remote sensing

MDPI

Article

Multiple Instance Learning Convolutional Neural Networks for Fine-Grained Aircraft Recognition

Xiaolan Huang [1], Kai Xu [1,*], Chuming Huang [1], Chengrui Wang [1] and Kun Qin [2]

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, China; hxl1997@cug.edu.cn (X.H.); huangcm98@cug.edu.cn (C.H.); ygwcr@cug.edu.cn (C.W.)
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; qink@whu.edu.cn
* Correspondence: xukai21@cug.edu.cn; Tel.: +86-135-4500-3637

Abstract: The key to fine-grained aircraft recognition is discovering the subtle traits that can distinguish different subcategories. Early approaches leverage part annotations of fine-grained objects to derive rich representations. However, manual labeling part information is cumbersome. In response to this issue, previous CNN-based methods reuse the backbone network to extract part-discrimination features, the inference process of which consumes much time. Therefore, we introduce generalized multiple instance learning (MIL) into fine-grained recognition. In generalized MIL, an aircraft is assumed to consist of multiple instances (such as head, tail, and body). Firstly, instance-level representations are obtained by the feature extractor and instance conversion component. Secondly, the obtained instance features are scored by an MIL classifier, which can yield high-level part semantics. Finally, a fine-grained object label is inferred by a MIL pooling function that aggregates multiple instance scores. The proposed approach is trained end-to-end without part annotations and complex location networks. Experimental evidence is conducted to prove the feasibility and effectiveness of our approach on combined aircraft images (CAIs).

Keywords: fine-grained image recognition; aircraft recognition; multiple instance learning; loss function

Citation: Huang, X.; Xu, K.; Huang, C.; Wang, C.; Qin, K. Multiple Instance Learning Convolutional Neural Networks for Fine-Grained Aircraft Recognition. *Remote Sens.* **2021**, *13*, 5132. https://doi.org/10.3390/rs13245132

Academic Editors: Sidike Paheding, Zahangir Alom, Maitiniyazi Maimaitijiang and Matthew Maimaitiyiming

Received: 2 November 2021
Accepted: 14 December 2021
Published: 17 December 2021

1. Introduction

Remote sensing images are proposed to capture the particulars of surface features with the advancement of remote sensing technology, thereby accelerating the development of high-resolution remote sensing image interpretation. However, it is necessary to obtain feature subcategories due to practical applications. For example, regulatory authorities need to use remote sensing images to know aircraft types (e.g., Airbus 330 and Boeing 737) to administer air transportation in airports. Therefore, fine-grained aircraft recognition has become one of the research emphases in remote sensing image recognition. The goal of aircraft recognition is aimed at mining distinguishing characteristics from subordinate categories. As shown in Figure 1a, UCMerced Land Use Dataset [1] only contains airplane scene semantics. The combined aircraft images (CAIs) subdivide the aircraft category into several subcategories in Figure 1b.

In the past few decades, many scholars proposed several approaches in the aircraft recognition task. The traditional method extracts handcraft features, which empirically selects or fuses obtained characteristics to constitute aircraft features. Handcraft features consist of scale-invariant feature transform [2], moment invariants [3–5], and Zernike moments [6]. The handcraft feature combination methods include principal component analysis [7], which automatically learns a group of weights from training samples. The classification approaches, such as back-propagation neural networks [4], independent component algorithm [6], and tree classifier [8], engraft extracted features to allow the recognition model to make precise decision-making. The above low-level information on an

aircraft combines with middle-level semantics, such as the bag of words [9]. Some scholars also use template matching for aircraft recognition [10,11]. However, the above-mentioned features possess weak generalization ability because the selection and combination of low or mid-level handcrafted features demand abundant professional and prior knowledge to generate rich characteristic expression. Thus, the classification results of traditional methods tend to be unstable and inaccurate.

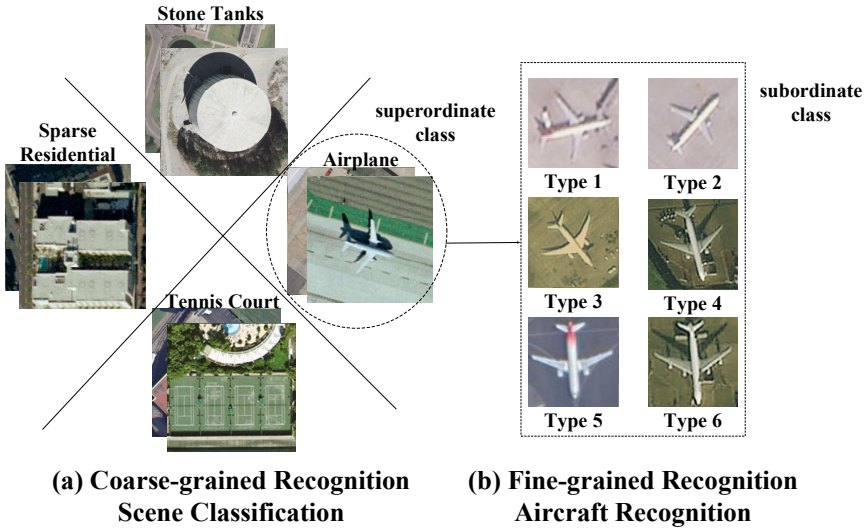


Figure 1. Schematic diagram of coarse-grained recognition and fine-grained recognition.

The rapid progress of deep learning has resulted in the emergence of some advanced convolutional neural networks (CNNs), such as AlexNet [12], VGGNet [13], GoogLeNet [14], and ResNet [15], in the computer vision community. Deep CNNs (DCNNs) automatically cultivate high-dimensional feature expressions. Therefore, previous researchers obtain aircraft features using DCNNs, thereby considerably increasing recognition performance [16–18] and enhancing the generalization ability of the recognition model [15–17]. Recent researchers have been locating components to highlight differences. The traditional locating methods include strong supervision and weak supervision (see Figure 2). The strong supervision utilizes box coordinates to local part regions [19], and the weak supervision performs part learning via a positioning attention mechanism to compensate for the lack of part annotation data. For example, K. Fu et al. [20] proposed multiple class activation mapping to locate discriminative features and recognize aircraft types without part information. Yunsheng Xiong et al. [21] proposed a non-locally enhanced feature fusion network to cultivate holistic representations and highlight part responses. The above approaches demand the complex design of locating networks to encode subcategory object features, which occupy many model parameters and consume much reasoning time.

To decrease the high consumption of obtaining part information, we attempt to introduce semi-supervised learning. Semi-supervised learning is dedicated to extracting explicit semantics by marked labels and mining implicit information with unlabeled samples. Meanwhile, the input data are composed of labeled images and part of unlabeled images, which can reduce the cost of acquiring training data. The main semi-supervised learning methods are divided into adaptive learning [22] and regularization methods. The main regularization methods include mean teacher [23], high-rankness regularization [24], and stack auto-encoder [25] to improve recognition performance by unlabeled data. Additionally, Bei Fang et al. [26] propose a novel dual-strategy sample selection co-training algorithm to achieve the same effect.

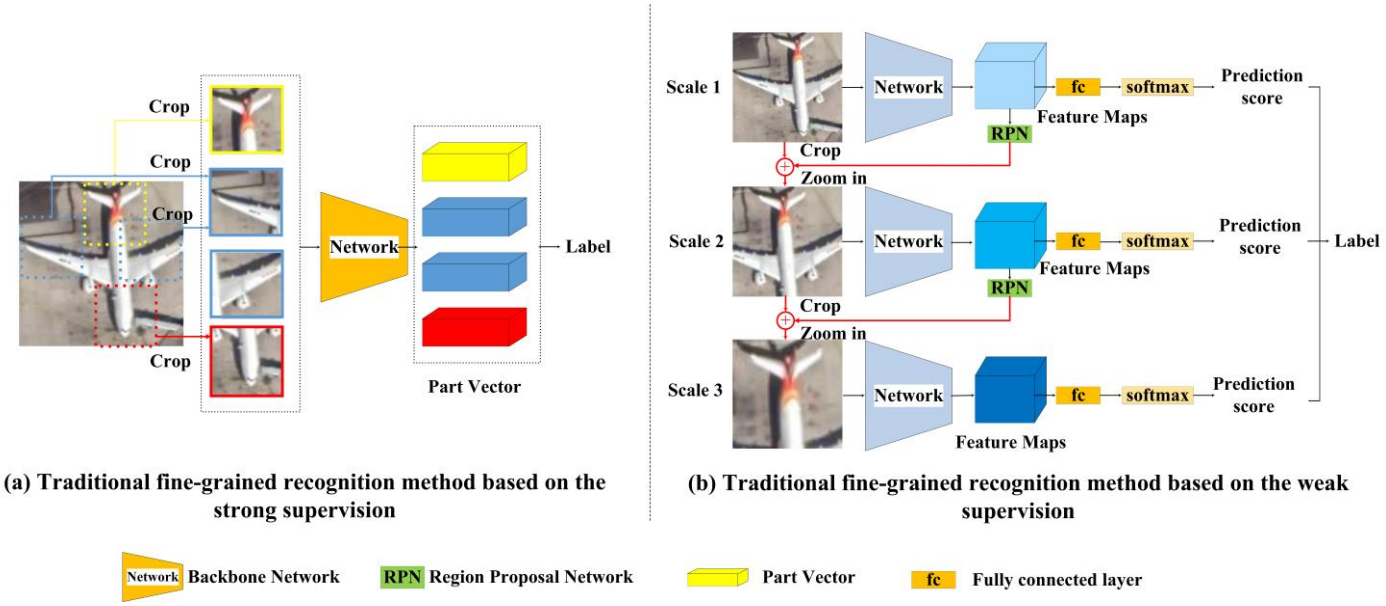


Figure 2. Traditional fine-grained aircraft recognition procedures using a strong supervision method and a weak supervision method separately. The process of aircraft recognition via strong supervision method is indicated in Figure (**a**), in which the local positions of aircraft components are obtained by the given box coordinates. The weak supervision scheme, shown in Figure (**b**), employs the cascade-attention model to mine distinguishable information.

The above approaches mainly tackle the scarcity of training data. However, in fine-grained recognition tasks, the label of the training data is already obtained, whereas the part annotations are unmarked. Therefore, we introduce multiple instance learning (MIL), which is used for image recognition to tackle obstacles in obtaining component information. MIL only adopts bag-level labels. In MIL, a bag comprises multiple instances, and a bag label is contributed by extracted instance features without instance information. MIL, originating from the activity detection of drug molecules [27], has been widely used in computer vision, such as image segmentation [28], image marking [29], and image retrieval [30]. For instance, Miao Sun et al. [31] proposed a loss function based on MIL to improve recognition performance. Mengran Fan et al. [32] enhanced foreground instances using an attention mechanism to enhance the expression ability of features. In the field of remote sensing image recognition, Zhili Li et al. [33] proposed an MIL scene feature-representation method, wherein a scene is represented as a bag of local patches. It can be facilitated to construct a robust scene representation. Accordingly, MIL is expected to show considerable potential in the recognition task.

Motivated by [33], we attempt to explore the effectiveness of MIL in fine-grained aircraft recognition. Meanwhile, the standard MIL is applied to represent scene semantics that contain several separate objects (see Figure 3a). Thus, we use generalized MIL applied in fine-grained aircraft recognition. In generalized MIL aircraft recognition, an aircraft is represented as a bag of component concepts, and a component concept denotes an aircraft part, such as the head, tail, or wing (see Figure 3b). Additionally, randomly arranging the order of the instances, which is termed permutation invariance of generalized MIL, is robust to aircraft recognition [34]. It facilitates the semantic information of robustness to various spatial transformations, such as translation, rotation, and mirroring. This scenario is conducive to expressing different poses of aircraft.

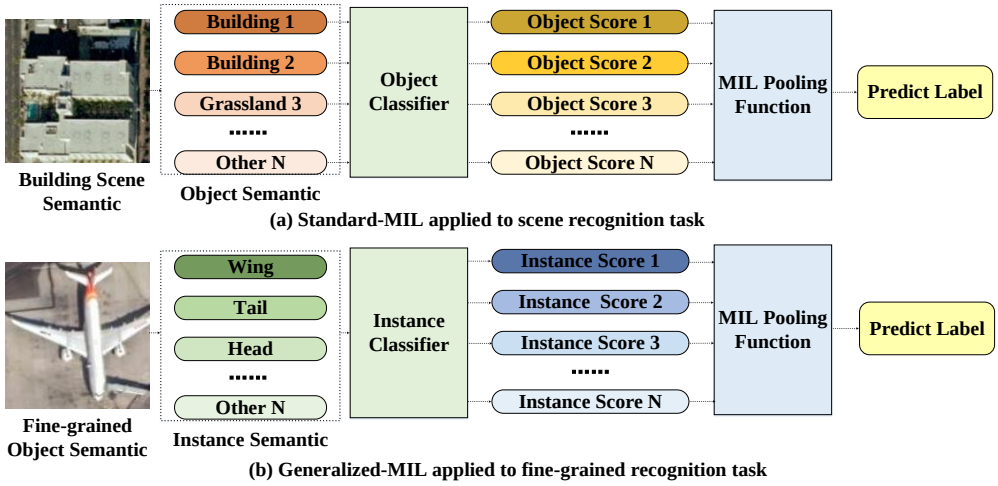


Figure 3. Schematic diagram of MIL classification recognition. Figure (**a**) illustrates the standard MIL classification recognition approach, and Figure (**b**) outlines the sketch map of the generalized MIL framework in fine-grained aircraft classification.

The structure of the generalized MIL recognition framework includes a feature extractor, an instance conversion component, and an MIL pooling component. The feature extractor adopts a pretrained network to obtain basic characteristics, transforming the input image into a set of local patch features (patch-level features). In Figure 4a, an aircraft component in an image is divided into several patches. In this case, a single patch cannot express the semantic information of specific parts, and the operation that directly feeds images into backbones destroys the characterization ability of an instance. Consequently, we design an instance conversion component, namely, instance loss function, to convert patch-level response patterns to instance-level response patterns (see Figure 4b).

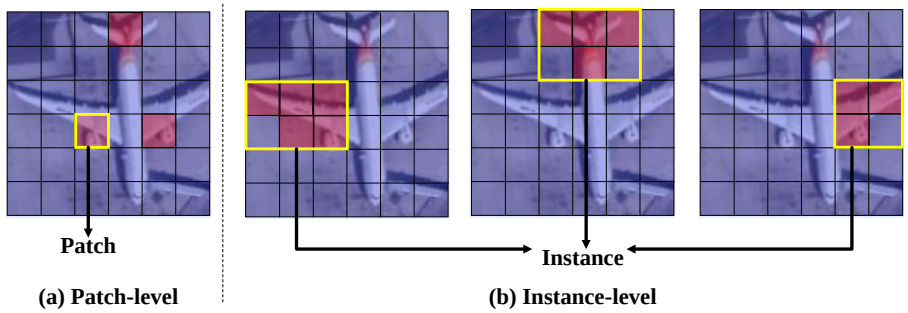


Figure 4. Sketch maps of the patch-level and instance-level feature maps. In Figure (**a**), the feature map is extracted by the backbone network, the size of which is H × W × N. In Figure (**b**), the multiple feature maps are inferred by the generalized MIL network, the size of which is H × W × N. The blue grids indicate the background responses, and the red grids represent the emphasized foreground value. The yellow boundary line is to spotlight the area comprised of the red grids.

The instance loss function is studied to learn the compactness and separability of multiple instances. Similar functions have been achieved in face recognition. For example, the contrastive loss [35] explores map input patterns into a target space, such that the L_1 norm in the target space approximates the semantic distance in the input space. The above-mentioned loss function optimizes feature distribution in a given metric space. Unlike metric learning, mutual-channel loss (MCL) [36] is aimed at gaining individual feature channels early on, as opposed to the convention of starting from a consolidated feature map, which provides a novel perspective of fine-grained recognition. Thus, the proposed instance loss function is devised to accomplish instance feature learning, referring to MCL. We employ a fixed percentage of randomly masked channels to generate several

instance blocks, which are termed instance masks. Then, the masked features flow into the instance conversion part composed of discriminability and diversity components. In the discriminability component, we apply cross-channel max pooling [37] (CCMP) to fuse the channels of each instance and maximize the mean of instance peaks, which can concentrate on mutually distinct local instances. In the diversity component, we leverage the softmax function to receive the normalized feature and adopt CCMP to maximize spatial decorrelation across instances belonging to the same category. Finally, the extracted instance-level features, aggerated by the MIL pooling function, infer classification probabilities, aiming to yield robust expressions considering all instance scores. Introducing additional information and complex training strategies is unnecessary.

The main contributions of this article include the following:

(1) We attempt to propose a generalized MIL fine-grained aircraft recognition method to focus on the discriminative regions and reduce the excessive computational cost of extracting fine-grained part features without marked part annotations. In generalized MIL, a fine-grained aircraft is combined with several component concepts, which are only known aircraft types. It can effectively spotlight part regions and suppress background response values.

(2) The patch-level output extracted from the MIL backbone network cannot simply present aircraft part semantics. Thus, we design an instance conversion part (instance loss function) to transform patch-level information to instance-level fine-grained semantic representations, which involves few model parameters and little testing time.

(3) We apply a self-made benchmark dataset CAIs in the remote sensing fine-grained images to demonstrate the effectiveness and universality of our method. Comprehensive experimental evaluations of basic DCNNs verify the effectiveness of the proposed method.

The rest of this article is organized as follows. Section 2 describes the theory and architecture of our proposed method in detail. Section 3 specifies the metric datasets and network hyperparameters and provides experimental results to demonstrate the feasibility and effectiveness of the proposed method. Section 4 discusses our proposed method. Section 5 summarizes the content of this article.

2. Materials and Methods

In this section, the theory of MIL recognition is illustrated in detail. Subsequently, we introduce the structure of the generalized MIL fine-grained recognition, which mainly describes the instance conversion component and the MIL pooling component.

2.1. Problem Statement

According to the relationship between the instances and bag labels, MIL includes the standard and generalized paradigms [38]. The standard MIL [27] suggests that the bag label depends on the most positive instance. Each instance exhibits an implicit mark $c \in \Omega = \{+1, -1\}$. In standard MIL, we assume $X = \{X_1, X_2, \ldots, X_n\}$ denotes a bag containing n instances $x_i \in \chi$ $(i = 1, 2, \ldots, n)$, and $c(x)$ represents an instance-level classifier. The formula of predicting labels is as follows:

$$f_{SMIL}(X) = \begin{cases} 1, \exists c(x_i) = 1 \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

The standard MIL recognition results are inferred by multiple individual instance scores, which is suitable to perform the object semantics in image understanding. The object semantics belongs to simple semantics, which is expressed by several closed regions in an image. The building object semantics are shown in Figure 5a, which contains complete individual buildings to signify building semantics. However, the aircraft parts are not demarcated by separate closed areas on an aircraft object, as shown in Figure 5b.

Aircraft components occur in an adjacency relationship. Thus, there are existing limitations applying the standard MIL assumption to fine-grained recognition.

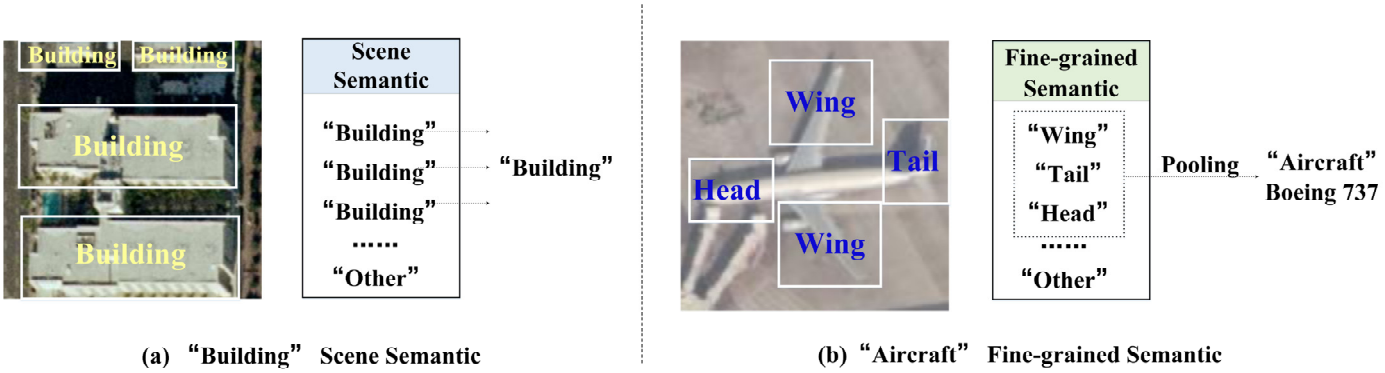


Figure 5. Diagrammatic sketch of the building object semantics and aircraft fine-grained semantics.

To break the limitations of standard MIL, we introduce the generalized MIL paradigm reference [39], which supposes a bag label is inferred by multiple instance concepts $C = \{c_1, c_2, \ldots, c_r\}(c_i: \chi \rightarrow \Omega)$, to optimize the standard MIL method in aircraft recognition. $N(X, c_i)$ signifies the number of instances corresponding to c_i in the bag X. The predicted formula can be written as follows:

$$f_{GMIL}(X) = \begin{cases} 1, & \forall c_i: N(X, c_i) \geq 1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

which indicates that an image marked as positive contains a positive instance referring to the instance concept c_i at least. The instance concept represents the sub-semantics of an aircraft, such as head, tail, and wing.

Thus, we establish the generalized MIL method to fine-grained recognition. Different recognition methods of response value distribution obtained by Grad-CAM [40] are displayed in Figure 6. In particular, Figure 6a shows that the features extracted by the baseline network are sparsely distributed, discarding aircraft part information to abstract ambiguous expressions. Then, Figure 6b displays the obtained results from the standard MIL network, the attention distribution of which can cover the whole aircraft. However, the attention pattern does not concentrate on fine-grained exclusive regions. Moreover, Figure 6c shows the attention distribution derived from a generalized MIL network, which can observe the response distribution of typical components and generate the instance-level attention distribution mode. Consequently, utilizing generalized MIL into aircraft recognition tasks is reasonable based on the above visualization results.

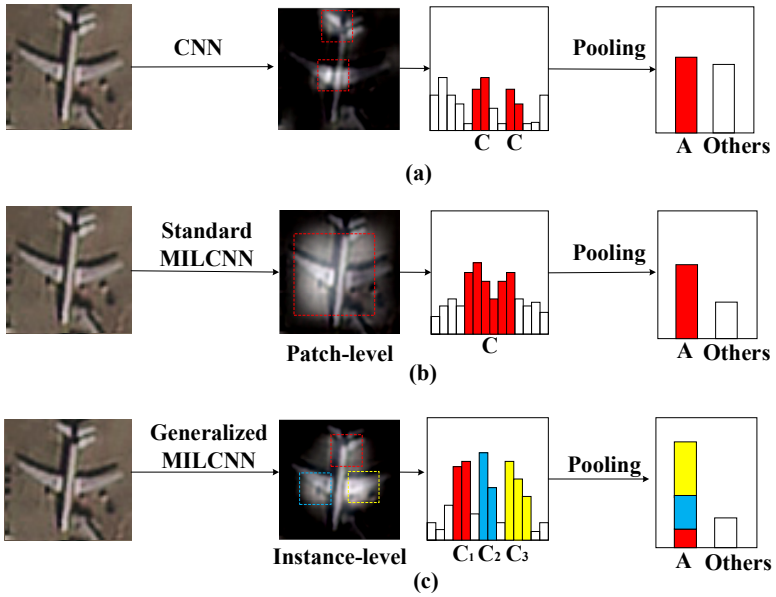


Figure 6. Schematic diagram of CNN, standard MILCNN, and generalized MILCNN in aircraft recognition. Figure (**a**) is the attention map extracted by CNN. The red box is the highlighted regions of the aircraft. c represents the category of aircraft. Figure (**b**) is the attention map extracted by standard MILCNN. The red columns describe the cth feature. The white columns indicate the surface object. Figure (**c**) is the attention map extracted by generalized MILCNN. The red, yellow, and blue boxes are the highlighted areas of the aircraft. The subscript number represents the number of aircraft components. The different colors of the columns denote different components. The red, yellow, and blue columns integrate aircraft expressions by MIL pooling function.

2.2. Generalized MIL Fine-Grained Recognition Network

As shown in Figure 7, the architecture of the proposed fine-grained recognition network contains three components: the feature extractor, the instance conversion component, and the MIL pooling component. The feature extractor uses a pretrained DCNN model to transform the fine-grained image into a set of local patches. An extracted patch-level feature represents the characteristic response distribution of fine-grained objects in the spatial domain, which cannot sufficiently characterize aircraft components. Therefore, we design the instance conversion component to accomplish interclass dispersion and intraclass compactness. The construction of the instance loss function is described in Section 2.3. Then, the MIL classifier is leveraged to calculate instance scores, and the MIL pooling function is utilized to aggerate the obtained instance-level scores and calculate class probabilities. The specific content of the MIL pooling part is introduced in Section 2.4.

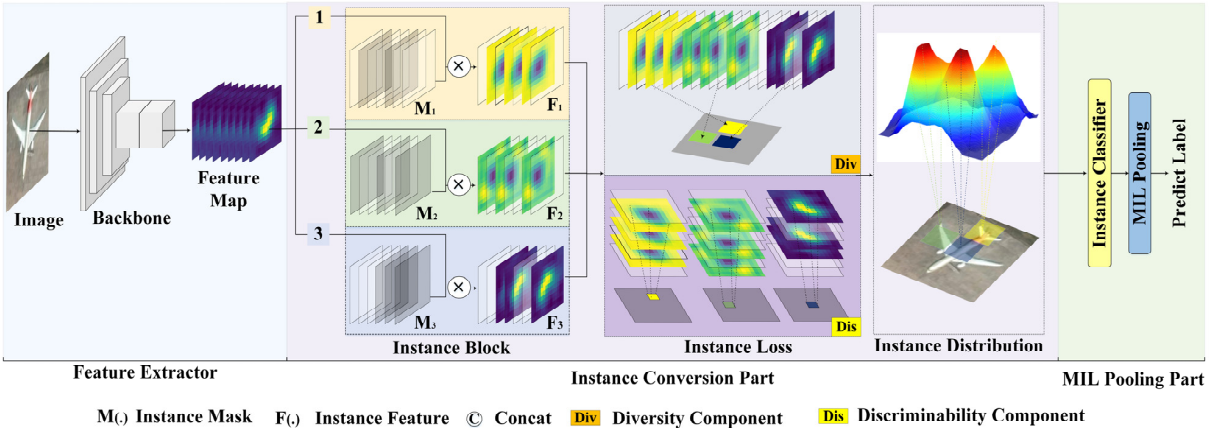


Figure 7. Structure of the generalized MIL network.

2.3. Instance Conversion Part (Instance Loss)

Instance loss function is devised to accomplish the feature conversion from patch level to instance level based on MCL. Given an input image, the feature map extracted by feeding into the backbone is denoted as $F \in \mathbb{R}^{N \times W \times H}$, where N is the number of feature map channels; W is the width of the feature map, and H is the height of the feature map. In MCL function, the feature map channels are equally divided to $c \times \xi$, where c represents the number of categories, and ξ is the set of channels that denote each object. The grouped feature channels corresponding to ith class are indicated by $F_i \epsilon \mathbb{R}^{\xi \times WH}$, $i = 0, 1, \ldots, c-1$. The subcategory channels, separated from one another, denote a particular semantic category without further subdivision into sub-semantics. Additionally, several channels individually expressing different subcategories are provided with similar feature responses, particularly in some fine-grained objects with complex structures. Although the number of categories increases, the number of channels belonging to each category decreases, possibly weakening the characterization ability of feature maps.

Therefore, instance masks are introduced to restrain the learning of different instances instead of directly splitting feature maps. The size of these masks is the same as the output from the backbone network. The architecture of the MIL network with the proposed instance loss function is shown in Figure 8. The channels of the feature map are randomly divided into m groups. The feature map is equally formed as $m \times \eta$, where m denotes the number of instances, and η describes the subset of instance channels that represent the components of fine-grained objects. The index of each group selects different channel combinations to generate instance masks. The weight of the selected ones is adjusted to 1, and the remainder are adjusted to 0. In this way, instance masks of m groups with specific channel suppression can be received. Mathematically, it can be presented as follows:

$$S_i = \left\{ F_{i \times \eta + 1}, F_{i \times \eta + 2}, \ldots, F_{i \times \eta + \eta} \right\} \tag{3}$$

$$H_i = \left\{ I_{i \times \eta + 1}, I_{i \times \eta + 2}, \ldots, I_{i \times \eta + \eta} \right\} \tag{4}$$

$$D_i = K_i - H_i \tag{5}$$

$$SI_i = T_0(D_i) \cap T_1(H_i) \tag{6}$$

$$SF_i = F \times SI_i \tag{7}$$

where S_i denotes the ith selected instance channel set; H_i is the index set of S_i, and K_i represents the array of positive integers from 0 to $N-1$. D_i, a difference set reckoned by K_i and H_i, indicates a set of unselected channels. T_0 implies to set the weight of the channel index D_i to 0, and T_1 stands for adjusting the weights of the channel index H_i to 1. SI_i performs the ith instance channel mask, and SF_i signifies the ith instance feature, $i = 0, 1, \ldots, m$.

Subsequently, $SF = \{SF_0, SF_1, \ldots, SF_{m-1}\}$ is driven by two loss functions for recognizable intentions, which involve the cross-entropy loss and the instance loss:

$$Loss(F) = L_{CE}(F) + \mu \times L_{instance}(F) \tag{8}$$

$$L_{instance}(F) = L_{dis}(F) - \lambda \times L_{div}(F) \tag{9}$$

where L_{CE} is the cross-entropy loss, and $L_{instance}$ intends the instance loss function. μ and λ are both the hyperparameters of instance loss. The hyperparameter setting is identified in Section 3.2.2.

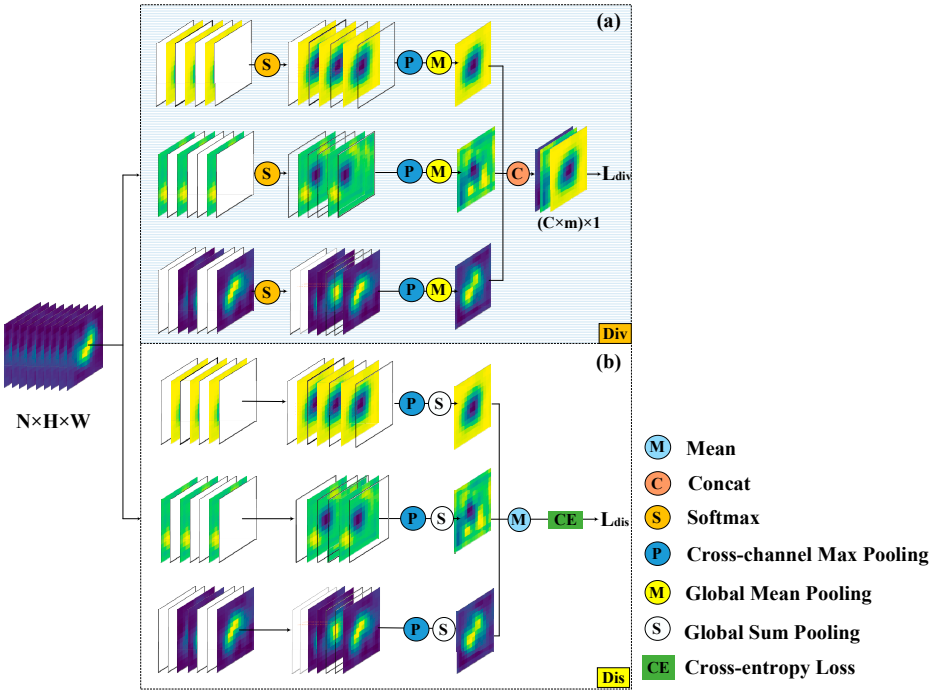


Figure 8. Overview of the instance loss, containing the diversity component and discriminability component. The number of instances is set to 3. In subblocks (**a**,**b**), the output feature map is divided into three groups, and each group denotes the set of instance channels. The flow diagram of the discriminability component is shown in subblock (**a**), and the outline pipeline of the diversity component is displayed in subblock (**b**).

After achieving masked instance particularities, the feature information flows into the discriminability and diversity components. The discriminability component aims to emphasize the feature expression of the m instance blocks. The mathematical formula of the discriminability component is as follows:

$$L_{dis}(F) = \frac{1}{m}\sum_{i=0}^{m-1} L_{CE}\left(y, \frac{\left[e^{g(SF_0)}, e^{g(SF_1)}, \ldots, e^{g(SF_{m-1})}\right]^T}{\sum_{i=0}^{c-1} e^{g(SF_i)}}\right) \tag{10}$$

$$g(SF_i) = \frac{1}{WH}\sum_{k=1}^{WH} \max_{j=1,2,\ldots,\eta} [SF_{i,j,k}] \tag{11}$$

where $g(SF_i)$ contains the global average pooling (GAP), cross-channel max pooling (CCMP), and channel-wise attention (CWA). The structures of CCMP and CWA are the same as those in MCL function [36]. After obtaining instance scores, we calculate the cross-entropy loss value of each instance and get the average instance value as the output of the discriminability component.

The diversity component denotes a distance measurement for feature channels to calculate total similarity, which is gained by CCMP and softmax function. This component focuses on exclusive object regions in instance blocks, rather than all the channels focusing on the most discriminative patch. The calculation formula of the diversity component is as follows:

$$L_{div}(SF_i) = \frac{1}{m}\sum_{i=0}^{m-1} F_{concat}(h(SF_0), h(SF_1), \ldots, h(SF_{m-1})) \tag{12}$$

$$h(SF_i) = \sum_{k=1}^{WH} \max_{j=1,2,\ldots,\eta} \left[\frac{e^{SF_{i,j,k}}}{\sum_{k'}^{WH} SF_{i,j,k}}\right] \tag{13}$$

where SF_i indicates the ith instance feature blocks, and $h(SF_i)$ denotes the inference logits of fine-grained categories. The obtained instance blocks are concatenated to generate instance feature blocks. They calculate the average of multiple instance values to expand the spatial distance among the instance peaks for mining various expressions.

2.4. MIL Pooling Part

After obtaining instance-level features driven by instance loss, we handle the MIL pooling part to anticipate labels, which contain the MIL classifier and MIL pooling function (see Figure 9). In the MIL pooling part, scoring instance blocks are employed by several 1×1 convolutions as an instance classifier. The MIL pooling function aims to aggregate the instance scores into object probabilities. Considering the contribution of each instance, GAP is adopted to reduce the sensitivity to abnormal instance blocks, which assign the equivalent weight of instances to predict labels. The calculation formula is as follows:

$$S_{i,j,c} = f_{conv1\times1\times C}(f_{conv1\times1\times N}(f_{conv1\times1\times N}(SF_{i,j,k})) \quad (14)$$

$$y_c = \frac{1}{W \times H} \sum_{i,j=1}^{WH} S_{i,j,c} \quad (15)$$

where $SF_{i,j,k}$ is the response value of instance in the ith column, jth row, and kth channel. $f_{conv1\times1\times N}$ denotes $1 \times 1 \times N$ convolution operation, and $f_{conv1\times1\times C}$ denotes $1 \times 1 \times C$ convolution operation and the shape of F_c is $N \times H \times W$. N represents the number of feature channels, and C denotes the number of types. $S_{i,j,c}$ indicates the instance score in the ith column, jth row, and cth channel position, and y_c signifies the label score of cth channel position, the shape of which is $1 \times 1 \times C$. H and W represent the height and width of the feature map.

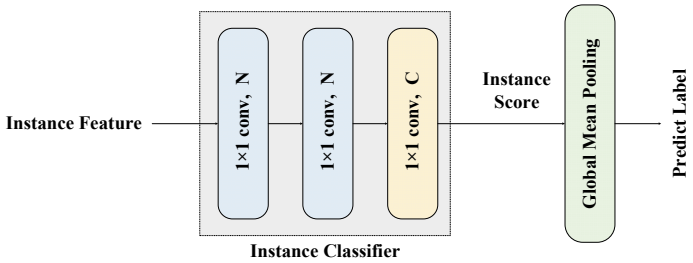


Figure 9. Construction of the MIL pooling component. The MIL pooling component comprises the instance classifier and the MIL pooling function. N represents the number of feature channels, and C denotes the number of aircraft types.

3. Experiments and Results

3.1. Dataset

At present, the public remote sensing image datasets contain a few categories with large intraclass variance, such as ship, tank, and harbor, which are distinguished readily. We collect an aircraft dataset from the 2020 Gaofen Challenge on Automated High-Resolution Earth Observation Image Interpretation [41], DIOR [42], and DOTA [43] to explore the fine-grained study. The aircraft images provided by the Gaofen competition are from GF-2, with spatial resolutions ranging from 0.5 m to 0.8 m (see Figure 10). We select additional types of aircraft images from DIOR and DOTA to extend the Gaofen competition dataset. The distribution of the CAIs, including 15 categories, is shown in Table 1. The fine-grained aircraft images are centrally cropped following the official coordinate annotations. Given the high similarity in the dataset, this dataset is considered the main experimental benchmark dataset.

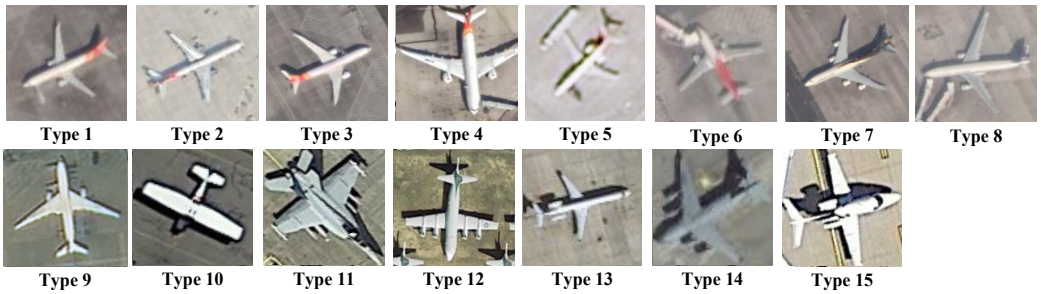


Figure 10. Aircraft samples in combined aircraft images (CAIs).

Table 1. Quantity distribution of combined aircraft images (CAIs).

Types	Images	Types	Images	Types	Images
Type 1	500	Type 2	480	Type 3	480
Type 4	374	Type 5	16	Type 6	500
Type 7	594	Type 8	263	Type 9	570
Type 10	500	Type 11	500	Type 12	500
Type 13	370	Type 14	493	Type 15	500

3.2. Implementation Details

3.2.1. Data Preprocessing

The training samples and testing samples are resized to 224 pixels by interpolation. During the training phase, data augmentation is applied by randomly left-right and up-down flipping and randomly rotating the images based on a set of angles (e.g., 0°, 90°, 180°, and 270°). The ratios of a training set to a test set are 5:5, 6:4, and 8:2 on CAIs.

3.2.2. Parameter Settings

The ImageNet pretrained model (VGGNet-16 and ResNet v2-50) is employed to initialize the backbone part in the recognition network and MILCNN. In CAIs, the optimizer in the model is Adam with a batch size of 32, and the weight decay is 0.0005. The initial learning rate is 0.00005, and the learning rate is multiplied by 0.1 every 30 epochs. The training process ends after 100 epochs. The number of instances m is set to 3; the coefficient of instance loss μ is adjusted to 0.005, and λ is adjusted to 10.

3.2.3. Evaluation Metrics and Experimental Platforms

Some common evaluation criteria are listed as follows to judge the performance of various models quantitatively: overall accuracy (OA) and confusion matrix. The confusion matrix evaluates the algorithm accuracy in each category and misclassification among different sorts. The evaluation reflects the classification performance of the proposed algorithm. OA performs the ratio of the number of correctly recognized sample to the number of test samples, thereby exhibiting the classification performance of the algorithm on the test metric dataset. The calculation formula is as follows:

$$\mathrm{OA} = \frac{\mathrm{N_{true}}}{\mathrm{N}} \tag{16}$$

where $\mathrm{N_{true}}$ denotes the number of the correctly classified samples, and N represents the number of the test images.

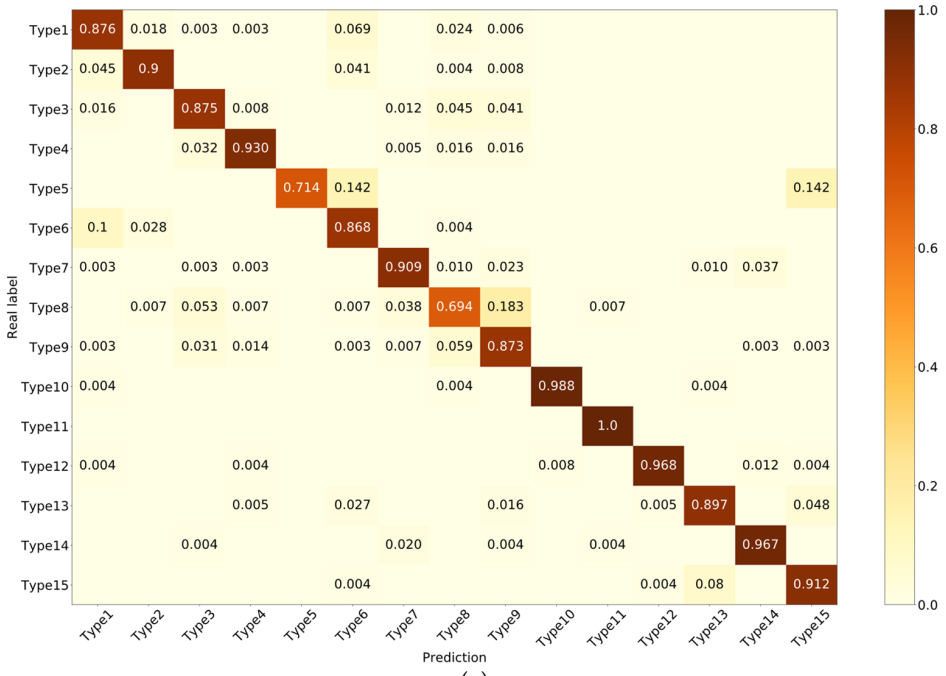
Additionally, we code via the TensorFlow framework and experiment on a workstation with an AMD Ryzen5 3600 CPU and NVIDIA GeForce GTX 2060 GPU.

3.3. Comparative Experiment of Baseline Networks and Standard MIL Networks

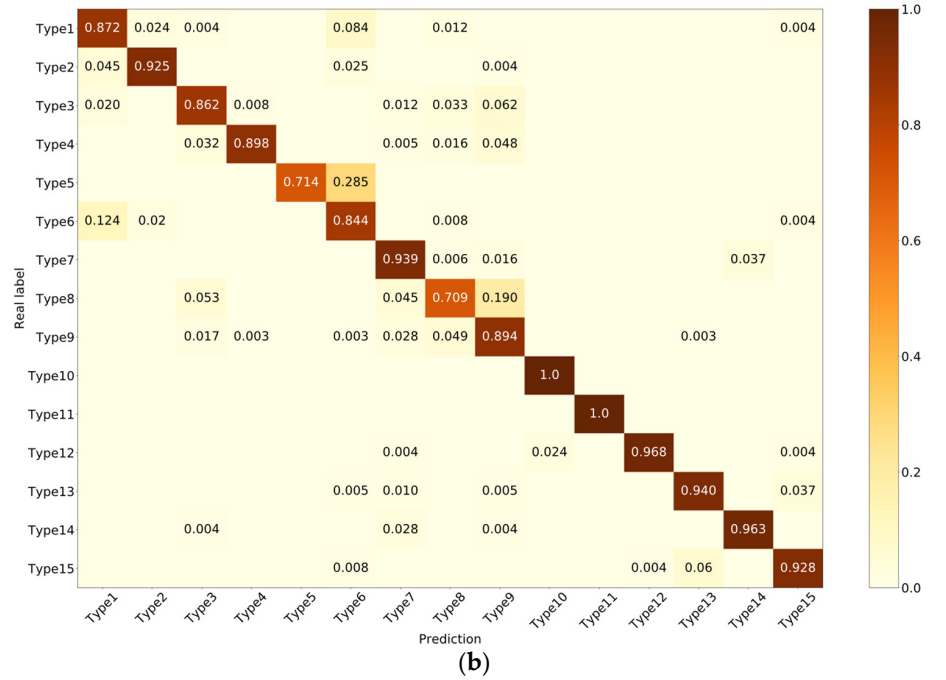
Sufficient experiments have been undertaken to compare the test performance of the baselines and standard MIL networks on the CAIs. The standard MIL network architecture

is the same as that in [33], which contains feature extractor and MIL pooling part. The test results of which are shown in Table 2.

In terms of the results of CAIs in Table 2, all the standard MIL networks outperform baselines in OA with fewer parameters. For instance, the standard MIL networks indicate that accuracy of 91.71%, 92.71%, and 93.58% on VGGNet and 92.01%, 92.94%, and 94.21% on ResNet with 50% training data, which proves the lightweight capability of MIL networks. The confusion matrixes of the baseline networks and the standard MIL networks are indicated in Figure 11a,b, and the per-class OA results on the baselines and standard MIL-VGGNet are shown in Figure 12, which shows that the number of misclassified samples drop.

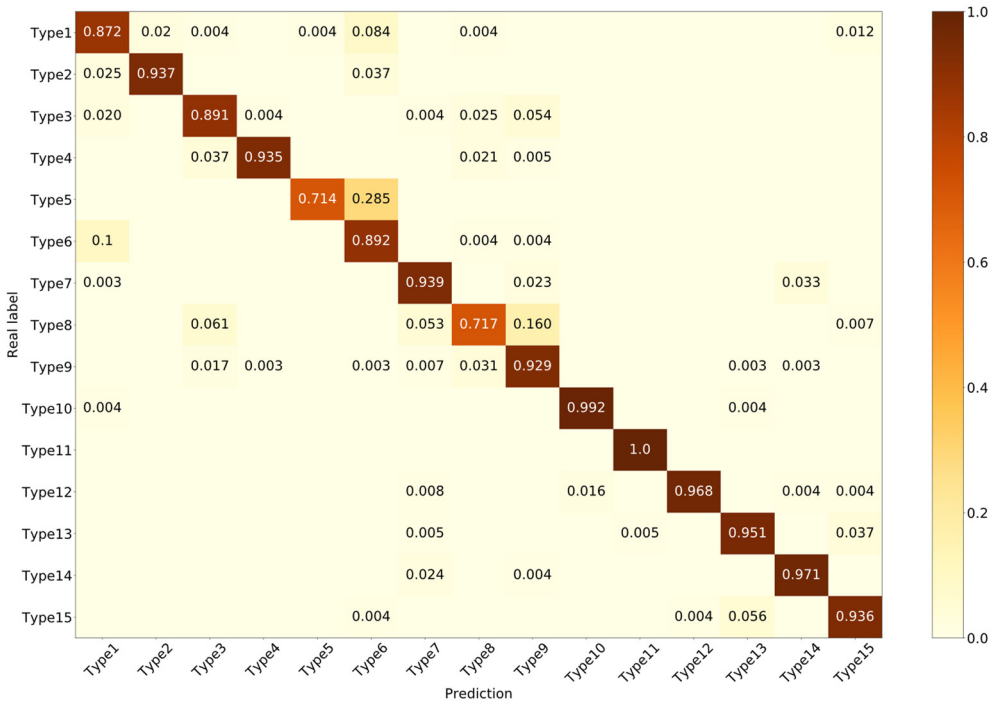


(**a**)



(**b**)

Figure 11. *Cont.*



(**c**)

Figure 11. The confusion matrix results from different recognition frameworks with 50% training data. Figure (**a**) represents the result of VGGNet; Figure (**b**) indicates the result of standard MIL-VGGNet (VGGMIL), and Figure (**c**) denotes the result of generalized VGGMIL. The black value in the figure represents per-class recognition accuracy. The white value in the figure signifies misclassification accuracy. For clarity, the confusion matrices only show the between-class misclassification greater than 0.001.

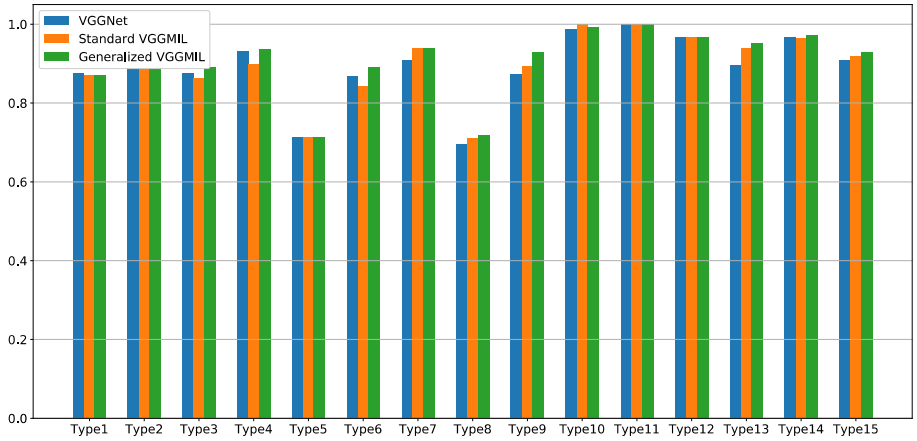


Figure 12. Comparison of the per-class recognition performances of the MIL methods with 50% training data (VGGNet and standard MIL-VGGNet (VGGMIL) and generalized VGGMIL).

Table 2. Experimental results of the different recognition loss function on the CAIs (the ratio of a training set is 50%, 60%, and 80%), which indicates the performance of OA, model parameters, training time, and testing time. Training time indicates the time consumption of model training, and testing time is the time consumption of predicting an image. VGGMIL denotes MIL network, whose backbone network is VGGNet, and ResMIL represents MIL, whose backbone network is ResNet. VGGMIL/ResMIL with CE Loss indicates the standard MIL networks, and VGGMIL/ResMIL with CE Loss and Instance Loss signifies the generalized MIL networks.

Baseline	Loss Function	Accuracy/%			Parameters/MB	Training Time/ms	Testing Time (ms/Image)
		Tr = 50%	Tr = 60%	Tr = 80%			
VGGNet	CE Loss	91.0000	91.5030	92.6038	134.32	5368.96	6.15
VGGNet	CE Loss and MCL	90.8981	90.4834	92.2264	134.32	5694.91	6.20
VGGNet	CE Loss and Instance Loss	90.5666	91.0498	92.9811	134.32	5791.13	6.10
VGGMIL	CE Loss	91.7119	92.7115	93.5849	56.91	**3216.95**	**3.80**
VGGMIL	CE Loss and MCL	92.8270	93.8444	93.3585	56.91	3270.45	3.95
VGGMIL	CE Loss and Instance Loss	**93.0380**	**94.0710**	**94.0377**	**56.91**	3337.66	3.90
ResNet	CE Loss	91.3201	92.7492	94.1887	89.84	2066.15	2.25
ResNet	CE Loss and MCL	92.3749	92.8625	94.2143	89.84	2099.63	2.15
ResNet	CE Loss and Instance Loss	91.1995	92.8625	94.3571	89.84	2438.67	2.25
ResMIL	CE Loss	92.0133	92.9381	94.2143	33.89	**1981.13**	1.85
ResMIL	CE Loss and MCL	92.3448	92.9381	94.4906	33.89	2077.18	1.85
ResMIL	CE Loss and Instance Loss	**92.4653**	**93.2024**	**94.6415**	**33.89**	2253.53	**1.75**

Meanwhile, the largest improvement types are contributed by Type 13 (+4.32%), Type 7 (+3.21%), and Type 2 (+2.50%), and the largest decline types include Type 4 (−3.20%), Type 6 (−2.40%), and Type 3 (−1.25%) in standard MIL-VGGNet. The attention visualization results of increased and dropped types are respectively exhibited in Figures 13 and 14. Considering increased categories, Type 7 has two jets on each side of the wing, which has typical traits compared with other Boeing aircraft. The attention results extracted by the standard MIL networks can capture the key parts of aircraft, which is demonstrated by the fact that, in standard MIL, it is easy to implement rich local semantic modeling to distinguish other subcategories. In terms of dropped categories, Type 3 belongs to large passenger planes without obvious local typical features, which is mainly predicted by the width of the fuselage and the length of the plane. This reflects the fact that the standard MIL is powerless to acquire non-local particularities of aircraft.

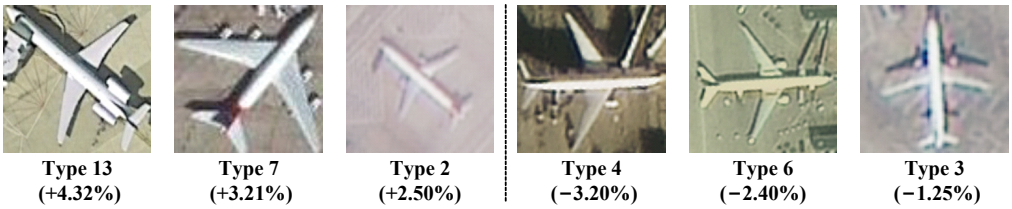


Figure 13. The increased and dropped aircraft samples compared with baselines and standard MIL networks on CAIs.

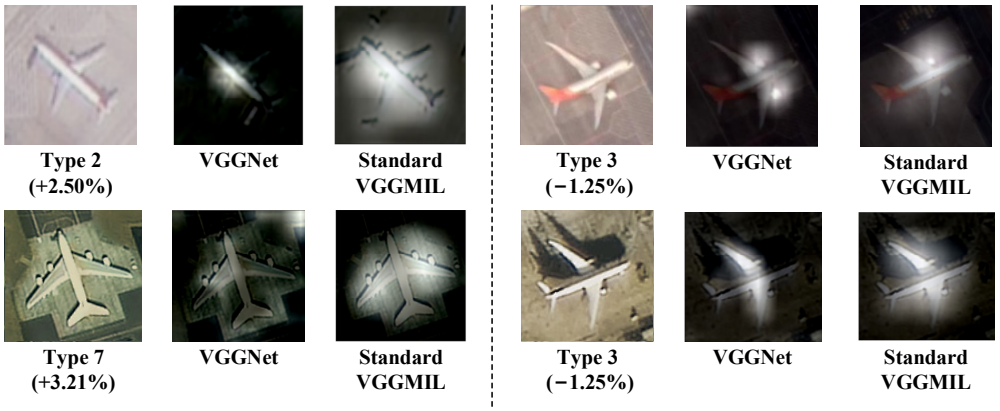


Figure 14. The attention visualization results of the increased and dropped aircraft samples compared with baselines and standard MIL networks on CAIs.

3.4. Comparative Experiment of the Standard MIL Networks and Generalized MIL Networks

Detailed experiments have been attempted to compare the recognition performance of standard MIL networks and generalized MIL networks on the CAIs. The generalized MIL network is composed of a feature extractor, an instance conversion component (instance loss), and an MIL pooling component.

According to the results of CAIs in Table 2, all of the generalized MIL networks outperform standard MIL networks in OA with fewer parameters. For example, the generalized MIL network reaches the best accuracy, 93.04%, 94.07%, and 94.04%, on VGGNet and 92.47%, 93.20%, and 94.64% on ResNet, with 50% training data. The memory parameters and testing time consumption reach the lower level as aircraft test accuracy increases. The confusion matrix results for standard and generalized MIL-VGGNet are displayed in Figure 11b,c to explore the contribution of instance loss. The per-class OA results on the standard and generalized MIL-VGGNet are revealed in Figure 12.

In addition, the largest improvement types are contributed by Type 9 (+5.61%), Type 13 (+5.41%) and Type 7 (+3.03%). The type most in decline is Type 1 (−0.45%) in generalized MIL-VGGNet. The aircraft samples and attention visualization outputs of increased and

dropped types are separately exhibited in Figures 15 and 16. Considering increased categories, the recognition accuracy of Type 7 has been further increased. The attention results of Type 9 in Figure 16 are displayed so that the output of the standard MIL spans larger coverage, which may emphasize the background regions to influence prediction results. The attention results extracted by generalized MIL only focus on the pivotal part of aircraft, which illustrates that generalized MIL networks can capture affluent local typical traits. In terms of dropped samples, Type 1 has a slender body, which has insufficient typical traits compared to other aircraft. The visualization result of Type 1 indicates that the highlighted regions of standard and generalized MIL networks are almost the same, which can further confirm the lack of long-distance feature modeling capability of the generalized MIL methods.

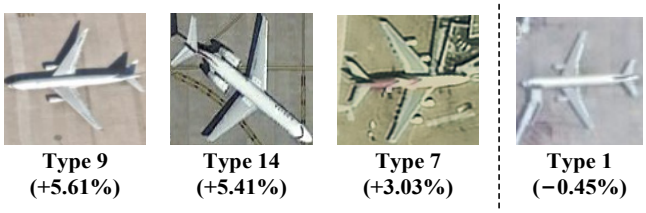


Figure 15. The increased and dropped aircraft samples to compare with standard MIL networks and generalized MIL networks on CAIs.

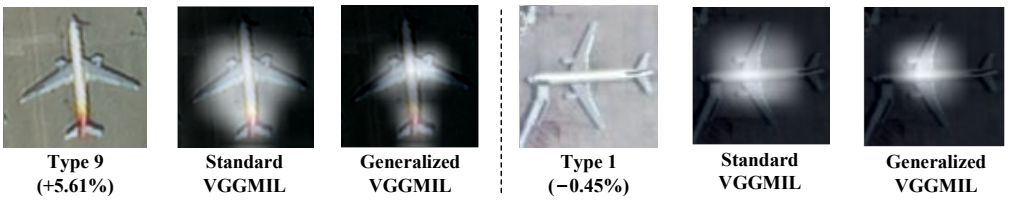


Figure 16. The attention visualization results of the increased and dropped aircraft samples to compare with standard MIL networks and generalized MIL networks on CAIs.

However, for aircraft with highly similar appearance characteristics, our proposed generalized MIL method still leads to some classification errors. Figure 17 exhibits the attention weights of misclassified samples to explore the limitation of the proposed generalized MIL network. The attention weights are obtained by Grad-CAM [40]. The calculation formula is as follows:

$$\omega_k^c = \frac{1}{W \times H} \sum_{i=0}^{W} \sum_{j=0}^{H} \frac{\partial y_c}{\partial A_{ij}^k} \tag{17}$$

$$W_{att} = f_{relu}\left(\sum_{k=0}^{D} \omega_k^c \cdot A^k\right) \tag{18}$$

where ω_k^c indicates the derivative score for the predicted class c and the kth feature map activations of a convolutional layer. y_c denotes the category c score. W_{att} signifies the attention weight, and D represents the number of channels of the feature maps. The gradients flowing back are GAP over the width and height dimensions to obtain the neuron importance weights, which are indicated the importance of the feature map for a target class.

In Figure 17, Type 1 is seriously confused with Type 6. The reason is that both of them are twin-engine turbofan short-range passenger aircraft, and their ratio of the wingspan to the fuselage is approximately 1. Therefore, it is challenging to distinguish Type 1 and Type 6 with global or local features. Furthermore, Type 8 is readily recognized as Type 9. This is because Type 8 and Type 9 both belong to the twin-engine turbofan long-distance aircraft of Boeing. Type 13 and Type 15 are seriously misclassified because they both possess two engines on the wing. The main point to distinguish between Type 13 and Type 15 is the

ratio of the plane and the wing. However, the input images resized to a fixed size may lead to a change in the proportion of aircraft.

In summary, a comparison of the generalized MIL network with other networks indicates that, when the baseline is utilized as feature extractors, the extracted features are an overall representation of an aircraft that may lack local semantics. Therefore, the final test accuracy is still not optimistic, even if some structures are introduced to drive the local semantic learning of objects. Although a single standard MIL network without any tricks is employed for aircraft recognition, which treats the patch-level features as aircraft expression, the extracted patch cannot express the part sub-semantics. However, the generalized MIL network applies several patches to construct aircraft explicit features and introduces the instance loss function to merge the patch-level feature to grab instance-level aggregation mode. In the conversion component of the generalized MIL network, the reason to modify it is that the MCL function randomly selects a few channels to gain the part feature of fine-grained objects. The instance loss function employs masked feature blocks to represent instance features, maximize the adoption of effective channels, and suppress useless background noise.

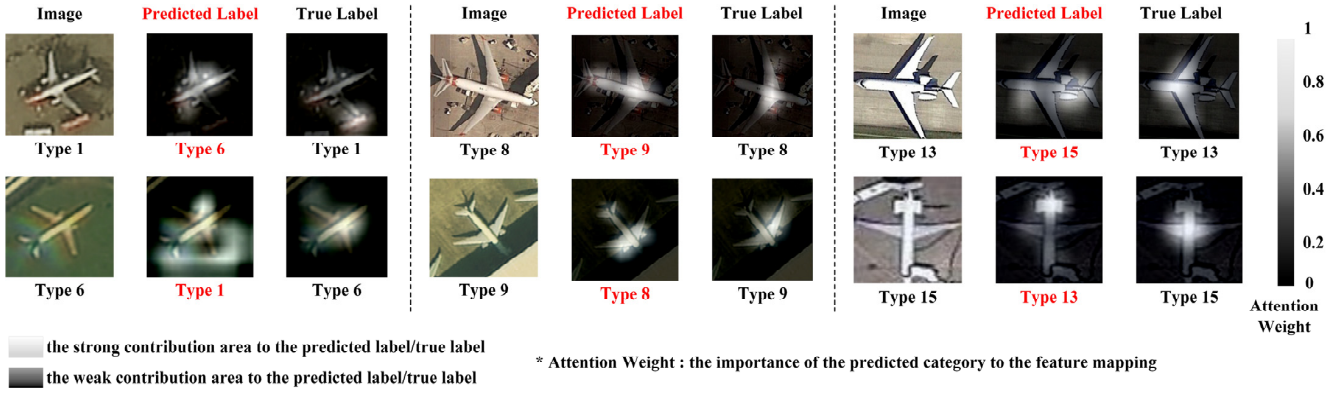


Figure 17. Examples of misclassified test images in one single experiment of CAIs using the proposed generalized VGGMIL network. The misclassified images inference the true label into the predicted label (labeled in red). The attention weights are obtained by Grad-CAM [40], which can explore the importance of the feature map for a predicted label/true label.

Another reason for the test accuracy improvement is the MIL pooling function. After obtaining the scores of multiple instances, this function jointly determines the aircraft label placed on instance sub-semantics. According to the permutation invariance in MIL, an object sample consists of instance concepts, and randomly scrambling the order of instances does not affect the object semantics. Thus, the MIL pooling function, robust to the space rotation characteristics of aircraft, can reduce the prediction error in aircraft recognition.

4. Discussion

In this section, we discuss the number of instances and the visualization results in generalized MIL networks based on CAIs.

4.1. Number of Instances

The aircraft components usually include two wings, a nose, a tail, and other unique structures (such as a high T-tail wing in ARJ21 and turbofan jet engines in Boeing 474). Thus, the number of instances must be discussed. The results of the different number of instances are shown in Table 3. Our proposed method has the best recognition results when m is set to 3. The performance of our pipeline reaches the lowest level when m is adjusted to 4, which implies that the higher the number of the instances performed, the better the recognition performance is. The test accuracy ranks second in Table 3 when m is set to 1. The reason for this phenomenon is that some instance features obtained by the instance extractor cannot motivate a typical aircraft feature result in instance images containing

multiple aircraft parts. The extracted instances cannot be encouraged to learn the most knowledgeable information to reduce recognition accuracy.

Table 3. The ablation experiment in the number of instances. VGGMIL denotes generalized MIL whose backbone network is VGGNet and ResMIL represents generalized MIL whose backbone network is ResNet.

Baseline	The number of Instances m	Loss Function	Accuracy/% (Tr = 50%)	Parameters/MB	Training Time/ms	Testing Time/ms
VGGMIL	1	CE Loss and Instance Loss	92.9475	56.91	3451.21	7.65
VGGMIL	2	CE Loss and Instance Loss	92.8270	56.91	3340.35	7.04
VGGMIL	3	CE Loss and Instance Loss	**93.0380**	**56.91**	**3337.66**	**3.90**
VGGMIL	4	CE Loss and Instance Loss	92.6160	56.91	3402.68	7.44
ResMIL	1	CE Loss and Instance Loss	92.4051	33.89	2253.53	5.18
ResMIL	2	CE Loss and Instance Loss	92.2242	33.89	2341.67	4.72
ResMIL	3	CE Loss and Instance Loss	**92.4653**	**33.89**	**2181.48**	**1.75**
ResMIL	4	CE Loss and Instance Loss	92.0434	33.89	2310.24	4.49

4.2. Visualization

The visualization results of the fifth convolutional feature map exported from the backbone network are shown in Figure 18 to illustrate the capability of the proposed method. The instance number is set to 3. The visualization results indicate that the characters derived from VGGNet exhibit a cramped attention span and few peak areas, and the attention explicit features of which are not disobeyed by multiple local attention modes. The features received from MIL-VGGNet possess discriminative characteristics, which comprise the scope of an aircraft. The attention distribution results reveal a single peak attention pattern. The generalized MIL network captures the salient regions of aircraft and effectively expands the highlight area coverage. Our consequences indicate discriminative and diverse attention mode, which can spotlight aircraft components and capture the different traits of a fine-grained object. Thus, the proposed method can emphasize the optimization of the extracted fine-grained characteristics.

Figure 19 displays the visualization consequences for the distribution of fine-grained attention weights by Grad-CAM. The second column represents the attention allocation generated from VGGNet, the high response values of which are small quantities and sparsely dispersed. The above extracted fine-grained traits cannot be fine-grained component semantics. The third column indicates the response value collected from the standard MIL-VGGNet. The attention range coverage in the third column is much wider than that in the second column, and it can enclose most regions of the fine-grained components. The above-mentioned circumstance proves the explicit characteristics of the objects extracted by MIL networks. However, the attention distribution pattern cannot effectively highlight components. The last column shows the attention output obtained from the generalized MIL-VGGNet, which exhibits the discriminating regions and the separability distributions among fine-grained samples.

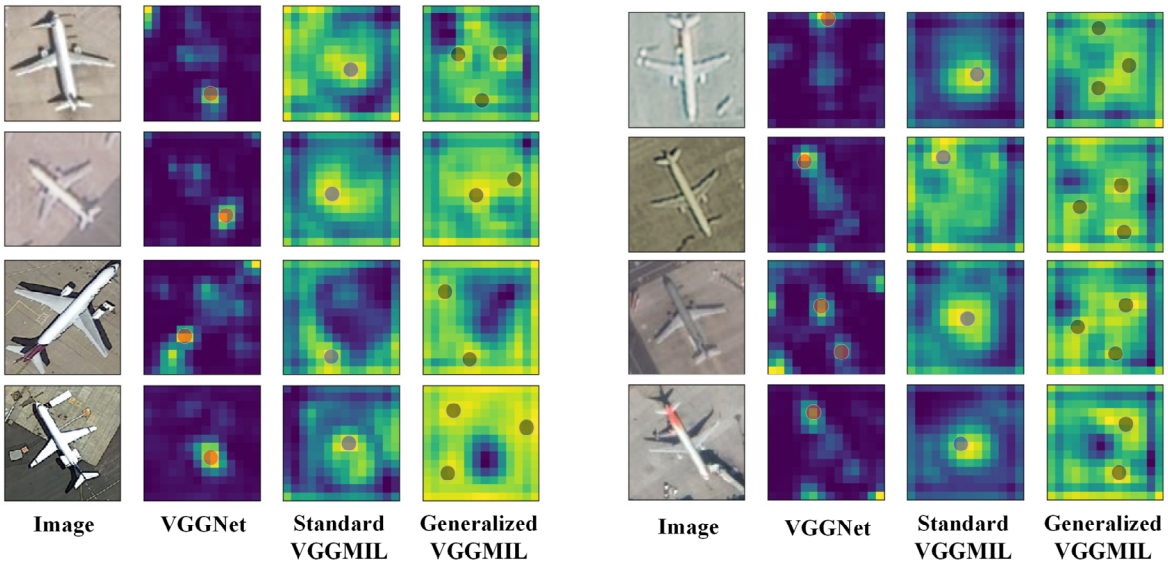


Figure 18. Visualization results of the fine-grained aircraft are generated by VGGNet, standard MIL-VGGNet (VGGMIL), and generalized VGGMIL. The first column is displayed as the fine-grained aircraft images, whereas the second column is the attention map of VGGNet. The visualization results of the standard MIL-VGGNet are indicated in the third column, and the attention feature maps of the generalized MIL-VGGNet are represented in the last column. The solid dots in the attention distribution are indicated to mark the center of the instance regions.

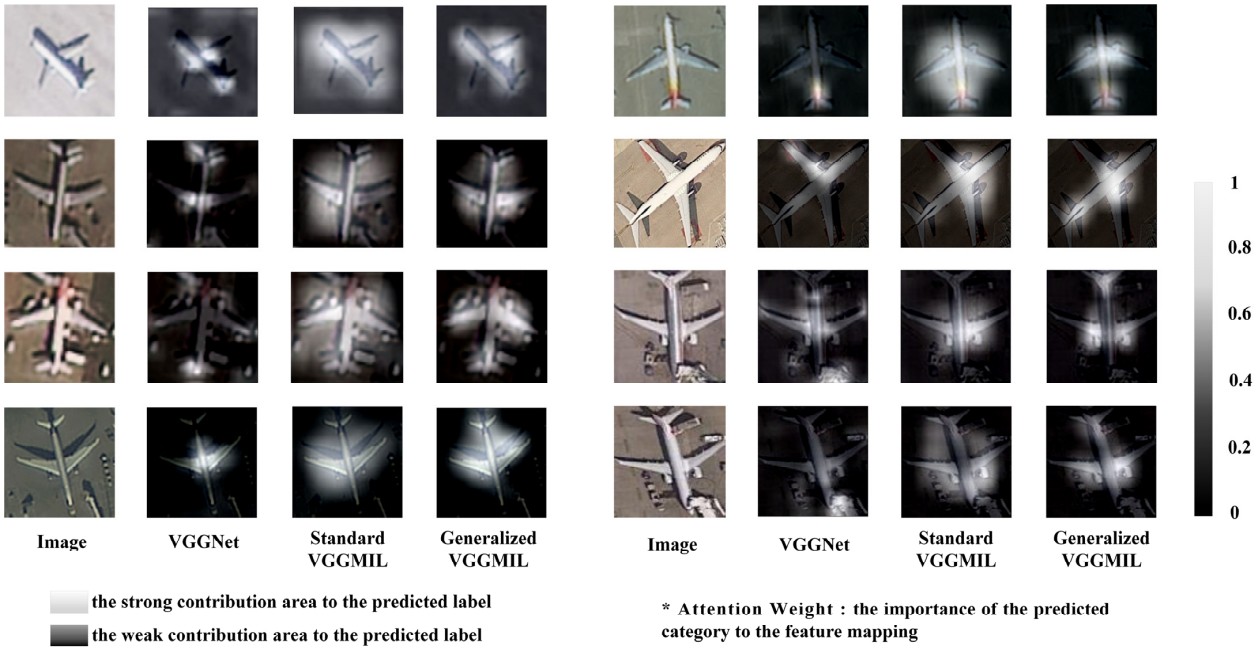


Figure 19. Visualization of the attention weight obtained by Grad-CAM [40] based on a VGGNet-16 model. The first column represents the original image. The second column represents the visualizations of the merged localization regions based on VGGNet-16. The third column represents the visualized results from the standard MIL networks, and the last column represents the results from the generalized MIL networks.

5. Conclusions

Fine-grained aircraft recognition is an essential topic with great practical demand in remote sensing image interpretation. The research focuses on obtaining the component features that can be distinguished from other subcategories. In this study, we introduce

the standard MIL framework, which consists of the feature extractor and the MIL pooling component. It can focus on typical components from a fine-grained object and suppress background noises. However, the feature extractor outputs a patch-level feature, which can result in the scattered distribution of the network interest regions. Thus, the generalized MIL is devised to encourage the learning of discriminative parts and different features among various subcategories without extra model parameters and part annotations.

The generalized MIL network includes a feature extractor, an instance conversion component, and an MIL pooling component. The instance conversion component (instance loss) is composed of discriminability and diversity components. The discriminability component is aimed to establish feature expressions of several instances, and the diversity component is explored to extend the response distribution of different component characteristics. After obtaining instance-level features, the MIL pooling component transforms instance scores to fine-grained aircraft object probabilities. Additionally, the performances of different neural networks on CAIs are compared, and ablation experiments are conducted to substantiate the validity of MIL network and instance loss function.

The use of deep learning for fine-grained image recognition is the mainstream method. Compared with that of the original model, the recognition accuracy of our proposed method is increased. However, it still needs further enhancement. Our pipeline has some defects, that is, the number of instances demands manual adjustment. Therefore, determining automatically the important recognition network hyperparameters is a vital direction for the application of deep learning to improve compatibility for aircraft recognition in future research. Meanwhile, we mainly discuss the reduction of parts annotations in this paper. Nevertheless, the research advancement in remote sensing is impeded by the lack of fine-grained aircraft public datasets. Consequently, considering open set domain or generalized zero-shot learning, using a few generic aircraft images to learn aircraft semantics in remote sensing images, is worth exploring in the future.

Author Contributions: X.H. and K.X. conceived of the study; X.H. collected the dataset, wrote the code, performed the analysis, and wrote the article; K.X. provided suggestions; C.H., C.W. and K.Q. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 41801265, in part by the National Natural Science Foundation of China (NSFC: U2033216) and in part by the National Natural Science Foundation of China (NSFC: 42171448).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The combined aircraft images we processed are obtained from the 2020 Gaofen Challenge on Automated High-Resolution Earth Observation Image Interpretation, DIOR, and DOTA.

Acknowledgments: The authors would like to thank the Aerospace Information Research Institute, Chinese Academy of Sciences, for providing the data used in this study and organizing the Gaofen Challenge on Automated High-Resolution Earth Observation Image Interpretation.

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

References

1. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 207–279.
2. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
3. Dudani, S.A.; Breeding, K.J.; Mcghee, R.B. Aircraft Identification by Moment Invariants. *IEEE Trans. Comput.* **2009**, *100*, 39–46. [CrossRef]
4. Zhang, Y.; Yao, G. Aircraft recognition model based on moment invariants and neural network. *Comput. Knowl. Technol.* **2009**, *14*, 3771–3772.

5. Hsieh, J.W.; Chen, J.M.; Chuang, C.H.; Fan, K.C. Novel aircraft type recognition with learning capabilities in satellite images. In Proceedings of the 2004 International Conference on Image Processing, Singapore, 24–27 October 2004; pp. 1715–1718.
6. Liu, F.; Yu, P.; Liu, K. Aircraft target recognition in remote sensing image using independent component analysis Zernike moments. *CAAI Trans. Intell. Syst.* **2011**, *6*, 51–56.
7. Wang, D.; Xin, H.; Wei, Z.; Yu, H. A method of aircraft image target recognition based on modified PCA features and SVM. In Proceedings of the 2009 9th International Conference on Electronic Measurement & Instruments, Beijing, China, 16–19 August 2009; pp. 261–265.
8. Ke, L.I.; Wang, R.S.; Wang, C. A Method of Tree Classifier for the Recognition of Airplane Types. *Comput. Eng. Sci.* **2006**, *28*, 136–139.
9. Zhu, X.; Ma, B.; Guo, G.; Liu, G. Aircraft Type Classification Based on an Optimized Bag of Words Model. In Proceedings of the 2016 IEEE Chinese Guidance Navigation and Control Conference, Nanjing, China, 12–14 August 2016; pp. 434–437.
10. Zhao, D.; Zhang, Y.; Wei, W. Aircraft recognition algorithm based on PCA and image matching. *Chin. J. Stereol. Image Anal.* **2009**, *14*, 261–265.
11. 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]
12. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
13. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
14. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
15. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
16. Diao, W.; Sun, X.; Dou, F.; Yan, M.; Wang, H.; Fu, K. Object recognition in remote sensing images using sparse deep belief networks. *Remote Sens. Lett.* **2015**, *6*, 745–754. [CrossRef]
17. Zuo, J.; Xu, G.; Fu, K.; Sun, X.; Sun, H. Aircraft type recognition based on segmentation with deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 282–286. [CrossRef]
18. Zhang, Y.; Sun, H.; Zuo, J.; Wang, H.; Xu, G.; Sun, X. Aircraft type recognition in remote sensing images based on feature learning with conditional generative adversarial networks. *Remote Sens.* **2018**, *10*, 1123. [CrossRef]
19. Zhang, N.; Donahue, J.; Girshick, R.; Darrell, T. Part-based R-CNNs for fine-grained category detection. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 834–849.
20. Fu, K.; Dai, W.; Zhang, Y.; Wang, Z.; Yan, M.; Sun, X. Multicam: Multiple class activation mapping for aircraft recognition in remote sensing images. *Remote Sens.* **2019**, *11*, 544. [CrossRef]
21. Xiong, Y.; Niu, X.; Dou, Y.; Qie, H.; Wang, K. Non-locally Enhanced Feature Fusion Network for Aircraft Recognition in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 681. [CrossRef]
22. Wu, H.; Pasad, S. Semi-Supervised Deep Learning Using Pseudo Labels for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2018**, *27*, 1259–1270. [CrossRef]
23. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv* **2017**, arXiv:170301780.
24. Kang, J.; Fernandez-Baltran, R.; Ye, Z.; Xiaohua, T.; Ghamisi, P.; Plaza, A. High-rankness regularized semi-supervised deep metric learning for remote sensing imagery. *Remote Sens.* **2020**, *12*, 2603. [CrossRef]
25. Protopapadakis, E.; Doulamis, A.; Doulamis, N.; Maltezos, E. Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery. *Remote Sens.* **2021**, *13*, 371. [CrossRef]
26. Fang, B.; Li, Y.; Zhang, H.; Chan, J. Semi-supervised deep learning classification for hyperspectral image based on dual-strategy sample selection. *Remote Sens.* **2018**, *10*, 574. [CrossRef]
27. Dietterich, T.G.; Lathrop, R.H.; Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **1997**, *89*, 31–71. [CrossRef]
28. Pinheiro, P.O.; Collobert, R. From image-level to pixel-level labeling with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 1713–1721.
29. Wu, J.; Yu, Y.; Huang, C.; Yu, K. Deep multiple instance learning for image classification and auto-annotation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 3460–3469.
30. Li, D.X.; Peng, J.Y.; Zhan, L.; Bu, Q. LSA based multi-instance learning algorithm for image retrieval. *Signal. Process.* **2011**, *91*, 1993–2000. [CrossRef]
31. Sun, M.; Han, T.X.; Liu, M.-C.; Khodayari-Rostamabad, A. Multiple instance learning convolutional neural networks for object recognition. In Proceedings of the 2016 23rd International Conference on Pattern Recognition, Cancun, Mexico, 4–8 December 2016; pp. 3270–3275.
32. Fan, M.; Chakraborti, T.; Eric, I.; Chang, C.; Xu, Y.; Rittscher, J. Fine-Grained Multi-Instance Classification in Microscopy Through Deep Attention. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging, Iowa City, IA, USA, 3–7 April 2020; pp. 169–173.

33. Li, Z.; Xu, K.; Xie, J.; Bi, Q.; Qin, K. Deep multiple instance convolutional neural networks for learning robust scene representations. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3685–3702. [CrossRef]
34. Ilse, M.; Tomczak, J.; Welling, M. Attention-based deep multiple instance learning. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2127–2136.
35. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively with application to face verification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 539–546.
36. Chang, D.; Ding, Y.; Xie, J.; Bhunia, A.K.; Li, X.; Ma, Z.; Wu, M.; Guo, J.; Song, Y.-Z. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Trans. Image Process.* **2020**, *29*, 4683–4695. [CrossRef] [PubMed]
37. Goodfellow, I.; Warde-Farley, D.; Mirza, M.; Courville, A.; Bengio, Y. Maxout networks. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 7–19 June 2013; pp. 1319–1327.
38. Li, D.X.; Zhao, X.Q.; Li, N. A Survey of Multi-instance Learning Algorithms for Image Semantic Analysis. *Control and Decision.* **2013**, *28*, 481–488.
39. Weidmann, N.; Frank, E.; Pfahringer, B. A two-level learning method for generalized multi-instance problems. In Proceedings of the European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, 22–26 September 2003; pp. 468–479.
40. 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.
41. Gaofen Challenge on Automated High-Resolution Earth Observation Image Interpretation. Available online: http://en.sw.chreos.org (accessed on 1 July 2020).
42. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
43. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.