


Article

EFCMF: A Multimodal Robustness Enhancement Framework for Fine-Grained Recognition

Rongping Zou ^{1,2,3} , Bin Zhu ^{1,2,3,*}, Yi Chen ^{1,2,3}, Bo Xie ^{1,2,3} and Bin Shao ^{1,2,3}¹ College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China² State Key Laboratory of Pulsed Power Laser Technology, Hefei 230037, China³ Key Laboratory of Infrared and Low Temperature Plasma of Anhui Province, Hefei 230037, China

* Correspondence: zhubin@nudt.edu.cn

Abstract: Fine-grained recognition has many applications in many fields and aims to identify targets from subcategories. This is a highly challenging task due to the minor differences between subcategories. Both modal missing and adversarial sample attacks are easily encountered in fine-grained recognition tasks based on multimodal data. These situations can easily lead to the model needing to be fixed. An Enhanced Framework for the Complementarity of Multimodal Features (EFCMF) is proposed in this study to solve this problem. The model's learning of multimodal data complementarity is enhanced by randomly deactivating modal features in the constructed multimodal fine-grained recognition model. The results show that the model gains the ability to handle modal missing without additional training of the model and can achieve 91.14% and 99.31% accuracy on Birds and Flowers datasets. The average accuracy of EFCMF on the two datasets is 52.85%, which is 27.13% higher than that of Bi-modal PMA when facing four adversarial example attacks, namely FGSM, BIM, PGD and C&W. In the face of missing modal cases, the average accuracy of EFCMF is 76.33% on both datasets respectively, which is 32.63% higher than that of Bi-modal PMA. Compared with existing methods, EFCMF is robust in the face of modal missing and adversarial example attacks in multimodal fine-grained recognition tasks. The source code is available at <https://github.com/RPZ97/EFCMF> (accessed on 8 January 2023).



Citation: Zou, R.; Zhu, B.; Chen, Y.; Xie, B.; Shao, B. EFCMF: A Multimodal Robustness Enhancement Framework for Fine-Grained Recognition. *Appl. Sci.* **2023**, *13*, 1640. <https://doi.org/10.3390/app13031640>

Academic Editors: Jing Zhang, Jipeng Qiang and Cangqi Zhou

Received: 15 October 2022

Revised: 18 January 2023

Accepted: 19 January 2023

Published: 27 January 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: fine-grained recognition; multimodal; modal missing; adversarial examples

1. Introduction

The purpose of fine-grained recognition is to distinguish subordinate categories (like owls, albatrosses, and seagulls in birds) with subtle differences in the same primary category (such as birds [1], Flowers [2], dogs [3], cars [4], and fruits [5]). These are applied to real-world scenes in different fields, such as species identification, vehicle identification, product identification [6,7] and so on. Since subcategories are all similar to each other, different subcategories can only be distinguished by subtle and subtle differences, which makes fine-grained identification a challenging problem.

Many fine-grained recognition methods have been proposed, which can be divided into two categories on a single visual modality. (1) One is a strongly-supervised method based on a localization and classification subnetwork, and the other is a weakly-supervised method for end-to-end feature encoding [8]. In intensely supervised methods, techniques such as object detection [9–11] or segmentation [12,13], can be used to locate parts of objects with crucial fine-grained features and enhance the effect of recognition, such as the use of segmentation models for assisted classification of part-stacked [14], and part-based RCNN [15] with detection models. (2) The other is weakly-supervised methods. In weakly-supervised methods, most of the classical classification networks such as ResNet [15], DenseNet [16] and other backbone structures are used as feature extraction models, among which VggNet [17] is used to construct dual-stream branches and fuse

them. The BCNN [18] and MOMN [19] methods are based on the BCNN method. These methods focus on improving the classification accuracy of the visual modality but are easily limited by a single visual modality.

Recently, some methods for fine-grained recognition based on multimodal data have been proposed. There are three data fusion methods in the data fusion modalities: vision and language, vision and speech, and vision and knowledge. Among them, the fusion method of vision and language is represented by CVL et al. [20–22]. In addition to using the two modalities of vision and language for fusion, Zhang et al. [23] also conducted corresponding research on vision and sound and fused the two modalities of vision and sound. In addition, related works introduce knowledge information [24–26] and fuse it with visual modalities. These methods effectively improve the accuracy of fine-grained recognition tasks by fusing data from multiple modalities. But its robustness in the face of modal missing and adversarial examples attacks is not well considered.

However, in practical fine-grained recognition tasks, modal missing and adversarial examples are often encountered, leading to the models based on multimodal data needing to be fixed. The reason for the modal missing is that in the data acquisition process, several modalities in a small part of the data need to be included due to factors such as instrument failure. Adversarial examples attacks refer to the unusual noise generated by the adversarial example method [27–32] that makes the model prone to fatal errors.

There are many application scenarios in which visual and language modalities exist in the actual usage process. For example, a product or a movie introduction often has visual modal information, such as pictures and videos, and language modality information, such as keywords and brief descriptions. The scenarios mentioned above are often prone to missing modalities. The standard solution is to train multiple models to cope with scenarios with only one modality, which is often more expensive. The EFCMF framework is proposed in this paper to utilize multimodal data better and reduce the cost.

This paper builds a multimodal fine-grained recognition framework EFCMF of visual language fusion with the same accuracy as the existing methods to solve the above problems. The framework adopts the technique of random modality deactivation for training while ensuring that original fusion accuracy remains unchanged. In this way, the model acquires the ability to cope with the modal missing and adversarial examples attacks, and dramatically improves its accuracy when attacked by adversarial examples.

The proposed framework's contributions are as follows: (1) The framework can deal with the modal missing problem without training additional single-modal fine-grained recognition models. (2) The framework can take advantage of multimodal data without adversarial training. The model accuracy is guaranteed to stay the same by using the modality that has not suffered from adversarial examples attacks. (3) Through a large number of experiments, hyperparameters to guide the use of random deactivation training methods are given.

2. Method

To address the problem of modal missing and adversarial example attacks in the multimodal fine-grained recognition task of vision and language fusion. The framework adopted in this study is shown in Figure 1, which consists of three parts: a visual feature extraction module, a language feature extraction module, and a feature fusion module.

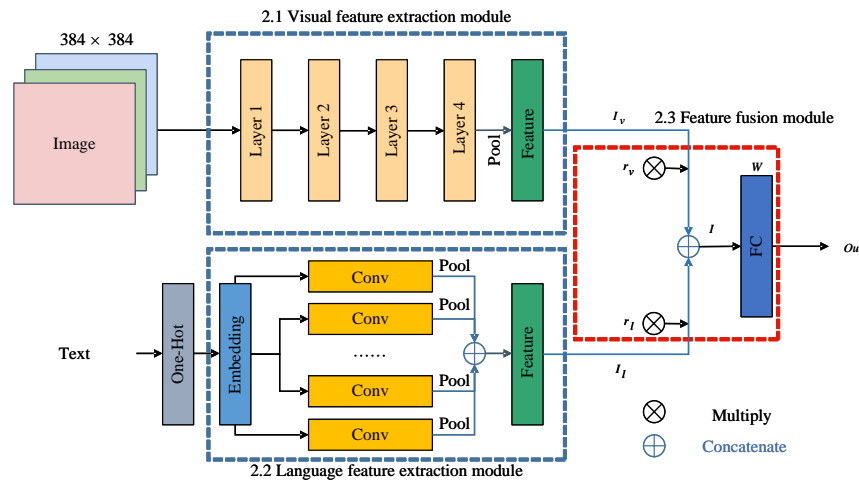


Figure 1. Structure of Enhanced Framework for Multimodal Feature Complementarity Based on the Modal Feature Random Deactivation Training Method.

2.1. Visual Feature Extraction Module

In the visual feature extraction module, the backbone network for feature extraction [33] is composed of four transformer layers. The input image is subjected to high-dimensional mapping of four feature extraction modules, and a high-dimensional vector I_v that can express the image features is obtained. The the input of the visual modality is represented as $Img_{3 \times 384 \times 384}$. The i -th layer of the network is denoted as $Layer_i$, and the number of channels of the output feature $Feat_i$ is c , and the size is $h \times w$. Then, the entire visual feature extraction module can be expressed as:

$$Layer_{1-4}(X) = Layer_4(Layer_3(Layer_2(Layer_1(X)))) \tag{1}$$

Its forward propagation process can be expressed as follows:

$$Feat_4_{1536 \times 144 \times 1} = Layer_{1-4} \left(\begin{matrix} Img \\ 3 \times 384 \times 384 \end{matrix} \right) \tag{2}$$

After the image is subjected to high-dimensional mapping through four feature layers, a feature matrix with 1536 channels is obtained, and then the output features are globally pooled, and the pooled features are deformed to obtain a feature vector I_v with a length of 1536.

The backbone network of language modalities consists of several convolutional layers [34] (denoted as *Conv*). Its forward propagation process can be expressed as follows:

2.2. Language Feature Extraction Module

The input text (denoted as $Text = [Word_1, \dots, Word_n]$) is a sequence of n words, and encodes words by constructing a dictionary to map words to the integer domain:

$$[num_1, \dots, num_n] = [Encode(Word_1), \dots, Encode(Word_n)] \tag{3}$$

Then, the one-hot encoding method is used to map from the integer domain to the sparse vector space, and the process is as follows:

$$\begin{bmatrix} Vec_1, \dots, Vec_n \\ n \times 1 \end{bmatrix} = [OneHot(num_1), \dots, OneHot(num_n)] \tag{4}$$

After that, take an embedding layer with weight W_E , map the sparse vector to a dense vector of 50 dimensions, and reshape it into a 3D embedding matrix Mat :

$$Mat_{1 \times n \times 50} = reshape \left(\left[\begin{matrix} W_E \times Vec_1, \dots, W_E \times Vec_n \\ 50 \times n & n \times 1 & 50 \times n & n \times 1 \end{matrix} \right] \right) \tag{5}$$

Next, feature extraction (represented as $Feat$) is performed by k convolution kernels with m convolutional layers of size $h \times w$ (the i -th convolutional layer is denoted as $Conv_i$), and the obtained features are pooled:

$$Feat_i_{k_i \times 1 \times 1} = Pool \left(Conv_i \left(Mat_{1 \times n \times 50} \right) \right) \tag{6}$$

2.3. Feature Fusion Module

Finally, all the features of the output are concatenated (represented as cat) and reshaped to obtain the features I_l for fusion:

$$Feat_{\sum_i^m k_i \times 1 \times 1} = cat \left(\left[\begin{matrix} Feat_1, \dots, Feat_m \\ k_1 \times 1 \times 1 & \dots & k_m \times 1 \times 1 \end{matrix} \right] \right) \tag{7}$$

$$I_l = Feat_{\sum_i^m k_i} = reshape \left(Feat_{\sum_i^m k_i \times 1 \times 1} \right) \tag{8}$$

In the feature fusion module, the study adopts a feature-level fusion strategy to obtain new features by concatenating the features of the visual modality and language modality

$$\begin{bmatrix} I_v \\ I_l \end{bmatrix} = cat(I_v, I_l) \tag{9}$$

The error is backpropagated through the classifier to jointly training the two feature extraction modules of the model.

The weight of the fully connected layer of the module is W , which can be regarded as the splicing of two weights using modal features for classification: $W = cat(W_v, W_l) = [W_v \mid W_l]$, and the matrix product with the feature to obtain the Out :

$$Out = W \times I = [W_v \mid W_l] \times \begin{bmatrix} I_v \\ I_l \end{bmatrix} \tag{10}$$

During training, in order to simulate modal missing to enhance the ability of the model to cope with modal missing and adversarial examples attacks, the input modal features are randomly deactivated. The random inactivation of features is a random event conforming to the Bernoulli distribution, and its Bernoulli random variable is $r_m \sim \text{Bernoulli}(p)$. When a modality is missing, the visual modality has a q probability of remaining intact, and the Bernoulli random variable for this event is $r_v \sim \text{Bernoulli}(q)$, and the language modality is $r_l = 1 - r_m r_v$. Then, the forward propagation process of the module is as follows:

$$I_v = r_v * I_v, I_l = r_l * I_l \tag{11}$$

$$I = cat(I_v, I_l) = \begin{bmatrix} I_v \\ I_l \end{bmatrix} \tag{12}$$

The output of the model appears as follows:

$$\begin{aligned}
 Out &= W \times I = [W_v \mid W_l] \times \begin{bmatrix} I'_v \\ I'_l \end{bmatrix} \\
 &= (1 - r_m(1 - r_v)) * W_v \times I'_v + r_l * W_l \times I'_l \\
 &= (1 - r_m(1 - r_v)) * Out_v + r_l * Out_l \\
 &= (1 - r_m(1 - r_v)) * Out_v + (1 - r_m r_v) * Out_l \\
 &= (1 - r_m) * Out + r_m(r_v * Out_v + (1 - r_v) * Out_l)
 \end{aligned} \tag{13}$$

The average expectation of the output is:

$$E(Out) = (1 - p) * Out + p * (q * Out_v + (1 - q) * Out_l) \tag{14}$$

At this time, the model can be considered as an ensemble model of three models with strong correlation. *Out* is a vector representing the confidence (denoted as *Conf_i*) of the classification result of the model for a total of *k* categories: *Out* = [*Conf₁*, ..., *Conf_i*] Finally, the output result is normalized by softmax to obtain the predicted value *Pred*, and the *Loss* is calculated as follows:

$$\begin{aligned}
 Pred &= \text{softmax}(Out) = \frac{1}{\sum_1^k e^{Conf_i}} [e^{Conf_1}, \dots, e^{Conf_i}] \\
 Loss &= Label * \log(Pred) + (1 - Label) * \log(1 - Pred)
 \end{aligned} \tag{15}$$

It can be inferred from the Formula (15) that when *p* = 0, the output expectation of the model is *E(Out)* = *q* * *Out_v* + (1 - *q*) * *Out_l*. At this time, if *q* = 0.5, it can be considered that *Out* is equal to the average of the outputs *Out_v* and *Out_l* of the visual modal classifier and the language modal classifier:

$$E(Out) = \frac{1}{2} (Out_v + Out_l) \tag{16}$$

Then, the model becomes a normal feature fusion model, and the average output of the model is expected

$$E(Out) = Out = W \times I \tag{17}$$

Since the error rate and correlation of the ensemble model are negatively correlated, we hope this method can weaken the correlation between submodels and improve the complementarity between modal features to improve the model's accuracy and robustness. This study performs related experiments in next Section to find the appropriate hyperparameters to optimize the model.

3. Experiment

3.1. Experimental setting

To verify the feasibility and universality of the method in this study, the experimental details are as follows:

The visual feature extraction module of the model has been pretrained on the ImageNet dataset [35], and the features output by the last layer of convolutions are pooled for feature-level fusion. The data of the visual modality adopt The Birds dataset [1] and Flowers dataset [2], the width and height of the image are scaled to 384 pixels, and some of the grayscale images are copied and synthesized into a three-channel image. Finally, the image is normalized using ImageNet data processing. The language modality adopts the text description extended by Reed et al. [36]. Because there are certain errors in the text description, the spelling correction operation is advanced on the text, and then lemmatizes words to form a new dictionary. Finally a 50-dimensional embedding layer is used for embedding.

The deep learning framework PyTorch [37] and the adversarial example toolbox torchattacks [38] are used in the experiments. The training parameters of each module of the model are shown in the Table 1.

Table 1. The training parameters of each module of the model.

	Optimizer	Learning Rate	Weight Decay	Dropout	Batch Size	Epoch
Vision	SGD	0.005	0.00001	0.5	8	50
Language	RAdam	0.01	0.0001	0.5	32	50
Fusion	SGD	0.005	0.00001	0.5	8	50

The parameters for generating adversarial examples are shown in the Table 2.

Table 2. Generating parameters for adversarial examples using torchattacks.

	Eps	Alpha	Steps	c	Kappa	lr	Random_Start
FGSM	0.014	/	/	/	/	/	/
BIM	0.01568	0.00392	0	/	/	/	/
PGD	0.03	0.00784	10	/	/	/	/
C&W	/	/	100	0.0001	0	0.1	True

3.2. Experiment Results

3.2.1. Analysis of the Robustness and Accuracy Performance of the Model

In this study, the robustness of the model is tested by attacking the model with adversarial examples and causing modal missing (deactivating the features of each modality of the model). The robustness results of the model are shown in Table 2. Accuracy represents the model's accuracy when adversarial examples do not attack it, and the modality is not missing. FGSM, BIM, PGD, and C&W are adversarial examples of attack methods. Vision Missing means visual modal features are missing (zeroing the input features of the visual modality), and Language Missing represents the absence of modal language features (zeroing out the input features of language modality).

The proposed method has excellent accuracy advantages compared with Bi-modal PMA. It can surpass the fine-grained recognition method of Bi-modal PMA in the Birds and the Flower datasets. At the same time, the method's robustness in this study is also excellent, and in the face of adversarial sample attacks, EFCMF can exceed Bi-modal PMA such as FGSM, BIM, PGD, and C&W in most cases. In the face of FGSM attacks, EFCMF's accuracy in the Birds dataset can exceed Bi-modal PMA by about 16% and on the Flowers dataset by about 70%. In addition, EFCMF performed well in the face of BIM attacks and PGD attacks, surpassing Bi-modal PMAs by about 9% on the Birds dataset and about 70% on the Flowers dataset. When attacking with powerful adversarial examples method PGD, the accuracy of Bi-modal PMA on both datasets has been reduced to the lowest point, which is lower than that of random decider (0.0050). However, EFCMF is still able to exercise some judgment, still having an accuracy of 0.0642 on the Birds dataset and 38% accuracy on the Flowers data set. The case of C&W attacks is unique, and Bi-modal PMA has high accuracy on the Birds dataset. Inactivation, which is a critical reason EFCMF can cope well with missing modality. In order to verify that random modal inactivation can effectively improve the model's ability to cope with modal loss, this study retrains Bi-modal PMA using this method. The model test results are shown in Table 3, and it can be seen that after random mode deactivation training, the model's ability to cope with modal loss can be effectively improved. After training with this method, Bi-modal PMA improved the model accuracy by 15% and 40% under the absence of visual modality in the Birds and Flowers datasets, respectively. At the same time, the accuracy of Bi-modal PMA in the face of FGSM and BIM counterattack attacks has also been improved. Random modality deactivation, which is a critical reason EFCMF can cope well with missing modality. In

order to verify that random modal deactivation can effectively improve the model's ability to cope with modal loss, this study retrains Bi-modal PMA using this method. The model test results are shown in Table 4, and it can be seen that after random mode deactivation training, the model's ability to cope with modal loss can be effectively improved. After training with this method, Bi-modal PMA improved the model accuracy by 15% and 40% under the absence of visual modality in the Birds and Flowers datasets, respectively. At the same time, the accuracy of Bi-modal PMA in the face of FGSM and BIM counterattack attacks has also been improved.

Table 3. The model's accuracy in the face of adversarial example attacks and modalities is missing.

		Accuracy	FGSM	BIM	PGD	C&W	Vision Missing	Language Missing
Birds [1]	Bi-modal PMA	0.8870	0.2260	0.1350	0.0013	0.7433	0.0486	0.7899
	EFCMF ($p = 0.8$, $q = 0.4$)	0.9114	0.3826	0.2231	0.0642	0.5730	0.5051	0.9099
Flowers [2]	Bi-modal PMA	0.9700	0.1503	0.0488	0.0019	0.7509	0.0098	0.8999
	EFCMF ($p = 0.8$, $q = 0.4$)	0.9931	0.8882	0.7656	0.3803	0.9509	0.6441	0.9941

Table 4. Comparison of the accuracy of the Bi-modal PMA method trained by the random modality deactivation method.

		FGSM	BIM	PGD	C&W	Vision Missing	Language Missing
Birds	Bi-modal PMA	0.2260	0.1350	0.0013	0.7433	0.0486	0.7899
	Bi-modal PMA ($p = 0.8$, $q = 0.4$)	0.3386	0.1824	0.0043	0.7239	0.2057	0.7297
Flowers	Bi-modal PMA	0.1503	0.0048	0.0019	0.7509	0.0098	0.8999
	Bi-modal PMA ($p = 0.8$, $q = 0.4$)	0.2109	0.0371	0.0000	0.6582	0.4362	0.8735

EFCMF has more accuracy advantages than Bi-modal PMA and higher accuracy than other fine-grained recognition models. As shown in Table 5, it can be seen that the EFCMF method has higher accuracy than the multimodal fine-grained recognition models such as CVL, TA-FGVC, KERL, and KGRF. Furthermore, EFCMF has higher accuracy in single-modal fine-grained recognition methods such as Inception-v3 [39], ViT-B [40], and PART [41].

Table 5. Comparison of the accuracy of each method.

Method	Data Field	Birds [1]	Flowers [2]
CVL [21]	Vision+Language	0.8555	\
TA-FGVC [22]	Vision+Language	0.8810	\
KERL [25]	Vision+Knowledge	0.8700	\
KGRF [26]	Vision+Knowledge	0.8849	\
Bi-modal PMA [42]	Vision+Language	0.8870	0.9740
Inception-V3 [39]	Vision	0.8960	0.9737
ViT-B [40]	Vision	\	0.9850
PART [41]	Vision	0.9010	\
EFCMF (ours)	Vision+Language	0.9114	0.9931

3.2.2. Performance of the Model under Different Hyperparameters

In this study, the two hyperparameters p and q were performed 36 experiments on each dataset at intervals of 0.2 from 0 to 1. In order to investigate the effects of hyperparameters p and q on the model’s robustness and accuracy, the experimental results are plotted in this study as heat maps shown in Figures 2–4.

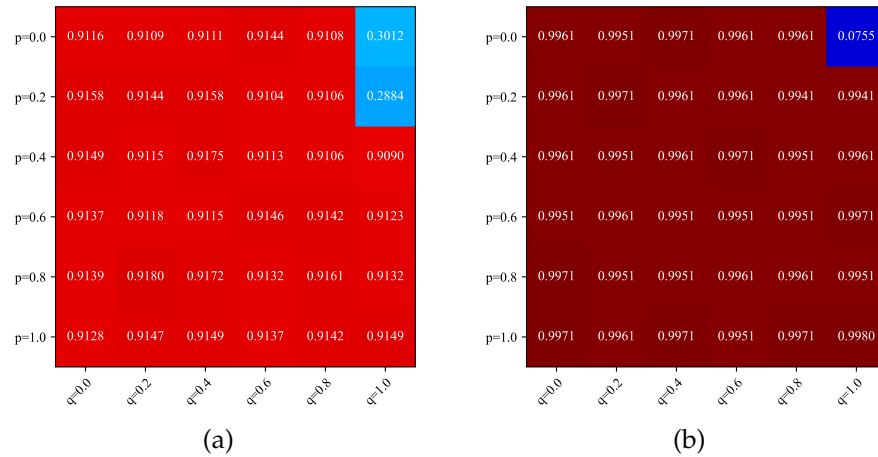


Figure 2. Heat map of model accuracy for different values of p and q . (a) Birds. (b) Flowers.

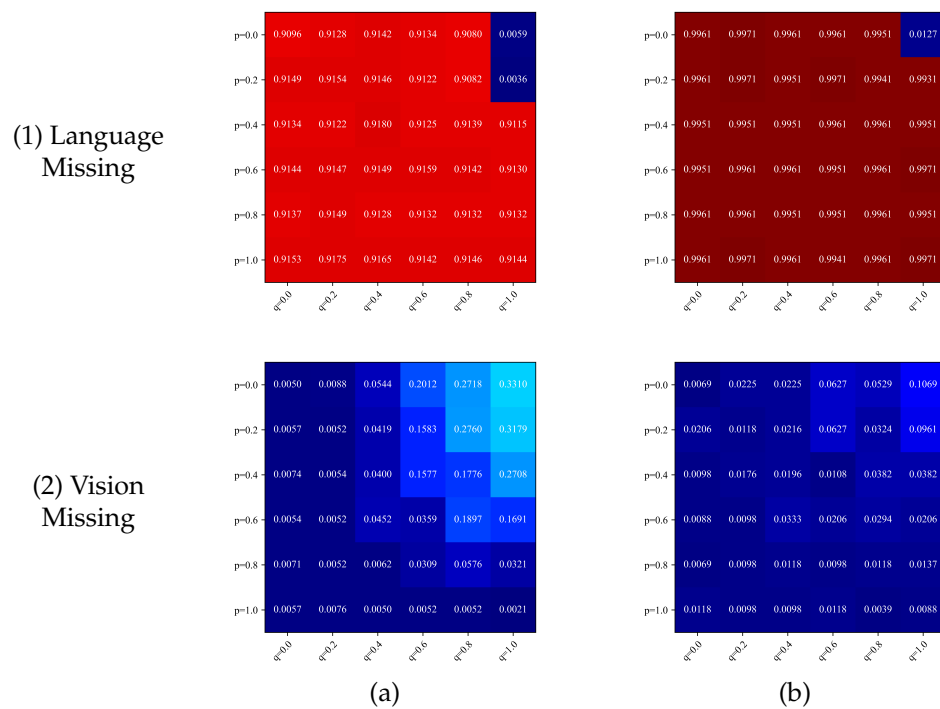


Figure 3. Heat map of accuracy of model facing modal missing for different values of p and q . (a) Birds. (b) Flowers.

Figure 2 shows the model’s accuracy for different values of p and q . The most special cases are at $p = 1, q = 0$, and $q = 0.2$. The reason is that for $p = 1$, the model only has data from the visual or language modality for each input sample. When $q < 0.2$, the probability of missing data of visual modality is higher, and the model is mostly training the extraction module of language modality at this time. $q < 0.2$ makes the model use only the information of linguistic modality for fine-grained recognition, and therefore the accuracy is lower.

Figure 3 shows the heat map of the accuracy of the model in the face of the missing modality. It can be seen that the accuracy of the model increases with the increase of the modal deactivation probability p and the decrease of the visual modality integrity probability q and stays at a low level when facing the missing visual modality. The above phenomenon is related to the degree of training in the language modality feature extraction module. The higher the training degree of the linguistic modality feature extraction module, the better the model can face the visual modality missing.

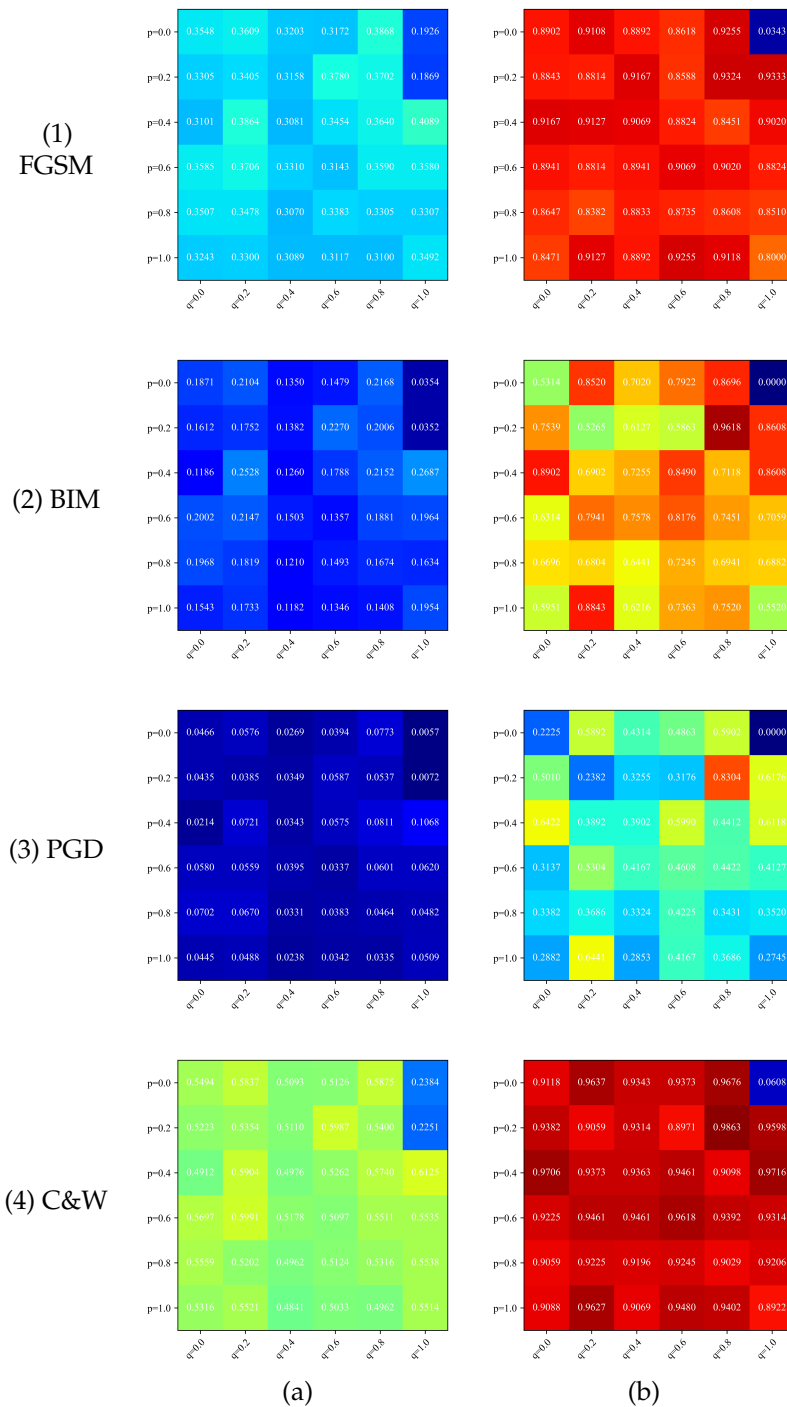


Figure 4. Heat map of the accuracy of the model against the adversarial examples attack with different values of p and q . (a) Birds. (b) Flowers.

Figure 4 shows the model’s accuracy in the face of the adversarial example attack under different values of p and q . The p and q parameters not only have a particular influence on the model’s ability to cope with the modal missing but also can impact the robustness of the adversarial sample attack. As shown in Figure 4b, with the appropriate selection of p and q parameters, the trained model has strong robustness and can maintain high accuracy against examples attacks.

In order to better demonstrate the effects of parameters p and q on the model, the data in Figures 2–4 are averaged on each axis to obtain the accuracy trend plots in Figures 5 and 6, respectively, and the results are given in the following.

(1) Analysis of the effect of the modality deactivation probability p on the model

Figure 5 shows the average model accuracy in various cases for the matrix data shown in Figures 2–4 for each p -value obtained by averaging over q -values. It can be seen that the impact of p -values in the face of adversarial sample attacks is similar for both datasets, with curves having extreme value points at $p = 0.2$ as well as $p = 0.8$ and a minimal value point at $p = 0.4$. The extreme value point implies that the random modal deactivation impacts the model’s robustness relative to the case when $p = 0$ is not performed randomly and can enhance the model’s ability to cope with counter-sample attacks at the appropriate value. Tables 6 and 7 correspond to the data in Figure 5a,b, respectively.

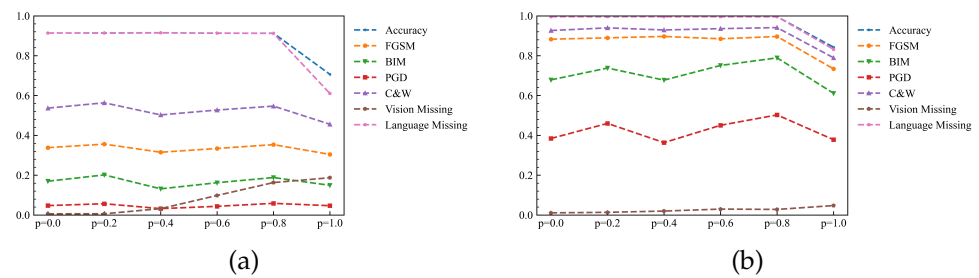


Figure 5. The accuracy trend of the probability of modality deactivation p . (a) Birds. (b) Flowers.

Table 6. Accuracy averages for different cases with different values of the modal deactivation probability p for the Birds data set.

	$p = 0.0$	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.8$	$p = 1.0$
Accuracy	0.9142	0.9137	0.9147	0.9129	0.9128	0.7065
FGSM	0.3418	0.3605	0.3152	0.3341	0.3534	0.3044
BIM	0.1716	0.2027	0.1315	0.1622	0.1882	0.1491
PGD	0.0456	0.0557	0.0321	0.0436	0.0587	0.0468
C&W	0.5366	0.5627	0.5027	0.5272	0.5467	0.4558
Vision Missing	0.0070	0.0059	0.0321	0.0982	0.1630	0.1872
Language Missing	0.9137	0.9143	0.9152	0.9136	0.9120	0.6103

Table 7. Accuracy averages for different cases with different values of the modal deactivation probability p for the Flowers dataset

	$p = 0.0$	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.8$	$p = 1.0$
Accuracy	0.9962	0.9958	0.9961	0.9959	0.9956	0.8426
FGSM	0.8828	0.8895	0.8966	0.8848	0.8962	0.7338
BIM	0.6786	0.7379	0.6773	0.7510	0.7891	0.6113
PGD	0.3843	0.4600	0.3636	0.4505	0.5026	0.3781
C&W	0.9263	0.9397	0.9291	0.9358	0.9410	0.7894
Vision Missing	0.0108	0.0136	0.0198	0.0297	0.0281	0.0474
Language Missing	0.9958	0.9964	0.9956	0.9956	0.9956	0.8317

Tables 6 and 7 show that the model’s accuracy mostly stays the same when p is less than 0.8. The accuracy of the model drops substantially when $p = 1$. The reason is that with $p = 1$, the model only uses the data of the visual modality or the data of the linguistic modality to update the weights each time, especially in $q = 0$. The model only uses the features of the linguistic modality for training, and the model degenerates into a recognition model of the linguistic modality, so the accuracy drops considerably.

In the face of the adversarial example attack, the model can achieve the maximum number of accuracy maxima in the face of the adversarial sample attack at $p = 0.8$. This phenomenon is because a higher p -value can better ensure the independence between two features, whose standard features can represent more information in an integrated way and, therefore, can achieve higher accuracy. In contrast, at $p = 0$, the model is not trained with random modal deactivation. Its average accuracy will be lower than the two maximum value points and higher than the minimum points. The results show that the parameter p can improve in robustness only by using the correct value. Otherwise, it may not only fail to improve the robustness but may also decrease it.

In the face of modality missing, it shows that both $p = 0$ and $p = 1$ correspond to poor average accuracy. These extreme cases represent training only the verbal modal feature extraction module and the visual modal feature extraction module, respectively, leading to model failure.

(2) Analysis of the effect of the visual modal integrity probability parameter q on the model

By averaging Figures 2–4 on the vertical axis, we can obtain the average accuracy of the visual modal integrity probability q in each case, as shown in Figure 6. There is a maximum point at $q = 0.4$ for both datasets, achieving the highest accuracy under most adversarial sample attacks. Equation (14) shows the reason for this point. The parameter p controls the probability of training the model with visual modal data in the case of missing modality. A large or small probability p will result in one of the modality feature extraction modules that cannot be trained effectively. Therefore, the value $q = 0.4$ is suitable to meet the theoretical expectation.

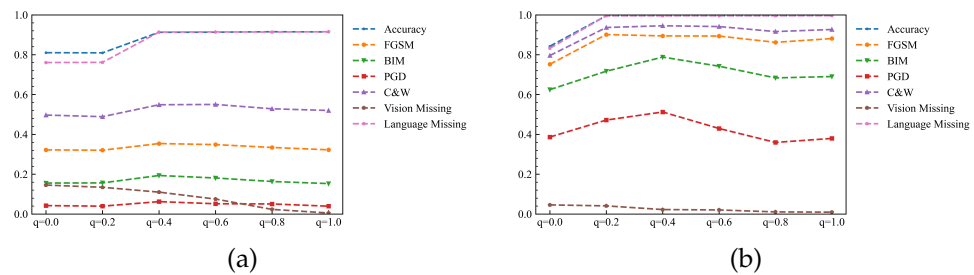


Figure 6. The accuracy trend of the visual modal integrity probability q . (a) Birds. (b) Flowers.

Table 8 and 9 shows the specific values of Figure 6. Tables show that the maximum number of maximum points for accuracy exists at $q = 0.4$ under the four adversarial examples attacks. The reason the accuracy maximum point appears at $q = 0.4$ when facing the adversarial example attack is that when the visual modality input data of the model is severely attacked, the unattacked linguistic modality features can describe the target better at this time. Lowering the p -value and making the model more biased to use the features of linguistic modality can improve some accuracy. Moreover, letting $q = 0$ and $q = 1$ both lead to a significant drop in the model’s accuracy in the face of modal deficits, which suggests that when training with the random modal deactivation method, it is best not to have the model extremely biased towards training a single modality. A proper bias towards training linguistic modalities enables the model to learn better standard features.

Table 8. Accuracy averages for different cases with different values visual modal integrity probability parameter q for the Birds dataset.

	$q = 0.0$	$q = 0.2$	$q = 0.4$	$q = 0.6$	$q = 0.8$	$q = 1.0$
Accuracy	0.8100	0.8092	0.9125	0.9130	0.9153	0.9142
FGSM	0.3221	0.3203	0.3538	0.3486	0.3342	0.3223
BIM	0.1554	0.1562	0.1934	0.1809	0.1633	0.1528
PGD	0.0423	0.0394	0.0622	0.0515	0.0505	0.0393
C&W	0.4968	0.4888	0.5487	0.5501	0.5284	0.5198
Vision Missing	0.1454	0.1342	0.1098	0.0751	0.0232	0.0051
Language Missing	0.7606	0.7615	0.9136	0.9145	0.9135	0.9154

Table 9. Accuracy averages for different cases with different values of visual modal integrity probability parameter q for the Flowers dataset.

	$q = 0.0$	$q = 0.2$	$q = 0.4$	$q = 0.6$	$q = 0.8$	$q = 1.0$
Accuracy	0.8426	0.9956	0.9959	0.9956	0.9958	0.9967
FGSM	0.7520	0.9011	0.8943	0.8935	0.8619	0.8810
BIM	0.6245	0.7170	0.7879	0.7420	0.6835	0.6902
PGD	0.3866	0.4717	0.5123	0.4294	0.3595	0.3796
C&W	0.7959	0.9364	0.9453	0.9412	0.9160	0.9265
Vision Missing	0.0458	0.0408	0.0224	0.0204	0.0106	0.0093
Language Missing	0.8322	0.9954	0.9954	0.9959	0.9956	0.9961

(3) Analysis of the impact of pre-training on the model

The above experiments reveal that the model is more challenging to update the weights of the linguistic feature extraction module without the modal random deactivation method. After training with random modal deactivation, the model's accuracy in the face of language modality still needs to meet the requirements. For this reason, the number of training iterations required for each feature extraction module during the training of the multimodal model is not consistent. The feature extraction module of the visual modality is often already pre-trained on a large dataset, so the feature extraction module of the linguistic modality requires more iterations for training. Experiments were done with $p = 0.8$ and $q = 0.4$ to verify the above idea. Table 10 shows the experimental results, where pre-trained indicates that the language modality of the model has been pre-trained.

Table 10. The effect of pre-training of the language feature extraction module on various aspects of the model.

Dataset	p	q	Pretrained	Accuracy	FGSM	BIM	PGD	C&W	Vision Missing	Language Missing
Birds	0.8	0.4	False	0.9105	0.3639	0.2152	0.0811	0.3639	0.1775	0.9138
	0.8	0.4	True	0.9114	0.3826	0.2231	0.0642	0.5730	0.5051	0.9099
Flowers	0.8	0.4	False	0.9951	0.8451	0.7118	0.4412	0.9098	0.0382	0.9961
	0.8	0.4	True	0.9931	0.8882	0.7656	0.3803	0.9509	0.6441	0.9941

Table 10 shows that when the language feature extraction module is trained, the model's accuracy is substantially improved due to visual modality deficiency. At the same time, it can achieve some improvement in the face of the adversarial sample attacks of FGSM, BIM, and C&W. This indicates that the standard features of the multimodal fine-grained recognition model based on feature fusion are more likely to be biased to represent modal features that have been pre-trained. Therefore, it is desirable to pre-train each feature extraction module of the multimodal model to ensure that the standard features do not tend to represent a particular modality more often.

3.3. Discussion

The above experiments show that both EFCMF and Bi-modal PMA are multimodal fine-grained recognition methods based on feature fusion. However, the robustness of the two differs significantly because of the structural complexity between them. Bi-modal PMA transforms the features of the visual module into attention to linguistic modalities through the QRM module, a process that makes the model structure more complex. EFCMF, on the other hand, performs feature fusion using connections, which is a better way to reduce the complexity of multimodal fine-grained recognition models.

Bi-modal PMA and EFCMF without pre-training of the linguistic modality feature extraction module showed lower accuracy when faced with the missing of linguistic modality. The reason is that both models use a pre-trained visual modal feature extraction module, which requires far fewer iterations to train than the linguistic modal feature extraction module because it is pre-trained. At the same time, EFCMF is trained using random modal deactivation, which simulates the modal deficit and allows the model to cope with counter-sample attacks. The random deactivation train method and pre-training of the linguistic modal feature extraction module make EFCMF more capable of coping with the missing modality.

Although EFCMF is able to achieve 91.80% and 99.80% on Birds and Flowers datasets, respectively, with specific parameters, the difference in accuracy of each parameter is not large, generally within 2%. Therefore, this study chooses to sacrifice some accuracy to improve the robustness of the model, which is also more beneficial to the application of the method in practical engineering.

In summary, it is because EFCMF employs various methods that facilitate the improvement of robustness that it has strong robustness in the face of modal deficiencies and against sample attacks.

4. Conclusions

In order to improve the ability of deep learning models to cope with modal missing and adversarial sample attacks, this study designs an enhanced framework for modal feature complementarity, EFCMF. The framework does not require additional expensive methods such as model distillation or adversarial training to train the models. The method effectively improves the models' robustness with appropriate parameter selection. Meanwhile, relevant experiments are conducted in the latest multimodal fine-grained classification methods using the training method of the framework, and the results show the validity of the findings. While ensuring the overall classification accuracy of the model and enhancing the ability of the model to extract features of each modality, the model gains the ability to cope with the lack of modality and some ability to cope with adversarial examples.

When facing the same level of adversarial sample attacks, EFCMF has a significant advantage over Bi-modal PMA in dealing with FGSM, PIM, and PGD adversarial sample attacks. It achieves a 15.56%, 8.81%, and 6.29% accuracy advantage on the Birds dataset and a 73.79%, 71.68%, and 37.84% accuracy advantage on the Flowers dataset, respectively. The average accuracy of EFCMF for both datasets is 52.85%, which is 27.13% higher than Bi-modal PMA when facing all four adversarial sample attacks. In the face of visual modal deficits, EFCMF achieves 45.65% and 63.43% higher accuracy on Birds and Flowers datasets, respectively. In the face of linguistic modal deficits, EFCMF achieved 12% and 9.42% higher accuracy on the Birds and Flowers datasets, respectively. The average precision of EFCMF for both datasets is 76.33%, which is 32.63% higher than that of Bi-modal PMA in the face of modal deficits. Regarding accuracy, EFCMF achieved 91.14% and 99.31% accuracy on the Birds and Flowers datasets. All these performances show that EFCMF has high accuracy and strong robustness.

Author Contributions: Conceptualization, R.Z., B.Z. and Y.C.; writing—original draft preparation, R.Z.; visualization, R.Z.; review and editing, Y.C., B.X. and B.S.; funding acquisition, B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Science Foundation of China (No. 61307025).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-Ucsd Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
2. Nilsback, M.E.; Zisserman, A. Automated flower classification over a large number of classes. In Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, Bhubaneswar, India, 16–19 December 2008; pp. 722–729. [[CrossRef](#)]
3. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.F. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*; Citeseer: Princeton, NJ, USA, 2011; Volume 2.
4. Krause, J.; Stark, M.; Deng, J.; Fei-Fei, L. 3D Object Representations for Fine-Grained Categorization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Sydney, Australia, 1–8 December 2013.
5. Hou, S.; Feng, Y.; Wang, Z. Vegfru: A domain-specific dataset for fine-grained visual categorization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 541–549.
6. Wei, X.S.; Cui, Q.; Yang, L.; Wang, P.; Liu, L. RPC: A large-scale retail product checkout dataset. *arXiv* **2019**, arXiv:1901.07249.
7. Peng, J.; Xiao, C.; Li, Y. RP2K: A large-scale retail product dataset for fine-grained image classification. *arXiv* **2020**, arXiv:2006.12634.
8. Wei, X.S.; Song, Y.Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; Belongie, S. Fine-Grained Image Analysis with Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8927–8948. [[CrossRef](#)]
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
10. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
11. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
12. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
13. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
14. Huang, S.; Xu, Z.; Tao, D.; Zhang, Y. Part-stacked cnn for fine-grained visual categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
15. Zhang, N.; Donahue, J.; Girshick, R.; Darrell, T. Part-based R-CNNs for fine-grained category detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014.
16. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
18. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear cnn models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1449–1457.
19. Min, S.; Yao, H.; Xie, H.; Zha, Z.J.; Zhang, Y. Multi-objective matrix normalization for fine-grained visual recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4996–5009. [[CrossRef](#)] [[PubMed](#)]
20. Niu, L.; Veeraraghavan, A.; Sabharwal, A. Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7171–7180.
21. He, X.; Peng, Y. Fine-grained image classification via combining vision and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5994–6002.
22. Li, J.; Zhu, L.; Huang, Z.; Lu, K.; Zhao, J. I Read, I Saw, I Tell: Texts Assisted Fine-Grained Visual Classification. In *Proceedings of the 26th ACM International Conference on Multimedia, MM '18*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 663–671. [[CrossRef](#)]
23. Zhang, H.; Cao, X.; Wang, R. Audio Visual Attribute Discovery for Fine-Grained Object Recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18*; AAAI Press: Washington, DC, USA, 2018.

24. Marino, K.; Salakhutdinov, R.; Gupta, A. The more you know: Using knowledge graphs for image classification. *arXiv* **2016**, arXiv:1612.04844.
25. Chen, T.; Lin, L.; Chen, R.; Wu, Y.; Luo, X. Knowledge-embedded representation learning for fine-grained image recognition. *arXiv* **2018**, arXiv:1807.00505.
26. He, Y.; Tian, L.; Zhang, L.; Zeng, X. Knowledge Graph Representation Fusion Framework for Fine-Grained Object Recognition in Smart Cities. *Complexity* **2021**, *2021*, 8041029. [[CrossRef](#)]
27. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
28. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
29. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 99–112.
30. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
31. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
32. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–24 May 2017; pp. 39–57.
33. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
34. Zhang, Y.; Wallace, B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv* **2015**, arXiv:1510.03820.
35. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. 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; pp. 248–255.
36. Reed, S.; Akata, Z.; Lee, H.; Schiele, B. Learning deep representations of fine-grained visual descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 49–58.
37. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
38. Kim, H. Torchattacks: A pytorch repository for adversarial attacks. *arXiv* **2020**, arXiv:2010.01950.
39. Cui, Y.; Song, Y.; Sun, C.; Howard, A.; Belongie, S. Large scale fine-grained categorization and domain-specific transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4109–4118.
40. Touvron, H.; Cord, M.; El-Nouby, A.; Verbeek, J.; Jégou, H. Three things everyone should know about Vision Transformers. *arXiv* **2022**, arXiv:2203.09795.
41. Zhao, Y.; Li, J.; Chen, X.; Tian, Y. Part-Guided Relational Transformers for Fine-Grained Visual Recognition. *IEEE Trans. Image Process.* **2021**, *30*, 9470–9481. [[CrossRef](#)] [[PubMed](#)]
42. Song, K.; Wei, X.S.; Shu, X.; Song, R.J.; Lu, J. Bi-modal progressive mask attention for fine-grained recognition. *IEEE Trans. Image Process.* **2020**, *29*, 7006–7018. [[CrossRef](#)]

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