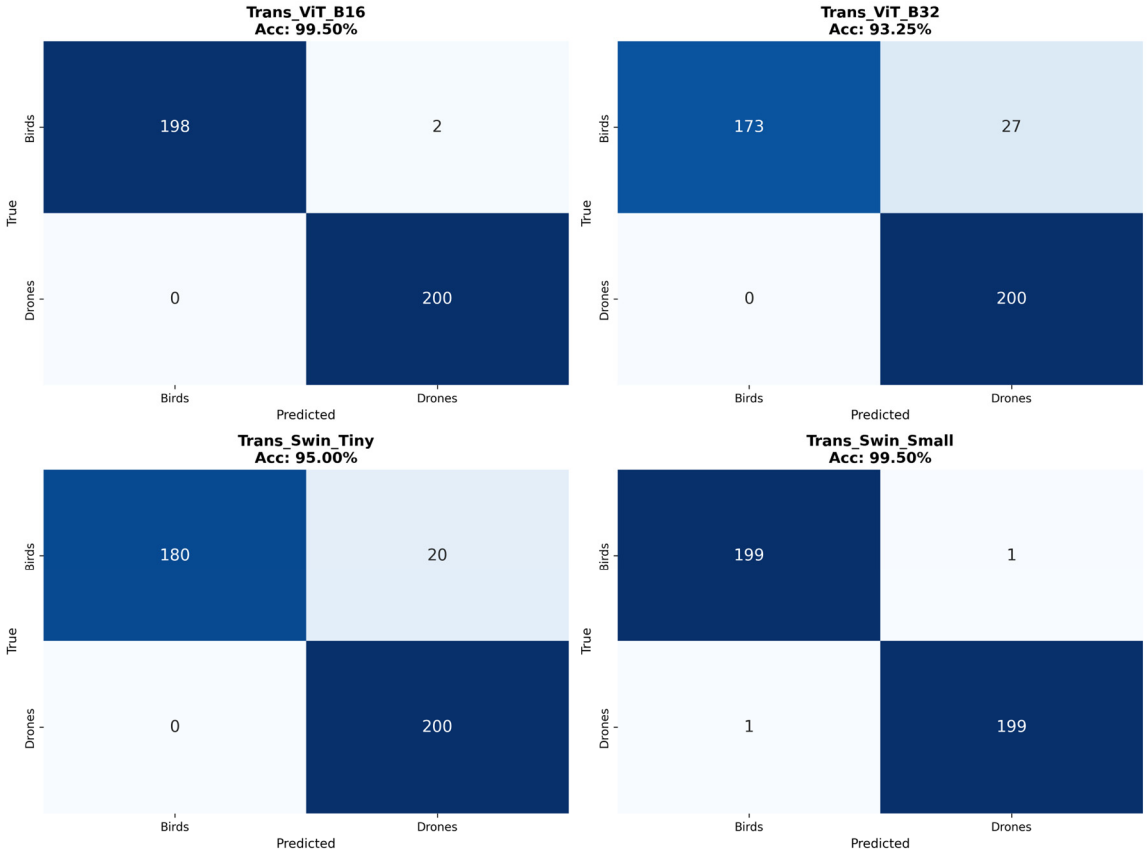
This consistency in results indicates that the models' predictive performance under closed-set conditions is generally well-balanced, and also suggests that the current dataset does not introduce any significant class bias.

The closed-set classification performance of all architectures is summarized in Table 4. Among all models, Swin_Small and ViT_B16 performed the best, with both achieving an accuracy of 99.50% and Macro-F1 scores close to perfect. ResNet50 followed closely with 99.25%, indicating that even amid the growing prevalence of Transformer models, mature convolutional neural network architectures remain highly competitive. Both VGG19 and DenseNet121 achieved an accuracy of 97.50%, while EfficientNet_B0 reached 96.25%, indicating that these CNN models are also capable of maintaining a high level of classification performance in the current task. In contrast, ViT_B32 performed the worst, with an accuracy of 93.25%. This may suggest that larger patch sizes weaken the model's ability to perceive fine-grained structural information, which is particularly crucial for distinguishing birds from drones.

As shown in Figure 3, the confusion matrix further reveals a distinct asymmetric misclassification pattern: instances where "Bird" is misclassified as "Drone" are more common, while instances where 'Drone' is misclassified as "Bird" are relatively rare. This trend is observable across multiple architectures, particularly in VGG19, DenseNet121, EfficientNet_B0, ViT_B32, and Swin_Tiny. This suggests that bird samples may exhibit greater intra-class variability, which is likely related to variations in pose, scale, and background conditions, making them relatively more challenging to model. Overall, all models demonstrated strong and relatively stable recognition capabilities in the closed-set task, providing a reliable baseline for subsequent open-set evaluation; in contrast, under open-set conditions, the models will face the stricter requirement of effectively rejecting unknown samples.



(**a**)



(**b**)

Figure 3. (**a**) Closed-set confusion matrices of CNN models. (**b**) Closed-set confusion matrices of Transformer models.

Table 4. Closed-set classification performance across architecture.

Model	Category	F1 Score (Macro)	Accuracy
ResNet50	CNN	0.992	99.25%
VGG19	CNN	0.974	97.50%
DenseNet121	CNN	0.974	97.50%
EfficientNet_B0	CNN	0.962	96.25%
ViT_B16	Transformer	0.994	99.50%
ViT_B32	Transformer	0.932	93.25%
Swin_Tiny	Transformer	0.949	95.00%
Swin_Small	Transformer	0.995	99.50%

4.4. Open-Set Results

In real-world deployments, recognition systems may encounter aerial targets that were not present during training; therefore, it is necessary to further evaluate the model within an open-domain recognition framework. Under this framework, the model must not only classify known categories but also minimize the misclassification of unknown inputs into known categories [7]. This paper evaluates the model's open-domain performance under three test-phase unknown-class settings: (i) aircraft as the unknown class (near-OOD), (ii) balloons as the unknown class (far-OOD), and (iii) both aircraft and balloons as unknown classes. For each setting, Tables 5–10 report the precision, recall, F1 score, and overall test accuracy for each category, while Figures 4–6 present the corresponding confusion matrices to analyze how the model handles unknown samples. Detailed results are presented in Sections

4.4.1. Plane as Unknown (Near-OOD)

When "Plane" was introduced as the sole unknown class, the model's open-set performance declined significantly compared to the closed-set baseline, indicating that effectively rejecting unknown samples that are semantically similar to known classes poses a significant challenge. As shown in Table 5, ViT_B16 achieved the highest test accuracy (86.67%), with Precision, Recall, and F1 scores for the Unknown class of 0.969, 0.630, and 0.764, respectively, demonstrating a balanced performance in both known-class recognition and unknown-class

rejection. Swin_Small's results are comparable, with a test accuracy of 85.50% and Precision, Recall, and F1 score for the Unknown class at 0.901, 0.635, and 0.745, respectively, indicating that this model can reject a significant portion of aircraft samples while maintaining stable recognition of known categories.

Table 5. Performance of different models under one unknown class (Plane).

Models	Category	Precision	Recall	F1 score	Testing Accuracy (%)
ResNet50	Birds	1	0.895	0.945	75.00
	Drones	0.613	0.965	0.75	
	Unknown	0.736	0.39	0.51	
VGG19	Birds	0.995	0.92	0.956	68.50
	Drones	0.537	0.985	0.695	
	Unknown	0.625	0.15	0.242	
DenseNet121	Birds	0.994	0.9	0.945	68.83
	Drones	0.543	1	0.704	
	Unknown	0.647	0.165	0.263	
EfficientNet_B0	Birds	0.993	0.675	0.804	77.67
	Drones	0.799	0.875	0.835	
	Unknown	0.637	0.78	0.701	
ViT_B16	Birds	0.995	0.97	0.982	86.67
	Drones	0.727	1	0.842	
	Unknown	0.969	0.63	0.764	
ViT_B32	Birds	0.96	0.84	0.896	63.67
	Drones	0.509	1	0.675	
	Unknown	0.438	0.07	0.121	
Swin_Tiny	Birds	0.976	0.825	0.894	63.33
	Drones	0.504	1	0.67	
	Unknown	0.441	0.075	0.128	
Swin_Small	Birds	0.985	0.955	0.97	85.50
	Drones	0.736	0.975	0.839	
	Unknown	0.901	0.635	0.745	

Under this near-OOD setting, the performance of the other Transformer models declined more significantly. ViT_B32 achieved a test accuracy of only 63.67%, with a recall and F1 score for the Unknown class of just 0.070 and 0.121, respectively; Swin_Tiny exhibited a similar pattern, with a test accuracy of 63.33%, a recall of 0.075 for the Unknown class, and an F1 score of 0.128. These results indicate that a large number of aircraft samples were not correctly classified as Unknown but were instead misclassified into known categories.

The performance of CNN models under this configuration was more varied. EfficientNet_B0 demonstrated moderate robustness, with a test accuracy of 77.67%, a recall of 0.780 for the "Unknown" class, and an F1 score of 0.701, indicating that, compared to most weaker models, it was more likely to classify aircraft samples as "Unknown". In contrast, VGG19 and DenseNet121 demonstrated weaker ability to reject aircraft samples: VGG19 achieved a test accuracy of 68.50%, a recall of 0.150 for the "Unknown" class, and an F1 score of 0.242; DenseNet121 has a test accuracy of 68.83%, a recall rate of 0.165 for the "Unknown" class, and an F1 score of 0.263. Even for ResNet50, although the precision for the "Unknown" class is relatively high (0.736), the recall remains limited (0.390), and the overall test accuracy drops to 75.00%. This indicates that a significant number of aircraft samples are still misclassified as known targets and are not successfully rejected.

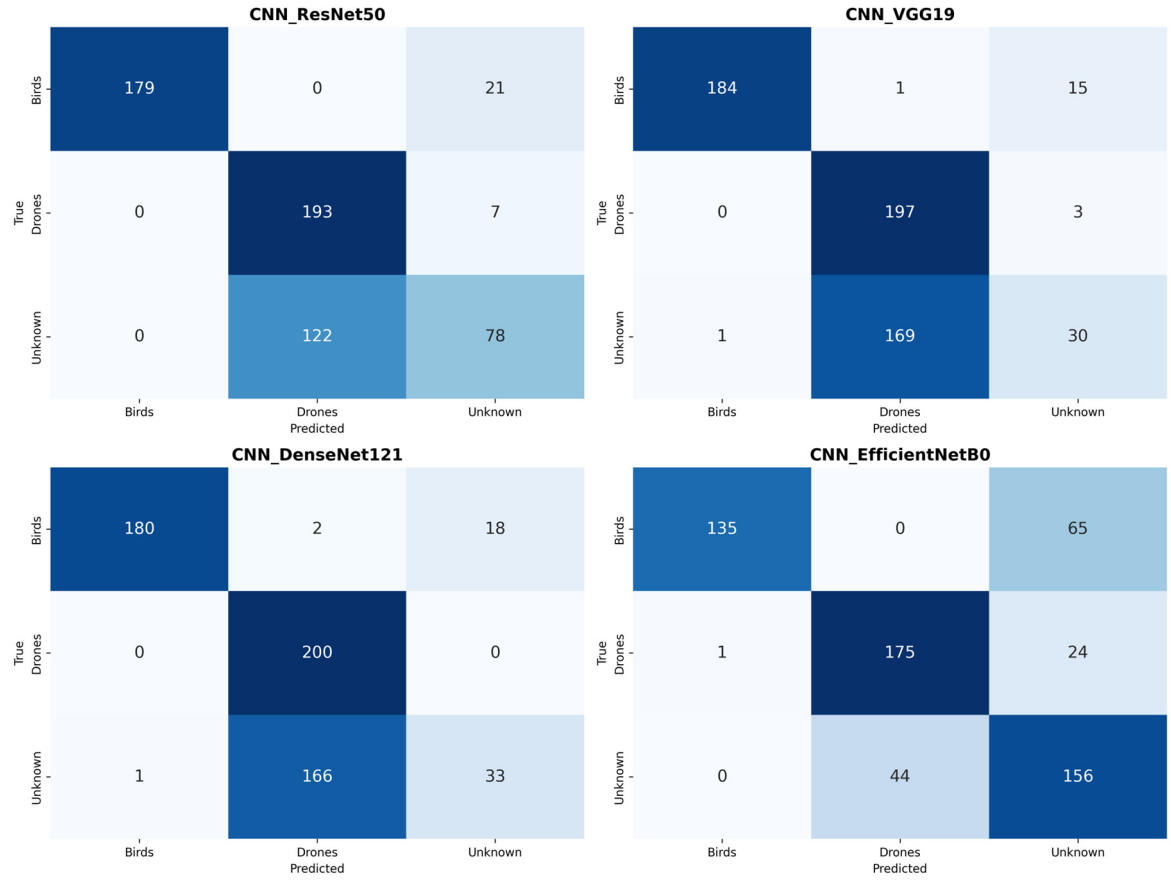
To supplement the results under fixed thresholds, this paper further reports threshold-independent detection metrics under the "Plane-as-Unknown" setting, including AUROC, AUPR, and FPR95, as shown in Table 6. These results are largely consistent with the fixed-threshold evaluation and further demonstrate significant differences among architectures in the distinguishability of known and unknown samples based on scores. ViT_B16 demonstrated the strongest overall detection performance, achieving AUROC and AUPR values of 0.9647 and 0.9318, respectively, indicating its ability to most clearly distinguish between known samples and near-OOD aircraft samples. Swin_Small also performed notably well, with an AUROC of 0.9596 and an AUPR of 0.8941, while achieving the lowest FPR95 (0.1775), indicating that this model exhibits greater stability under conditions requiring high recall. Among the CNN models, VGG19 demonstrated the strongest threshold-independent detection capability, with an AUROC of 0.9067, an AUPR of 0.7229, and an FPR95 of 0.2000; in contrast, ResNet50, DenseNet121, and EfficientNet_B0 showed significantly weaker performance in score-based discrimination. ViT_B32 and Swin_Tiny have the lowest AUROC and AUPR, while their FPR95 is relatively high, further indicating that these two models face the greatest difficulty in distinguishing aircraft samples that are

semantically close to known categories. Overall, the results from the score-based metrics align with those from fixed-threshold evaluations: the Plane setting is a highly challenging near-OOD scenario, and strong closed-set classification performance does not necessarily imply that a model can reliably handle unknown samples.

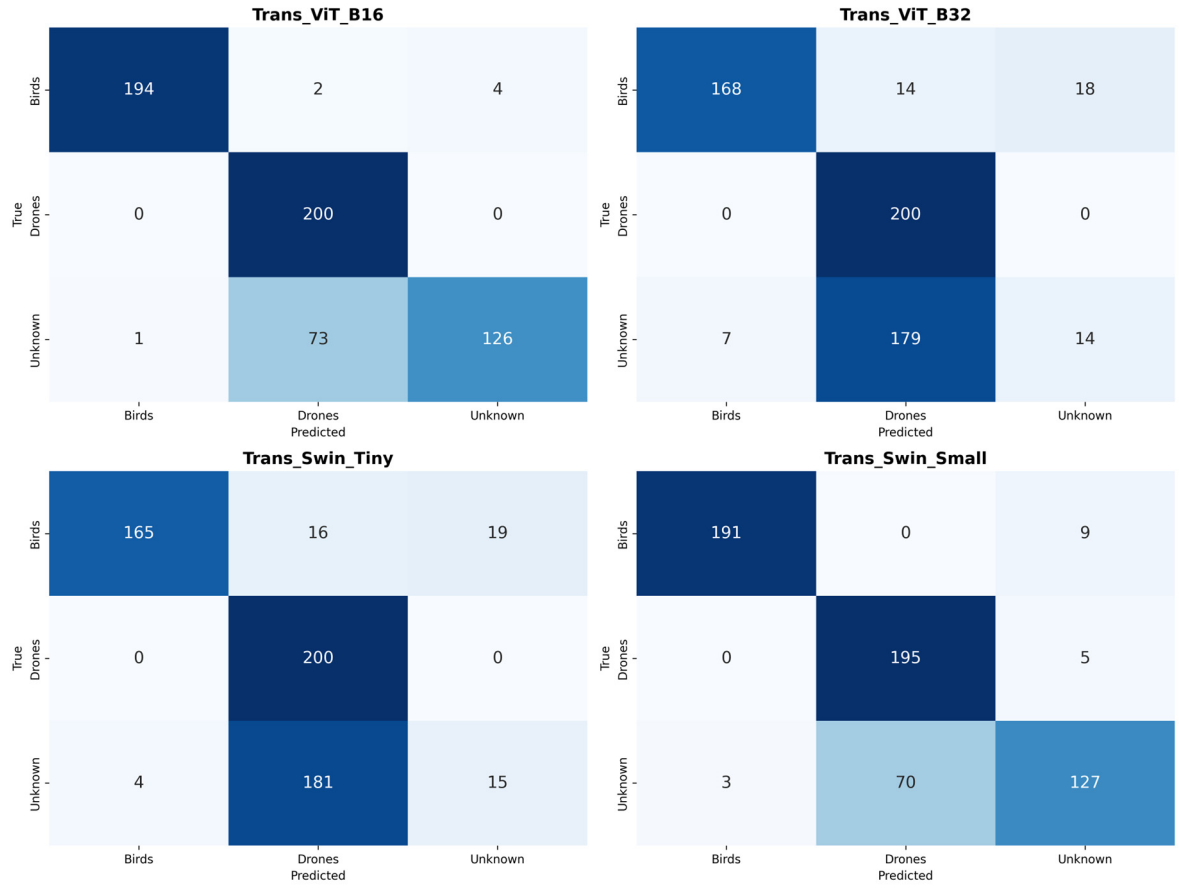
Table 6. Score-based open-set detection performance under one unknown class (Plane).

Model	AUROC	AUPR	FPR95
ResNet50	0.8309	0.6958	0.5675
VGG19	0.9067	0.7229	0.2
DenseNet121	0.8463	0.6326	0.38
EfficientNet_B0	0.836	0.6116	0.435
ViT_B16	0.9647	0.9318	0.21
ViT_B32	0.779	0.5295	0.5375
Swin_Tiny	0.7612	0.5292	0.555
Swin_Small	0.9596	0.8941	0.1775

The confusion matrices in Figure 4a,b further reveal a consistent failure pattern in the Plane-as-Unknown setting: unknown aircraft samples are more often predicted as Drones than as Birds, rather than being correctly rejected as Unknown. For example, in ResNet50, 122 plane samples are classified as Drones, whereas only 78 are rejected as Unknown. Similar behavior is also observed in VGG19, DenseNet121, ViT_B32, and Swin_Tiny. This suggests that, under near-OOD conditions, visually similar flying-object features systematically pull unknown aircraft toward the Drone class.



(a)



(**b**)

Figure 4. (**a**) Confusion matrices of CNN models in the Plane-as-Unknown setting. (**b**) Confusion matrices of Transformer models in the Plane-as-Unknown setting.

4.4.2. Balloon as Unknown (Far-OOD)

When Balloon is treated as the only unknown class, overall open-set performance improves compared with the near-OOD aircraft setting, indicating that visually distinctive unknowns are easier to reject under the same softmax-based thresholding rule. As reported in Table 7, Swin_Small achieves the best overall robustness with the highest testing accuracy (93.50%) and strong unknown-class precision/recall/F1 of 0.926/0.875/0.900, showing that most balloon samples are correctly rejected while known-class recognition remains stable.

Models in the second tier also demonstrated good stability. ResNet50 achieved a test accuracy of 90.67%, with Precision, Recall, and F1 scores of 0.860 for the "Unknown" class. This indicates that when balloons are classified as the "Unknown" class, the model exhibits a balanced performance in identifying unknown samples, showing neither significant under-recall nor a large number of unknown samples being incorrectly classified into known categories. ViT_B16 achieved a test accuracy of 87.50%, with a high Precision of 0.970 for the "Unknown" class, but a low Recall of 0.655 and an F1 score of 0.782. This suggests that once this model classifies a sample as "Unknown", it is usually correct; however, some balloon samples were not successfully rejected and were instead classified into known categories.

When focusing further on the trade-off between precision and recall for the "Unknown" class, the differences among models become more pronounced. EfficientNet_B0 achieved the highest recall rate for the "Unknown" class (0.960), but its precision was relatively low (0.683), resulting in an F1 score of 0.798. This suggests that under this configuration, the model tends to classify samples as "Unknown", thereby identifying most balloon samples, but at the cost of misclassifying more samples from known classes as "Unknown". In contrast, Swin_Tiny exhibits a more conservative behavior, with a test accuracy of 83.33% and Precision, Recall, and F1 scores for the "Unknown" class of 0.877, 0.675, and 0.763, respectively. VGG19 achieved a test accuracy of 82.33%, but its Recall for the Unknown class dropped to 0.565, with an F1 score of 0.683, indicating that a significant portion of balloon samples were still incorrectly classified into known categories.

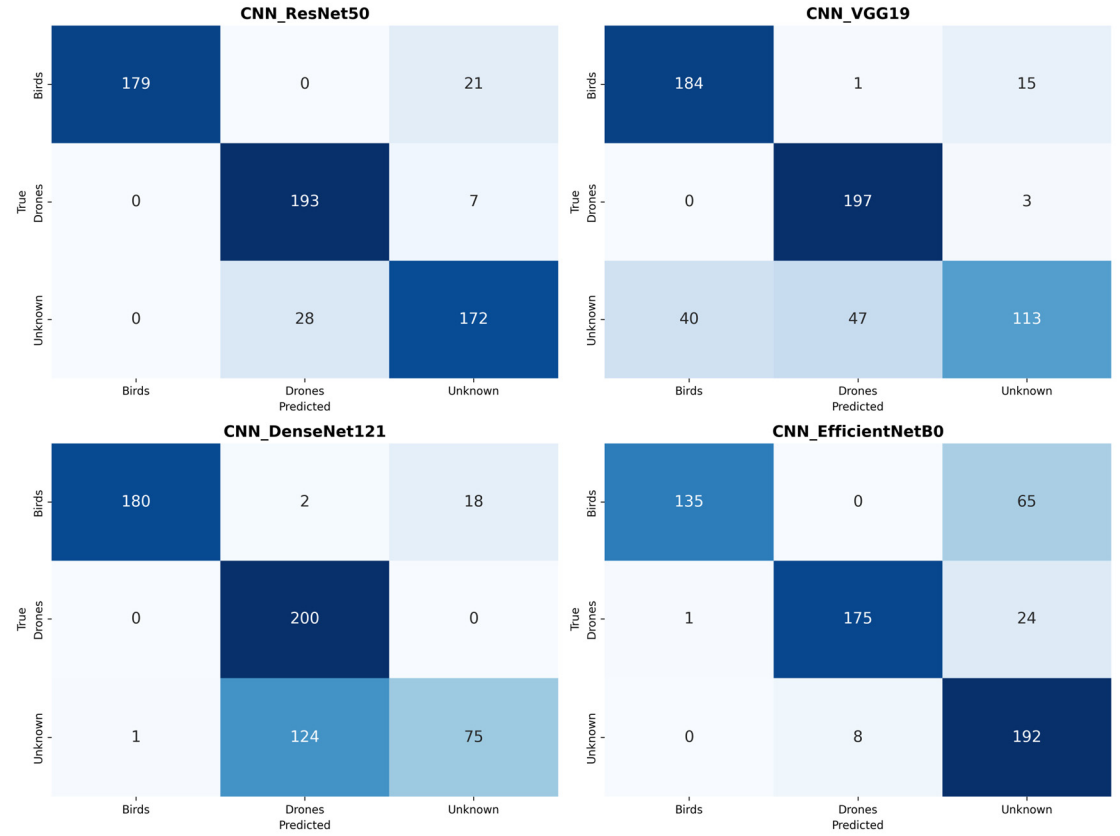
Table 7. Performance of different models under one unknown class (Balloon).

Models	Category	Precision	Recall	F1 score	Testing Accuracy (%)
ResNet50	Birds	1	0.895	0.945	90.67
	Drones	0.873	0.965	0.917	
	Unknown	0.86	0.86	0.86	
VGG19	Birds	0.821	0.92	0.868	82.33
	Drones	0.804	0.985	0.885	
	Unknown	0.863	0.565	0.683	
DenseNet121	Birds	0.994	0.9	0.945	75.83
	Drones	0.613	1	0.76	
	Unknown	0.806	0.375	0.512	
EfficientNet_B0	Birds	0.993	0.675	0.804	83.67
	Drones	0.956	0.875	0.914	
	Unknown	0.683	0.96	0.798	
ViT_B16	Birds	0.951	0.97	0.96	87.50
	Drones	0.766	1	0.868	
	Unknown	0.97	0.655	0.782	
ViT_B32	Birds	0.87	0.84	0.855	71.33
	Drones	0.608	1	0.756	
	Unknown	0.769	0.3	0.432	
Swin_Tiny	Birds	0.954	0.825	0.885	83.33
	Drones	0.733	1	0.846	
	Unknown	0.877	0.675	0.763	
Swin_Small	Birds	0.905	0.955	0.929	93.50
	Drones	0.975	0.975	0.975	
	Unknown	0.926	0.875	0.9	

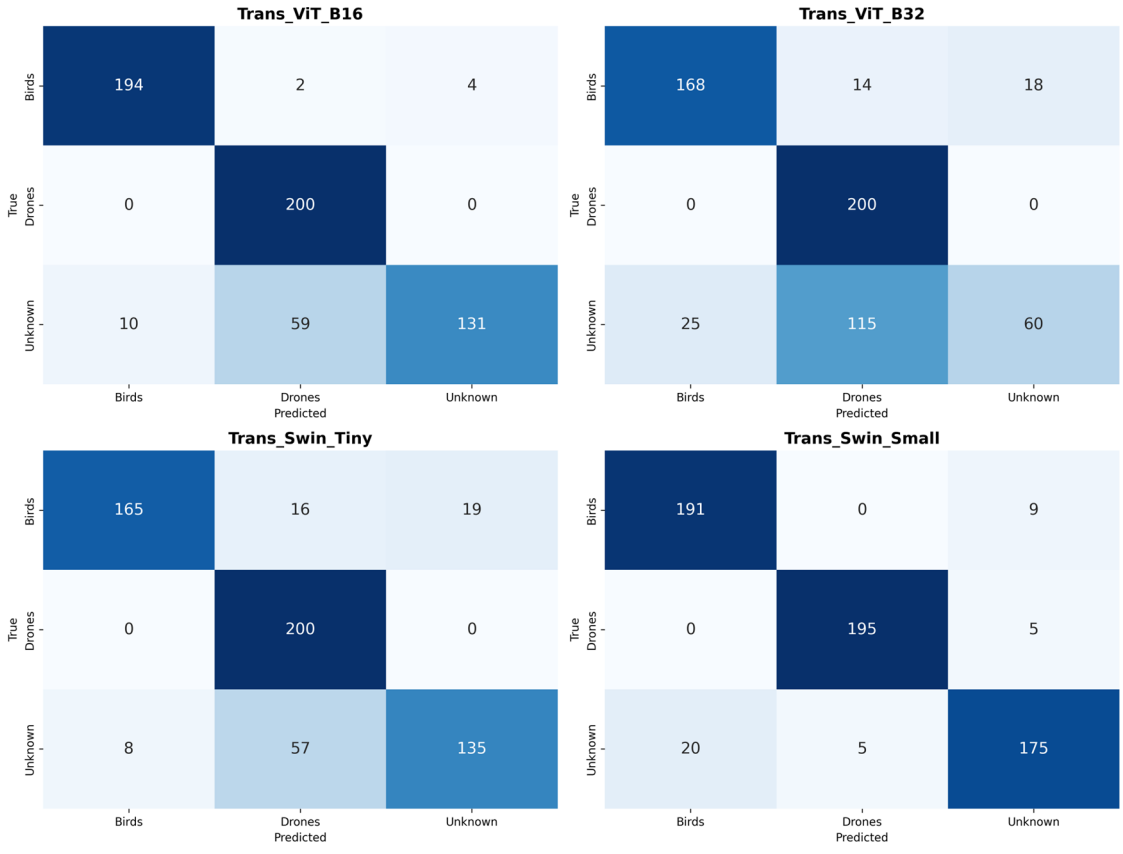
Despite the larger distribution shift, several models remain vulnerable even in this far-OOD setting. ViT_B32 reaches only 71.33% accuracy with an unknown recall of 0.300 and an unknown F1 of 0.432, and DenseNet121 achieves 75.83% accuracy with an unknown recall of 0.375 and an unknown F1 of 0.512, indicating that many balloon samples are not rejected and are instead forced into known predictions.

A similar pattern can also be observed from the threshold-independent detection metrics, but the separation between known and unknown samples is noticeably stronger in the Balloon setting. As summarized in Table 8, ViT_B16 again achieves the highest AUROC (0.9849) and AUPR (0.9557), indicating the clearest score-based distinction between known classes and far-OOD balloon inputs. Swin_Small performs comparably and attains the lowest FPR95 (0.0400), together with a high AUROC of 0.9808 and an AUPR of 0.9277, suggesting particularly strong robustness when the operating condition requires high recall. Among CNN models, VGG19 shows the strongest threshold-independent detection performance, with an AUROC of 0.9618, an AUPR of 0.8889, and an FPR95 of 0.0800, while ResNet50 also remains competitive with an AUROC of 0.9568 and an AUPR of 0.8991. In contrast, ViT_B32, DenseNet121, and EfficientNet_B0 achieve lower AUROC and AUPR values together with relatively higher FPR95, indicating weaker score-based separability. Overall, these results are consistent with the fixed-threshold evaluation and further confirm that the Balloon setting is easier than the Plane setting for unknown rejection, since visually distinct far-OOD samples are more readily separated from the known classes.

As shown in Figure 5a,b, rejection performance is clearly stronger than that observed in the Plane setting. Since balloon samples are visually more distinct from the known Bird and Drone classes, several models are able to assign a larger proportion of unknown inputs to the Unknown category. This pattern is particularly visible in ResNet50, EfficientNet_B0, ViT_B16, Swin_Tiny, and Swin_Small, whereas VGG19, DenseNet121, and ViT_B32 still misclassify a noticeable portion of balloon samples as known classes. Among all models, Swin_Small achieves one of the best overall balances between known-class recognition and unknown rejection. These results indicate that visually distant unknown objects are more separable from the known-class space under the same threshold.



(a)



(b)

Figure 5. (**a**) Confusion matrices of CNN models in the Balloon-as-Unknown setting. (**b**) Confusion matrices of Transformer models in the Balloon-as-Unknown setting.

Table 8. Score-based open-set detection performance under one unknown class (Balloon).

Model	AUROC	AUPR	FPR95
ResNet50	0.9568	0.8991	0.16
VGG19	0.9618	0.8889	0.08
DenseNet121	0.9195	0.7755	0.205
EfficientNet_B0	0.9087	0.7493	0.195
ViT_B16	0.9849	0.9557	0.0575
ViT_B32	0.9054	0.7499	0.2325
Swin_Tiny	0.9578	0.8802	0.1225
Swin_Small	0.9808	0.9277	0.04

4.4.3. Both Plane and Balloon as Unknown

When both airplanes and balloons are introduced as unknown classes, the model's overall performance declines compared to when only balloons are designated as unknown. This is because the mixed unknown scenario contains both near-OOD airplane samples and far-OOD balloon samples; the former are semantically closer to known classes and are therefore more difficult to correctly reject. As shown in Table 9, Swin_Small remains the most robust model overall, achieving the highest test accuracy (89.67%). Its Precision, Recall, and F1 score for the unknown category reached 0.916, 0.760, and 0.831, respectively, indicating that the model can correctly identify a significant portion of mixed unknown samples while maintaining stability in recognizing known categories.

The models in the second tier exhibit moderate robustness, with their primary issue being that a certain proportion of unknown samples are still misclassified into the known category space. ViT_B16 achieved a test accuracy of 87.50%, with Precision, Recall, and F1 score for the Unknown class at 0.970, 0.655, and 0.782, respectively. This indicates that the model's predictions for the Unknown class remain highly reliable, though its coverage of unknown samples is only moderate. ResNet50 achieved a test accuracy of 84.17%, with Precision, Recall, and F1 score for the Unknown class at 0.826, 0.665, and 0.737, respectively. This indicates that its ability to handle unknown inputs is weaker than that of Swin_Small, and a significant number of unknown samples are still incorrectly classified into known categories.

Table 9. Performance of different models under two unknown classes (Plane + Balloon).

Models	Category	Precision	Recall	F1 score	Testing Accuracy (%)
ResNet50	Birds	1	0.895	0.945	84.17
	Drones	0.742	0.965	0.839	
	Unknown	0.826	0.665	0.737	
VGG19	Birds	0.92	0.92	0.92	76.33
	Drones	0.646	0.985	0.78	
	Unknown	0.811	0.385	0.522	
DenseNet121	Birds	0.994	0.9	0.945	71.17
	Drones	0.565	1	0.722	
	Unknown	0.723	0.235	0.355	
EfficientNet_B0	Birds	0.993	0.675	0.804	79.83
	Drones	0.85	0.875	0.862	
	Unknown	0.655	0.845	0.738	
ViT_B16	Birds	0.99	0.97	0.98	87.50
	Drones	0.743	1	0.853	
	Unknown	0.97	0.655	0.782	
ViT_B32	Birds	0.918	0.84	0.877	68.17
	Drones	0.559	1	0.717	
	Unknown	0.695	0.205	0.317	
Swin_Tiny	Birds	0.971	0.825	0.892	74.17
	Drones	0.604	1	0.753	
	Unknown	0.808	0.4	0.535	
Swin_Small	Birds	0.936	0.955	0.946	89.67
	Drones	0.848	0.975	0.907	
	Unknown	0.916	0.76	0.831	

When further examining the trade-off between precision and recall for the "Unknown" class, the differences among models become more pronounced. EfficientNet_B0 still exhibits a recall-biased pattern in mixed unknown scenarios, with a recall of 0.845 and an F1 score of 0.738 for the Unknown class, but a precision of only 0.655.

This indicates that the model tends to classify more samples as "Unknown", thereby identifying more unknown samples, but at the cost of causing more known categories to be misclassified as "Unknown."

Some models exhibited a significant performance drop under mixed unknown conditions, primarily due to low recall in the "Unknown" class—that is, a large number of unknown samples were not successfully rejected but were instead misclassified as known categories. The test accuracy of ViT_B32 dropped to 68.17%, with a recall rate for the "Unknown" class of only 0.205 and an F1 score of 0.317; DenseNet121 achieved a test accuracy of 71.17%, with a recall rate for the "Unknown" class of 0.235 and an F1 score of 0.355. These results indicate that both models exhibit significant deficiencies in identifying the "unknown" class under mixed unknown sample conditions. VGG19 and Swin_Tiny performed at an intermediate level, with "unknown" class F1 scores of 0.522 and 0.535, respectively, suggesting that while they possess some rejection capability, a significant number of unknown samples still leak into the known label space.

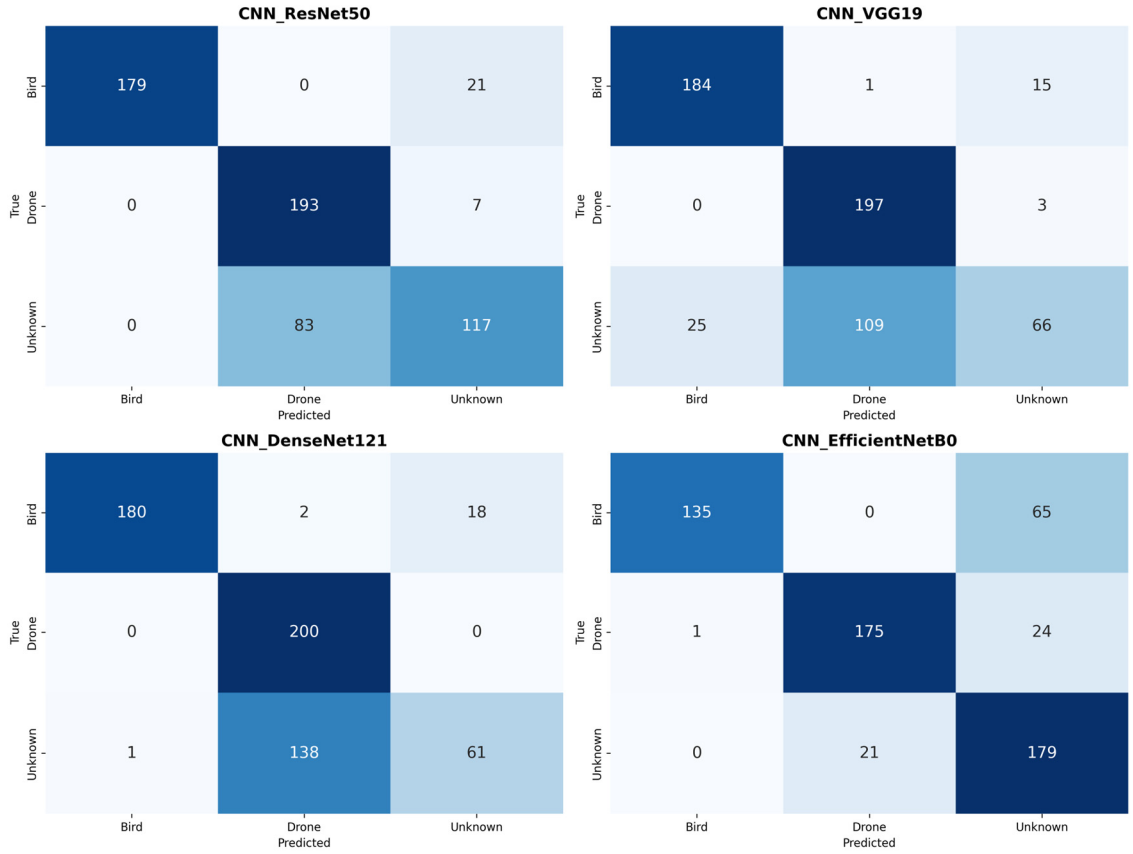
When Plane and Balloon are merged into a unified unknown sample pool, the score-based detection results generally fall between those of the two single-unknown settings, reflecting the complexity of simultaneously handling near-OOD and far-OOD samples. As shown in Table 10, ViT_B16 once again demonstrates the strongest overall threshold-independent detection performance, with AUROC and AUPR reaching 0.9746 and 0.9446, respectively, indicating its clearest ability to distinguish between known samples and mixed unknown samples. Swin_Small also performed outstandingly, with an AUROC of 0.9689 and an AUPR of 0.9116, while maintaining a low FPR95 (0.1100), indicating strong stability even under high recall conditions. Among the CNN models, VGG19 performed best in fractional detection, with an AUROC of 0.9339, an AUPR of 0.8223, and an FPR95 of 0.1650; ResNet50 performed at an intermediate level, with an AUROC of 0.8950 and an AUPR of 0.8130. In contrast, ViT_B32 and Swin_Tiny exhibited lower AUROC and AUPR values, along with a higher FPR95, indicating that they still face greater difficulty in distinguishing mixed unknown samples from known categories. Overall, these threshold-independent metrics align with the results obtained using fixed thresholds and further illustrate that the difficulty of the mixed-unknown scenario lies between the two single-unknown settings: it is more challenging than the balloon-only scenario because it contains aircraft samples that are semantically closer to the known classes; yet it is less challenging than the aircraft-only scenario because balloon samples are visually easier to distinguish from the known classes.

Table 10. Score-based open-set detection performance under two unknown classes (Plane + Balloon).

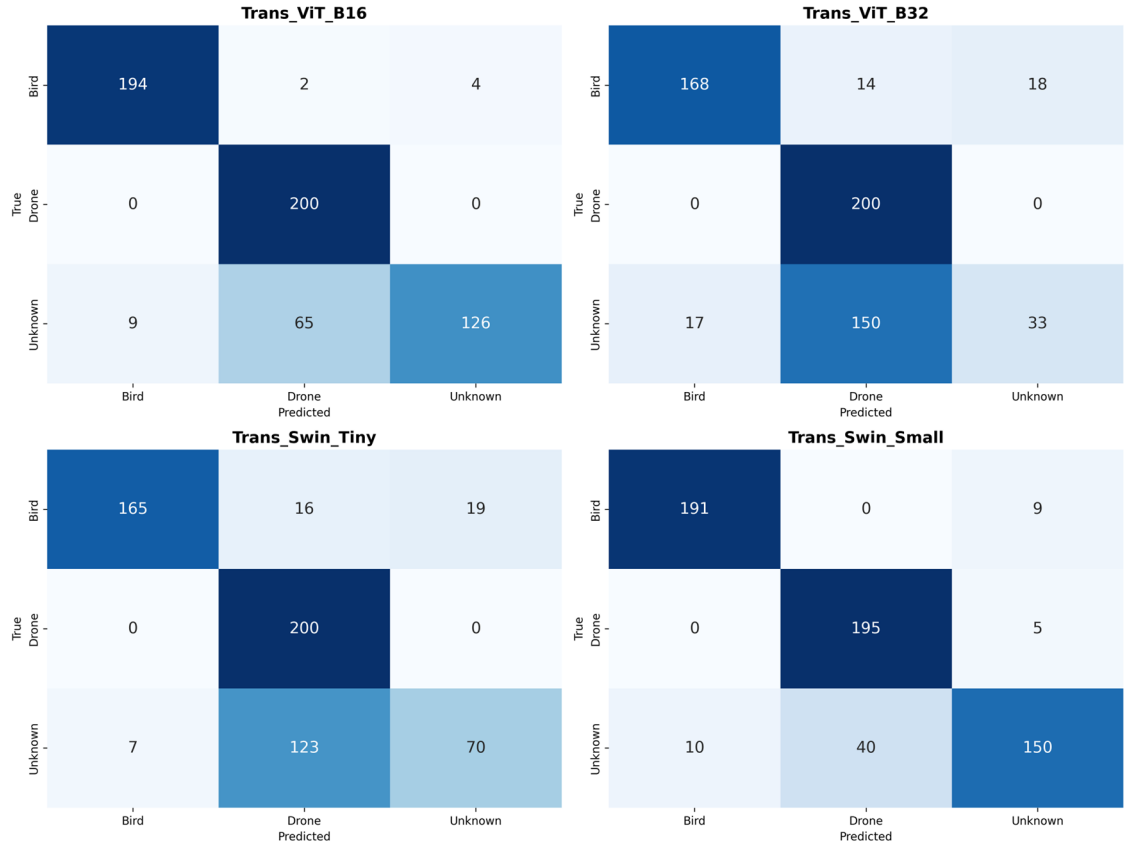
Model	AUROC	AUPR	FPR95
ResNet50	0.895	0.813	0.4475
VGG19	0.9339	0.8223	0.165
DenseNet121	0.8781	0.6986	0.2975
EfficientNet_B0	0.8656	0.679	0.3825
ViT_B16	0.9746	0.9446	0.1
ViT_B32	0.8369	0.6482	0.495
Swin_Tiny	0.8646	0.741	0.4825
Swin_Small	0.9689	0.9116	0.11

The confusion matrices shown in Figure 6a,b further indicate that the mixed-unknown scenario remains challenging for many models. A large number of unknown samples are either correctly classified as "Unknown" or misclassified as "Drone", while the proportion misclassified as 'Bird' is relatively low. This pattern is particularly evident in VGG19, DenseNet121, ViT_B32, and Swin_Tiny, further illustrating that under mixed-unknown conditions, unknown samples are more likely to cluster toward the "Drone" category.

In comparison, EfficientNet_B0, ViT_B16, and Swin_Small show better rejection performance and keep more mixed unknown samples in the Unknown class. These results indicate that when plane and balloon samples are combined, the ability to reject unknown inputs remains different across models under the same threshold.



(**a**)



(**b**)

Figure 6. (**a**) Confusion matrices of CNN models in the Mixed-Unknown setting. (**b**) Confusion matrices of Transformer models in the Mixed-Unknown setting.

4.5. Discussion

Experiments in open-world settings demonstrate that strong performance in confined environments does not necessarily translate into reliable recognition of unseen objects in real-world deployment [7]. Figure 7 summarizes testing accuracy across the three unknown-class settings and shows a consistent ordering across architectures: performance is highest in the Balloon (far-OOD) setting, lowest in the Plane (near-OOD) setting, and the Mixed setting generally falls between them. This indicates that semantic similarity to drones substantially increases the difficulty of unknown rejection, whereas visually more distinct unknowns are easier to separate under the same confidence-based decision rule.

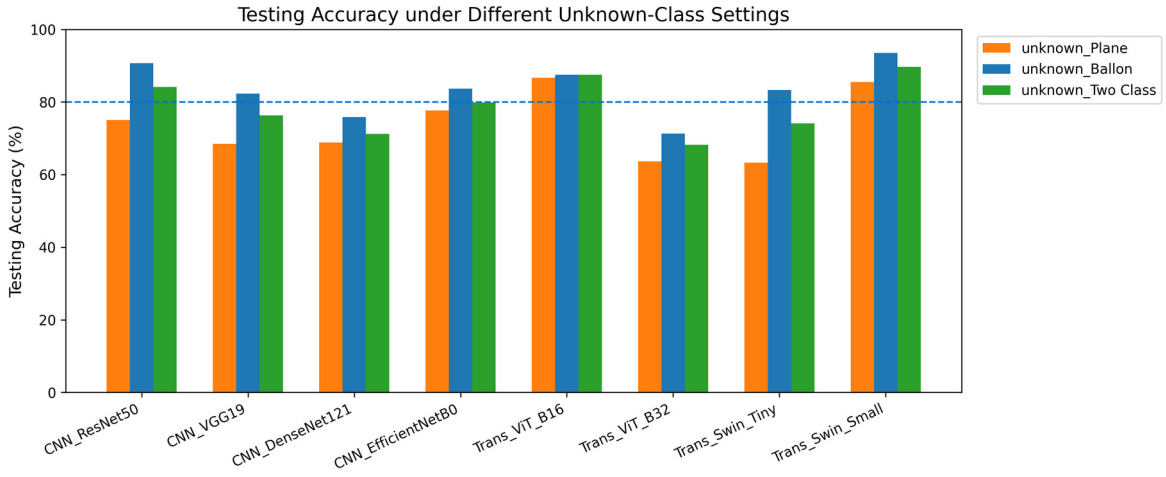


Figure 7. Testing accuracy under three unknown-class settings (Plane, Balloon, and Two-Class Mixed).

In addition to the overall accuracy, the confusion matrix analysis in Sections 4.4.1 through 4.4.3 also reveals a consistent error pattern: when unknown samples are not successfully rejected, they are more likely to be misclassified as drones (Drone) rather than birds (Bird). This suggests that, under open-set conditions, the "drone" class acts to some extent as a "lure class" for unknown aerial targets. The representative error samples shown in Figures 8 and 9 further corroborate this phenomenon from a qualitative perspective. In Swin_Small, representative errors include aircraft samples misclassified as "Bird", aircraft samples misclassified as "Drone", balloon samples misclassified as "Bird", and balloon samples misclassified as "Drone", with corresponding MSP values of 0.924, 0.960, 0.836, and 0.827, respectively. In ViT_B32, errors of the same type are even more pronounced, with corresponding MSPs further increasing to 0.973, 1.000, 0.993, and 0.998. For ResNet50, representative misclassification samples are similarly concentrated in the direction where unknown samples are absorbed into the Drone class, with MSPs reaching 0.985 and 0.965 when airplane samples are misclassified as Drones and balloon samples are misclassified as Drones, respectively. Since these values are all significantly higher than the unified rejection threshold $\tau = 0.8$, these unknown samples are absorbed into the known class space with high confidence rather than being classified as "Unknown".

Meanwhile, the representative correctly rejected samples shown in Figure 10 demonstrate that, under the same mixed-unknown setting, the model is indeed capable of successfully classifying unknown targets as "Unknown". The MSPs for the four examples are 0.507, 0.579, 0.542, and 0.567, respectively, all of which are below the unified threshold $\tau = 0.8$ and are therefore correctly assigned to the "Unknown" class. It is worth noting that these successful rejection samples include both near-OOD aircraft and far-OOD balloons, indicating that the model is not entirely incapable of rejecting unknown classes, but rather can effectively suppress the maximum confidence in known classes for certain samples, thereby achieving successful rejection. When considering these success stories alongside the aforementioned failures, the primary limitation of MSP-based open-set detection lies not in the model's complete inability to reject unknown classes, but rather in the significant variability in the reliability of rejection results across different models and under different unknown conditions. For some unknown samples, the MSP can be suppressed to approximately 0.51–0.58, causing the model to correctly classify them as "Unknown"; for other unknown samples, however, the MSP may rise to 0.83 or even approach 1.00, leading the model to misclassify them as 'birds' or "drones" with high or even extremely high confidence. Overall, these qualitative examples further support the conclusions of the quantitative analysis discussed earlier: unknown targets that are semantically closer to known classes remain more difficult to reject, and a model's open-set behavior depends not only on overall accuracy but is also closely related to how confidence is distributed across unknown aerial targets in different architectures.

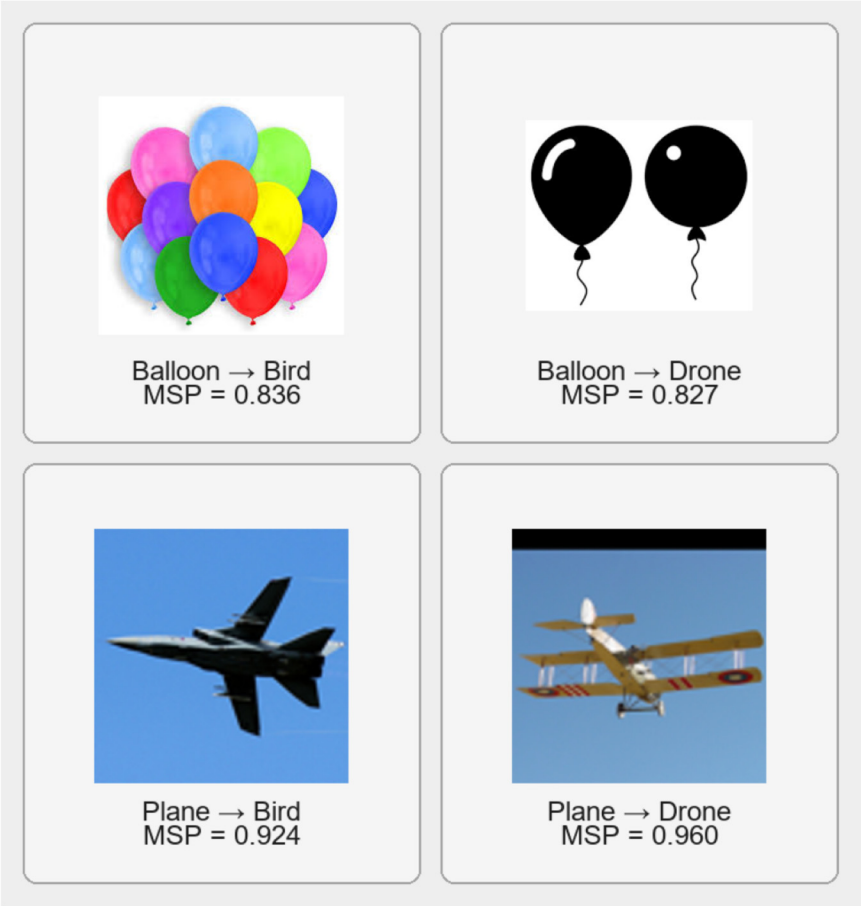


Figure 8. Representative open-set failure cases of Swin_Small.

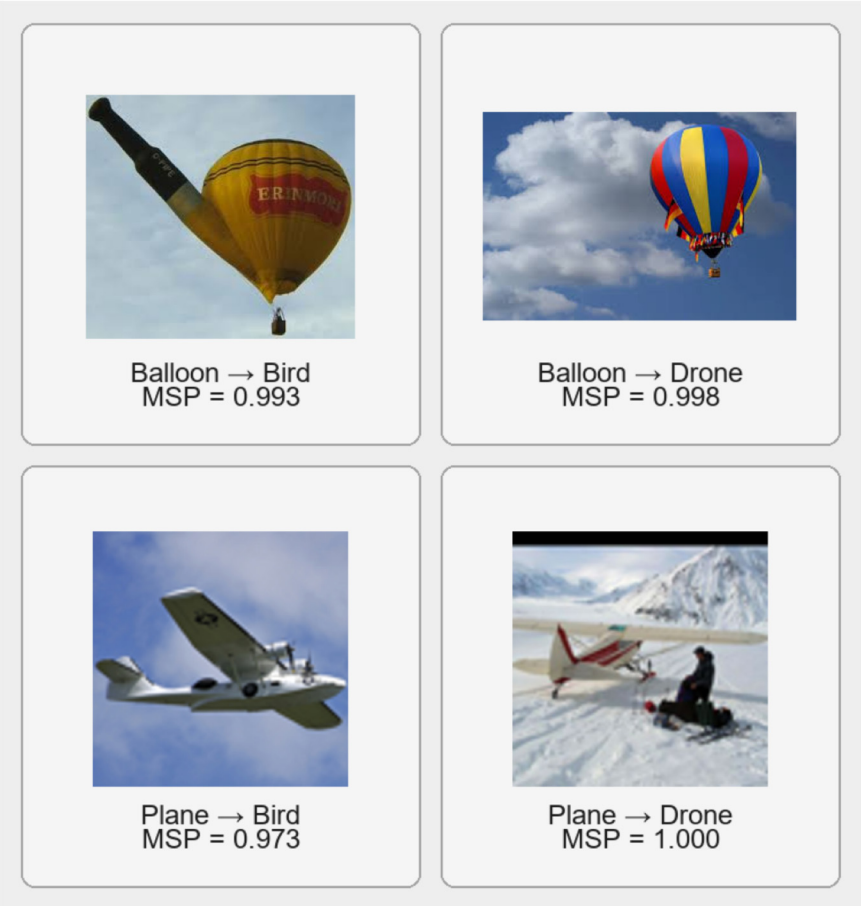


Figure 9. Representative open-set failure cases of ViT_B32.

Representative misclassified samples from ResNet50 show a pattern similar to that of Swin_Small, but are concentrated mainly in the Unknown-to-Drone direction. Since these error examples are highly similar in visual pattern to those already shown for Swin_Small, they are not presented as a separate figure here. In the selected exported cases, the Plane → Drone and Balloon → Drone errors reach MSP values of 0.985 and 0.965, respectively, while no representative Unknown → Bird case was observed. This further suggests that, for ResNet50, open-set errors are more strongly biased toward absorbing unseen aerial objects into the Drone class.

Model selection in practical deployment is also influenced by computational efficiency. Table 11 compares the training and inference times of various models under uniform hardware and identical experimental settings. The results indicate significant differences in computational overhead among different architectures. Among the CNN models, ResNet50 and EfficientNet_B0 demonstrate superior training efficiency. ResNet50 has the shortest training time per epoch, while EfficientNet_B0 maintains low training overhead, indicating that these two convolutional architectures offer good computational efficiency for the current task. Although VGG19 does not have the shortest training time, it has the lowest inference latency per image, indicating that this model has a certain advantage in terms of forward prediction speed. DenseNet121's training and inference costs are at an intermediate level, with its overall computational burden falling between lightweight and high-cost models.

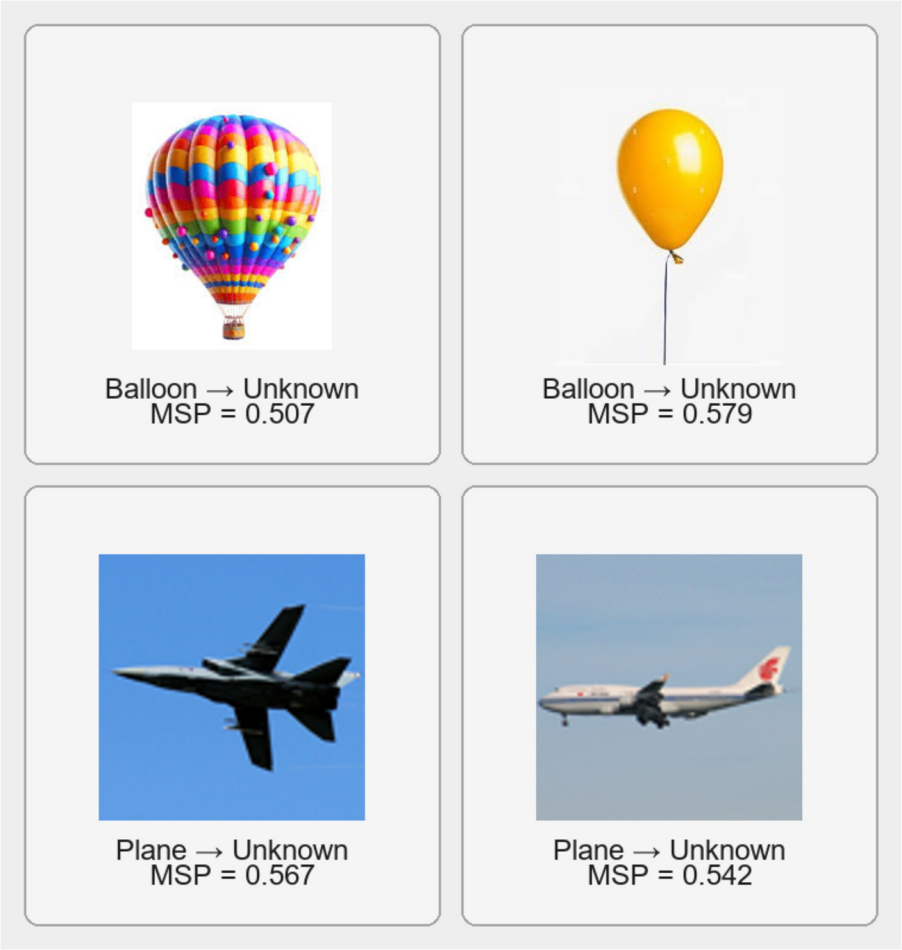


Figure 10. Representative open-set correctly rejected unknown samples.

Table 11. Computational cost comparison across architectures.

Model	Category	Training Time (per Epoch)	Inference Time (per Image)
ResNet50	CNN	35.04	10.274
VGG19	CNN	48	5.1151
DenseNet121	CNN	39.44	24.5074
EfficientNet_B0	CNN	36.72	12.2736
ViT_B16	Transformer	42.53	7.1363
ViT_B32	Transformer	45.25	19.1277
Swin_Tiny	Transformer	45.25	19.1277
Swin_Small	Transformer	49.5	34.4139

Among Transformer models, ViT_B32 demonstrates high efficiency in both training and inference, suggesting that larger patch sizes help reduce computational costs to some extent. In contrast, the Swin series, particularly Swin_Small, exhibits higher computational overhead during both training and inference, with Swin_Small having the highest overall cost among all models. ViT_B16's efficiency falls between the two; although its training and inference costs are higher than those of some CNN models, they remain significantly lower than those of Swin_Small. Overall, Table 11 clearly illustrates the trade-offs in computational efficiency among different models: some models have lower training and inference costs, while others require higher computational overhead to achieve more complex representational capabilities. Therefore, in practical deployment, model selection must not only consider recognition performance but also balance training costs with inference latency, and weigh efficiency requirements based on specific application scenarios.

5. Conclusions and Future Work

In this paper, we conduct a systematic comparison of the performance of CNN and Vision Transformer architectures on the bird–drone recognition task within a unified experimental framework. We also perform closed-set and open-set evaluations under three deployment-oriented unknown-class scenarios, including near-OOD, far-OOD, and mixed unknown scenarios. The experimental results indicate that high accuracy under closed-set conditions does not directly reflect a model's actual reliability in open-set scenarios. This aligns with the core principle of open-set recognition, namely that classification performance on known classes alone is insufficient to ensure a model's ability to effectively manage open-space risk [7]. In this study, the near-OOD scenario corresponding to the "Plane" class was consistently the most challenging, while the far-OOD scenario corresponding to the "Balloon" class was relatively easier to distinguish; the overall difficulty of the mixed unknown scenario typically fell between the two. This further supports the finding in the general OOD/OSR literature that semantically similar unknown classes are more difficult to handle [9]. Analysis of the confusion matrix revealed that, if not correctly rejected, unknown samples are more likely to be misclassified as "drones" rather than "birds". This demonstrates that, under open-set conditions, different architectural approaches exhibit varying degrees of overgeneralization to known categories when encountering unknown aerial targets. In conjunction with the analysis of computational efficiency, the selection of a model for practical deployment should depend not only on classification performance under closed-set conditions, the model's ability to handle unknown examples under open-set conditions, and typical misclassification patterns, but also on the computational cost during the training and inference phases. Under the experimental conditions of this work, we found that ResNet50 exhibits a good overall balance between recognition performance and computational cost; conversely, when classification reliability in an open environment is a greater priority, ViT_B16 and Swin_Small are the more attractive choices despite higher training and computational costs, due to their superior recognition performance.

In future research, we plan to investigate methods that go beyond simple confidence thresholds. These include OpenMax-like recalibration methods [16], ODIN-like temperature scaling and input distortion methods [17], Mahalanobis detection methods based on feature space distances [18], and energy-based OOD evaluation frameworks [19]. Furthermore, future research will consider the introduction of hybrid methods such as ViM [20] to further improve the ability to distinguish between known and unknown samples by combining logit information with feature representations, while simultaneously utilizing post-calibration techniques to increase the alignment between model confidence and prediction accuracy [21].

At the data level, future work will incorporate more OOD-like air targets to expand the set of unknown categories and create evaluation scenarios that more closely resemble real-world operational environments. Through these improvements, future research is expected to further investigate, at acceptable training and inference costs, how the model's open-domain robustness can be enhanced in real-world application scenarios.

Author Contributions

Z.T.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, and writing—original draft preparation. V.L.: supervision and writing—review and editing. H.M.: supervision and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding. The APC was funded by the authors.

Institutional Review Board Statement

Not applicable. This study did not involve human participants or animal experiments.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data and code supporting the findings of this study are available from the corresponding author upon reasonable request. The image data used in this study were collected from publicly available online resources, and their redistribution is subject to the licenses and terms of the original sources.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

While writing this article, the authors used ChatGPT to assist with language editing, sentence refinement, and enhancing academic expression, and used DeepL to translate the content from the authors' native language, Chinese, into the official language, English. After using these tools, the authors carefully reviewed the content and made revisions as necessary, and assumes full responsibility for the content of the published article.

References

1. Coluccia, A.; Fascista, A.; Schumann, A.; et al. Drone vs. bird detection: Deep learning algorithms and results from a grand challenge. *Sensors* **2021**, *21*, 2824. https://doi.org/10.3390/s21082824.
2. Coluccia, A.; Fascista, A.; Dimou, A.; et al. The Drone-vs-Bird Detection Grand Challenge at IJCNN 2025. In Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN), Rome, Italy, 30 June–5 July 2025.
3. Shandilya, S.K.; Srivastav, A.; Yemets, K.; et al. YOLO-based segmented dataset for drone vs. bird detection for deep and machine learning algorithms. *Data Brief* **2023**, *50*, 109355. https://doi.org/10.1016/j.dib.2023.109355.
4. Kaur, D.; Battish, N.; Bhavsar, A.; et al. YOLOBirDrone: Dataset for Bird vs Drone Detection and Classification and a YOLO based enhanced learning architecture. *arXiv* **2026**, arXiv:2601.08319.
5. Rahman, S.; Robertson, D.A. Classification of drones and birds using convolutional neural networks applied to radar micro-Doppler spectrogram images. *IET Radar Sonar Navig.* **2020**, *14*, 653–661. https://doi.org/10.1049/iet-rsn.2019.0493.
6. Akyon, F.C.; Akagündüz, E.; Altinuc, S.O.; et al. Sequence Models for Drone vs. Bird Classification. In Proceedings of the Sixteenth International Conference on Machine Vision (ICMV 2023), Yerevan, Armenia, 15–18 November 2023.
7. Scheirer, W.J.; de Rezende Rocha, A.; Sapkota, A.; et al. Toward open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1757–1772.
8. Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2017.
9. Yang, J.; Zhou, K.; Li, Y.; et al. Generalized out-of-distribution detection: A survey. *Int. J. Comput. Vis.* **2024**, *132*, 5635–5662. https://doi.org/10.1007/s11263-024-02117-4.
10. He, K.; Zhang, X.; Ren, S.; et al. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
11. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
12. Huang, G.; Liu, Z.; Van Der Maaten, L.; et al. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
13. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual, 3–7 May 2021.
15. Liu, Z.; Lin, Y.; Cao, Y.; et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.

16. Bendale, A.; Boult, T.E. Towards open set deep networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
17. Liang, S.; Li, Y.; Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 30 April–3 May 2018.
18. Lee, K.; Lee, K.; Lee, H.; et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31, Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada, 3–8 December 2018*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2018.
19. Liu, W.; Wang, X.; Owens, J.; et al. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems 33, Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Online, 6–12 December 2020*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2020; pp. 21464–21475.
20. Wang, H.; Li, Z.; Feng, L.; et al. Vim: Out-of-distribution with virtual-logit matching. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022.
21. Guo, C.; Pleiss, G.; Sun, Y.; et al. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017.
22. Sun, J.; Dong, Q. A survey on open-set image recognition. *arXiv* **2023**, arXiv:2312.15571.
23. Deng, J.; Dong, W.; Socher, R.; et al. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
24. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
25. Powers, D.M. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *arXiv* **2020**, arXiv:2010.16061.
26. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.
27. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006.