electronics

MDPI

Article

A Study of Weather-Image Classification Combining VIT and a Dual Enhanced-Attention Module

Jing Li [1,*] **and Xueping Luo** [2]

[1] Department of Computer Science and Engineering, Southwest Minzu University, South Section 4, First Ring Road, Wuhou District, Chengdu 610041, China
[2] Department of Mathematics, Southwest Minzu University, South Section 4, First Ring Road, Wuhou District, Chengdu 610041, China
* Correspondence: li.wade707@outlook.com

Abstract: A weather-image-classification model combining a VIT (vision transformer) and dual augmented attention module is proposed to address the problems of the insufficient feature-extraction capability of traditional deep-learning methods with the recognition accuracy still to be improved and the limited types of weather phenomena existing in the dataset. A pre-trained model vision transformer is used to acquire the basic semantic feature representation of weather images. Dual augmented attention combined with convolutional self-attention and Atrous self-attention modules are used to acquire the low-level and high-level deep-image semantic representations, respectively, and the feature vectors are spliced and fed into the linear layer to obtain the weather types. Experimental validation is performed on the publicly available standard weather-image datasets MWD (Multi-class Weather Database) and WEAPD (Weather Phenomenon Database), and the two datasets are combined to enhance the comprehensiveness of the model for weather-phenomena recognition. The results show that the model achieves the highest F1 scores of 97.47%, 87.69% and 92.73% on the MWD, WEAPD and merged datasets, respectively. These scores are higher than the scores of recent deep-learning models with excellent performance in the experimental comparison, thereby, proving the effectiveness of the model.

Keywords: weather-image classification; vision transformer; convolutional self attention; atrous self attention; feature fusion

check for updates

Citation: Li, J.; Luo, X. A Study of Weather-Image Classification Combining VIT and a Dual Enhanced-Attention Module. *Electronics* **2023**, *12*, 1213. https://doi.org/10.3390/electronics12051213

Academic Editor: Giovanni Ramponi

Received: 2 February 2023
Revised: 28 February 2023
Accepted: 1 March 2023
Published: 3 March 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/).

1. Introduction

Weather-phenomenon detection plays a critical role in various applications, such as weather forecasting, road monitoring, transportation and agriculture [1–3]. Limited research has been performed on weather-image classification, which is different from ordinary image-classification tasks and more difficult. Since the same objects sometimes appear in different classes of weather images, it is difficult to extract effective features for weather-image classification when using traditional feature extraction for image feature extraction.

Unlike other scene-classification problems, weather-image classification is susceptible to factors that do not have structural information, such as illumination and reflection, and these factors usually have high correlation with each other while changing all the time, making it necessary for the classification model to obtain more abstract features of the image. The richness of the scene structure leads to weak information about the changes in weather conditions presented on the whole image, making the traditional scene classification [4] methods inappropriate for weather-image classification studies.

Weather-image-classification methods based on feature engineering [5] usually have a complex tuning process and require strong expertise of the researcher. Each method is also application-specific, thus, resulting in poor model generalization ability as well as poor robustness.

Aiming at the current problems and challenges in the field of weather-image classification, this paper mainly explores and analyses various weather-image-classification methods based on deep learning and proposes a weather-image-classification model combining VIT and a double enhanced-attention module with high classification accuracy and a good generalisation ability to result in good application value. The main contributions and innovations of this paper are as follows.

(1) To address the problems of low accuracy of existing weather-image-classification methods and difficulties in training network models, we fine-tune the weather-image-classification task using features learned by migration with the pre-trained model vision transformer (VIT) to obtain a global feature representation of the image.

(2) To fully capture the low and high level semantic features of images, a dual enhanced attention model consisting of convolutional self-attention and atrocious self-attention is constructed.

(3) To ensure the model can identify as many weather phenomena as possible, the multi-classification weather-image datasets MWD and WEAPD are used for separate experiments, and the two are combined to obtain a weather-image dataset with a classification number of 15.

2. Related Work

Traditionally, we mainly rely on human visual observation or physical sensors for weather identification [5]; however, as human visual observation is generally subjective and infrequent, and observation times and locations are easily restricted by objective conditions, the final conclusions are also inaccurate.

With the rapid development and deepening of smart-city construction, traditional weather-recognition methods cannot meet our needs for city management. In recent years, with the continuous progress of smart-city construction, a variety of intelligent monitoring devices have entered various areas of cities, and we can quickly and easily obtain outdoor weather images, which has made outdoor-weather-image-recognition methods based on image processing and machine learning increasingly popular.

Elhoseiny et al. [6] used a convolutional neural network (CNN) to study weather classification from images, first pre-training an AlexNet model using the Image Net large image dataset, then transferring the trained model parameters to the new AlexNet model and finally fine-tuning the AlexNet model to achieve the classification of two types of weather images, sunny and foggy. The experimental results showed that the model had excellent classification performance and achieved better weather-image classification results.

Zhu et al. [7] used convolutional neural networks to solve the problem of extreme-weather recognition by collecting and organising a large extreme-weather dataset, Weather Dataset, in which 16,635 extreme weather images were classified into four categories: sunny, heavy rain, snowstorm and fog, which covered most of the complex scenes. The classification and recognition of extreme weather consisted of two main steps: model pre-training and fine-tuning. In the model pre-training stage, the Goog Le Net model was first trained with the large-scale image dataset Image Net, and the Goog Le Net model was adjusted according to the specific task requirements. Then, the model was further trained with the extreme weather-image dataset, and finally a more accurate extreme-weather-recognition model was obtained. The experimental results showed that the authors' proposed method achieved a high accuracy rate of 94.5%, which can meet the needs of some practical applications.

Li et al. [8] proposed a data-enhancement method using a generative adversarial network (GAN) that complements and enriches the diversity of image data. Specifically, the authors designed a framework that uses deep convolutional generative adversarial networks (DCGAN) as a generator to generate images to balance an unbalanced dataset, and then they performed classification experiments on a dataset that was processed by the generative adversarial network to balance the number of categories. Classification experiments were then performed on a dataset with a balanced number of categories

after processing by generative adversarial networks. The experimental results showed that high-quality weather images could be generated on the weather-image dataset using deep convolutional generative adversarial networks. The classification accuracy of the model was effectively improved after using the data-enhancement technique based on deep convolutional generative adversarial networks.

Guo et al. [9] studied outdoor-weather-image classification and used AlexNet to learn deep-image features and fuse them with traditional features to verify the effectiveness of the model on a common five-class weather dataset, AMOS. Xiao et al. [10] proposed a weather-recognition model based on the MeteCNN deep convolutional network to build a database of weather phenomena with 11 classifications, and this had a significant improvement in classification performance and more comprehensive categories of weather phenomena compared to the traditional image pre-trained model. Zuo et al. [11] designed a weather-recognition model, VGG16-TL, based on image chunking and feature fusion to fuse shallow features of weather images with deep features to classify weather phenomena, including fog, rain, snow and sunshine.

Mesut et al. [12] combined image features extracted from the GoogLeNet and VGG-16 models and used the SNN (Spiking Neural Network) network for classification. They experimentally validated their method on cloudy, rainy, sunny and sunrise weather-image data; however, the model struggled to learn the shallow information of weather images. Chen et al. [13] introduced a lightweight convolutional neural network and multi-headed self-attentive mechanism to capture both local and global features of images and achieved good performance on remote-sensing image-classification tasks. Y et al. [14] constructed a dataset of nine weather phenomena images for classification and recognition by fusing deep residuals and dense convolutional networks.

The datasets used in the above deep-learning model algorithms contained few types of weather and could not accurately identify most possible weather phenomena. In addition, the training relied on advanced computing devices and took a long time to train. Therefore, to obtain a satisfactory deep-learning model, one must be able to afford certain financial and time costs. The pre-trained models used or the subsequent feature-learning modules all used a CNN as the core, which makes it difficult to learn both high- and low-level features and presents limited feature-extraction capabilities. Although these deep-learning-based weather-image-classification methods have shown good performance on certain datasets, there are still many challenges to be solved, which can be summarised as follows.

First, the current publicly available weather-image datasets are small in size and too homogeneous to meet the situation where the training of deep-learning models requires a large amount of labelled data for support. Second, the existing weather-image-classification methods have limited capabilities—in particular, the accuracy rate is not high when classifying multiple types of weather images, which cannot meet the actual demand. Third, as a deep-learning network model becomes more complex and the number of model parameters increases, the computational power and time required for model training increases, and one must rely on high-performance graphics cards to train the model.

3. Weather-Image-Classification Model Combining VIT and a Dual Enhanced-Attention Module

3.1. Overall Model Structure

The overall structure of the weather-image-classification model combining VIT and a dual enhanced-attention module is shown in Figure 1. It mainly consists of the combination of the pre-trained model VIT, convolutional self-attention module, Atrous self-attention module, feature fusion layer and classification output layer. The original weather image is pre-processed by image chunking and embedding and is then input to the VIT module for fine-tuning to obtain the global semantic feature representation of the current image. After the secondary high- and low-feature learning by the dual enhanced-attention module constructed by the convolutional self-attention module and the Atrous self-attention module,

the feature fusion operation is performed, and the weather type belonging to each image is obtained by the classification output layer.

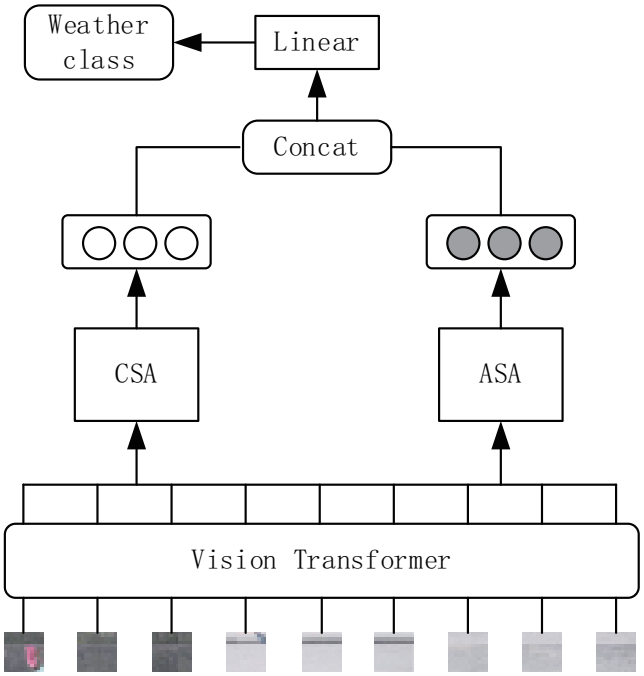

Figure 1. The overall model structure.

3.2. Transfer Learning

Transfer learning [15] is a research problem in machine learning that focuses on transferring the knowledge learned in the source domain to the target domain and improving the performance of the model in the target domain. A deep transfer-learning network model is designed and built by exploiting the properties of different layers of convolutional neural networks.

The same network is first trained on a very large dataset to achieve a better classification result. The convolutional layer structure of the network and the weight of the trained part of the network are then used to classify the data. The weights obtained from the training are extracted, and the fully connected layer structure for classification of the new task is added after the convolutional layer structure. In the training process of the new task, the network is trained with the fully connected layer structure. In the training process of the new task, the previously trained convolutional weights are loaded into the deep transfer-learning network and then retrained by weight fine-tuning or weight freezing.

Researchers in many fields use migration learning to solve special problems, such as for the MRI image abnormal-brain-detection problem. Lu et al. [16] modified the batch normalization layer of the pre-trained model AlexNet and introduced the chaotic bat algorithm for an automatic parameter search of the extreme-learning machine, which achieved the best performance. The authors in [17] proposed a novel interpretable new coronary pneumonia diagnostic system CGENet based on graph embedding. They introduced graph theory into the k-nearest-neighbour-based ResNet-18 model, used an extreme-learning machine for the classification operation and proved that CGENet achieved 97.78% accuracy using a five-fold cross-validation method. The graph-embedding approach effectively improved the classification performance.

3.3. Vision Transformer Pre-Trained Model

A VIT [18], as a visual classification network with a pure transformer-encoder structure, is able to achieve better performance than CNNs by performing pre-training on large-scale

general-purpose image datasets and migrating to in-domain fine-tuning for classification tasks on small- and medium-scale datasets [19].

The core process of a VIT includes the main components of image chunking (make patches), image block embedding (patch embedding) as well as position coding, transformer encoding [20], MLP head classification processing, etc. The structure of the VIT model is shown in Figure 2.

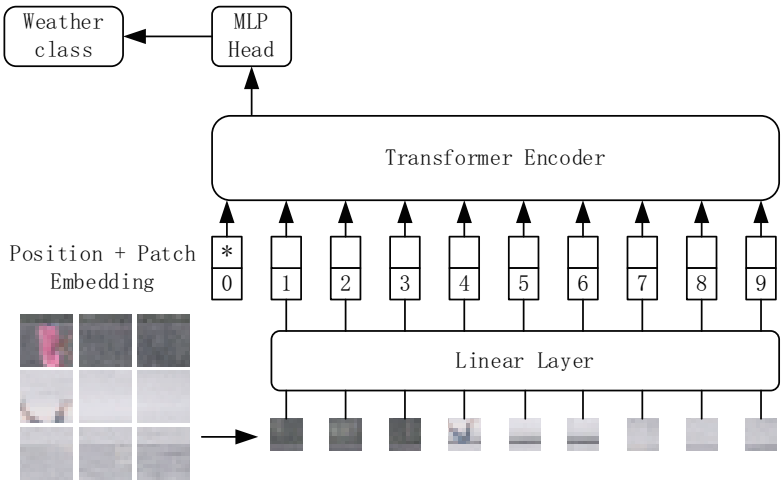

Figure 2. Model structure of a VIT.

While traditional convolutional neural networks can directly convolve images in two dimensions without special image chunking and block embedding processes [21], the standard transformer encoder accepts a one-dimensional vector sequence input. The current input image dimension size is $x \in R^{H \times W \times C}$. $H \times W$ denotes the image resolution size, and C is the number of channels. Assuming that the block image size is set to $P \times P$, then N blocks are obtained after dividing the image, which is the length of the sequence input to the transformer encoder. The calculation process is shown in Equation (1).

$$N = \frac{H \times W}{P \times P} \tag{1}$$

The dimension size of all block images is $N \times P \times C \times C$, and the spreading operation is performed for each image block with dimension size $Patch \in R^{P \times R \times C}$. The corresponding data dimensions can be written as $N \times (P^2 \times C)$. The embedding operation is performed on the flattened vector $Patch$ by the linear transformation layer, and the related process is shown in Equation (2).

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos} \tag{2}$$

where $E \in R^{P^2 \times C \times D}$ denotes the linear layer with input dimension size $P^2 \times C$, and the output dimension size is D. By referring to the BERT model, a trainable vector parameter x_{class} is introduced specifically for classification. $E_{pos} \in R^{(N+1) \times D}$ denotes the position encoding of the image block. In order to maintain the spatial position information between the input images $Patch$, one-dimensional learnable position-embedding variables are used [22].

The transformer-encoder module mainly consists of a multi-head self-attention mechanism (Multi-Head Self Attention), layer normalization (Layer Norm), residual connection and MLP layer combination, where the MLP module consists of a fully connected layer (Linear), GELU activation function and random deactivation layer (Dropout). The main transformer-encoder module structure is shown in Figure 3.

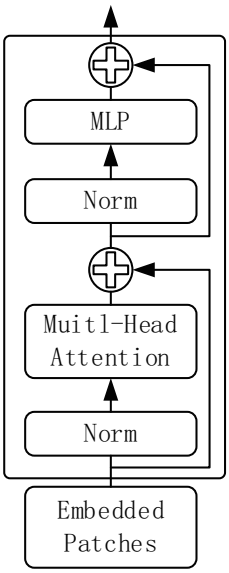

Figure 3. Structure of the transformer encoder.

In this paper, instead of using the MLP head module to output the weather-image classification probabilities, the encoding of each block is output by the transformer encoder $[O_p^1, O_p^2, \ldots, O_p^N]$. A block fusion operation is performed and then fed to the convolutional self-attention and Atrous self-attention modules to further capture low- and high-level image semantic features, respectively.

3.4. Convolutional Self-Attention Module

Convolutional layers are excellent layers for handling low-level features, and a number of studies have been proposed to combine self-attentiveness with convolutional global sensing fields [23–25]. The convolutional self-attention (CSA) module [26] integrates local self-attentiveness into the in-kernel convolution process of size 3×3, thereby, defining a window-based self-attentive layer to enhance the model's learning of the underlying location features with better generalization capability. The structure of the CSA model is shown in Figure 4.

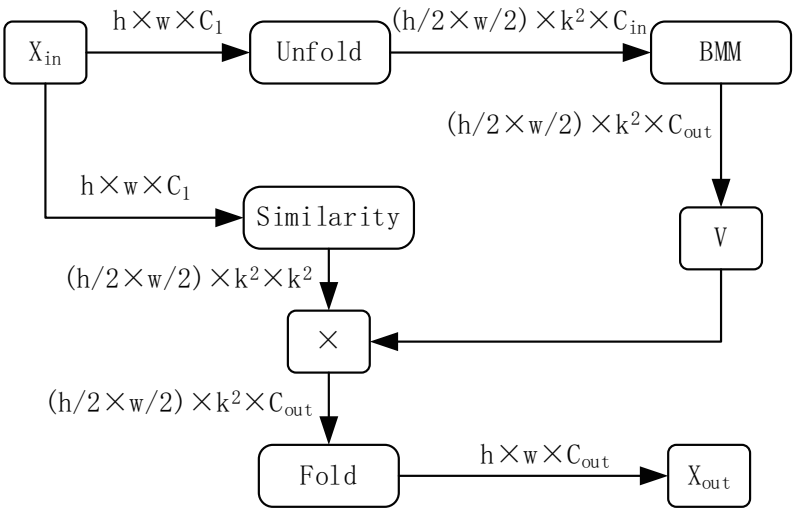

Figure 4. Convolutional self-attentive module structure.

Here, x_{in} denotes the feature vector obtained from the transformer-encoder output after a block fusion operation; $Unfold$ is a manually implemented sliding window operation

with a step size of 2; $fold$ denotes the inverse operation of $Unfold$; BMM denotes batch matrix multiplication; $Similarity$ denotes the similarity calculation process; $H \times W$ denotes the image resolution size; and C is the number of channels. x_{out} denotes the feature output after CSA module calculation.

The process of generalizing self-attention and convolution into unified convolutional self-attention is shown in Figure 5, and the related calculation process is shown in Equation (3).

$$y_i = \sum_{j \in N(i)} a_{i \to j} W_{i \to j} x_j \tag{3}$$

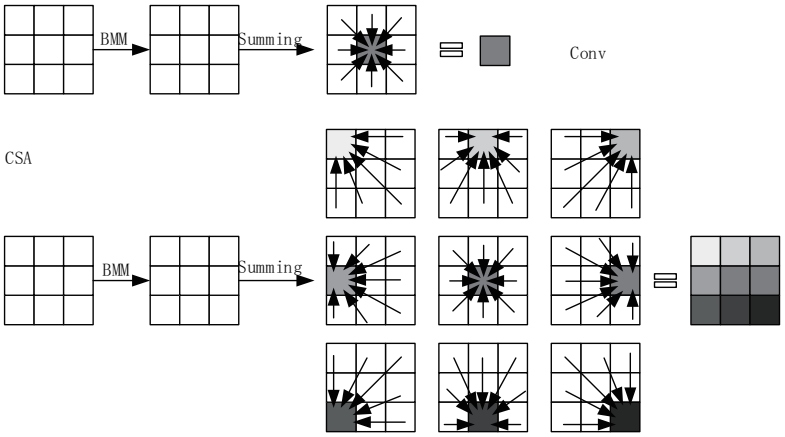

Figure 5. Convolutional self-attentive promotion process.

Among these, $x, y \in R^d$; the x and y are the feature inputs and outputs, respectively; and d denotes the number of channels. i, j is the index space position. $W_{i \to i} \in R^{d \times d}$ denotes the projection matrix, and $i \to j$ denotes the relative spatial relationship from position i to j. $a_{i \to j} \in (0, 1)$ is a is a scalar that controls the magnitude of the contribution of the value $W_{i \to i} x_j$ to the sum. $N(i)$ denotes the spatial position in the local neighbourhood defined by the kernel centred at position i, and when the size of the convolutional kernel is 3×3, $|N(i)|$'s size is 9, indicating a total of nine projection matrices.

The CSA module outputs vector x_{out} after layer normalization and the MLP layer to obtain the weather-image low-level feature representation CSA_{out}.

3.5. Atrous Self-Attention Module

Multi-scale features are beneficial for image-classification and target-recognition tasks, and the Atrous convolution module [27] can capture multi-scale image-context features with the same number of module parameters as the standard convolution kernel, while the weight-sharing operation can improve the model performance. In this paper, the Atrous self-attention module is used to capture the multi-scale semantics in the self-attentive similarity-mapping computation. The structure of the Atrous self-attention (ASA) module is shown in Figure 6, and the computation procedure is shown in Equations (4)–(6).

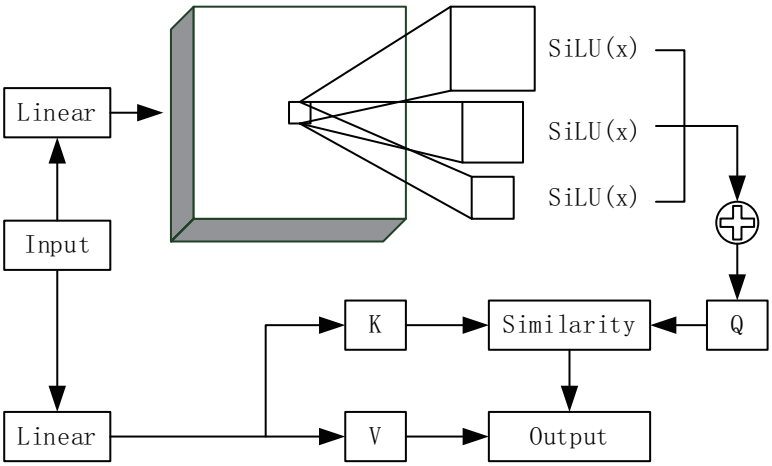

Figure 6. Structure of the Atrous self-attention module.

The Atrous self-attentive module uses a 1×1 convolution module to apply linear projections and captures multi-scale contexts using three convolutions with different expansion rates but shared weight kernels. The parameter overhead is further reduced by setting the number of groups to match the number of feature channels. Finally, the parallel-feature results for different scales $x * \mathrm{sigmoid}(\mathrm{x})$ are weighted and summed to obtain the ASA module vector output, and, after layer normalization and the MLP layer output, the high-level feature representation of the weather image ASA_{out} is output.

$$Q = \sum_{r \in \{1,3,5\}} \mathrm{SiLU}\left(\mathrm{Conv}\left(\widehat{Q}, W_q^{k=3}, r, g = d\right)\right) \tag{4}$$

$$\widehat{Q} = \mathrm{Conv}\left(\widehat{Q}, W_q^{k=1}, r = 1, g = 1\right) \tag{5}$$

$$\mathrm{SiLU}(m) = m * \mathrm{sigmoid}(m) \tag{6}$$

where $X, Q \in R^{d \times H \times W}$ is the image feature representation; $W_q^k \in R^{k^2 \times d \times d/g}$ is the learnable convolution kernel parameter matrix; d denotes the number of feature channels; and k, r and g denote the kernel size, expansion rate and number of groups of the convolution, respectively. *Conv* denotes the convolution operation, and *Sigmod* is the nonlinear activation function SiLU in [28]. To implement the self-calibration function, the activation intensity is used to determine the weight size of each scale.

3.6. Feature Fusion and the Classification Result Output Layer

The image feature representations $\mathrm{CSA_{out}}$ and $\mathrm{ASA_{out}}$ are stitched together and mapped to the weather-image instance classification space by a linear layer. The classification probability is calculated as P, and finally the maximum probability of each line corresponding to the label is taken by the function $MaxIndex$ as the weather-type recognition result $Result$. The calculation process is shown in Equations (7)–(9).

$$out = \mathrm{Concat}([CSA_{\mathrm{out}}, ASA_{\mathrm{out}}]) \tag{7}$$

$$\mathrm{Z} = Softmax(\tanh(Wout + \mathrm{b})) \tag{8}$$

$$Result = MaxIndex(\mathrm{Z}) \tag{9}$$

4. Experiments and Analysis of Results

4.1. Experimental Dataset and the Task-Evaluation Metrics

To verify the effectiveness of the VIT-DA model, the multi-class weather environment image dataset MWD [29] (Multi-class Weather Database) released by the team of the Visual Computing Research Center of Shenzhen University at the IEEE Image Processing Conference and the weather-phenomena dataset WEAPD (Weather Phenomenon Database)

used in the literature [9] were used for experimental validation. The dataset MWD contains 60,000 images of six common categories with weather types of sunny, cloudy, rain, snow, mist and thunderstorm, which are common weather types. WEAPD contains 6877 images of different types of weather, which are dew, fog, frost, glaze, hail, lightning, rain, rainbow, mist, dust storm and snow, totalling 11 types—mostly of severe weather. The training set, test set and validation set were randomly divided according to the ratio 8:1:1, and we ensured that there was no image overlap in each set. The detailed composition of the datasets MWD and WEAPD is shown in Table 1.

Table 1. Detailed composition of the datasets.

Dataset	Total/Sheet	Type	Quantity/Sheet
MWD	60,000	Sunny	10,000
		Cloudy	10,000
		Rain	10,000
		Snow	10,000
		Misty	10,000
		Thunderstorm	10,000
WEAPD	6877	Dew	700
		Fog	855
		Frost	475
		Glaze	639
		Hail	592
		Lightning	378
		Rain	527
		rainbow	238
		Mist	1160
		Sandstorm	692
		Snow	621

In this paper, we first conducted experimental validation on the two datasets separately to obtain the individual classification-performance results; subsequently, to further validate the performance of the model on each type of weather-image recognition, the same types of rain and snow data from the MWD and WEAPD datasets were combined for classification experiments.

To reflect the feasibility and effectiveness of the VIT-DA model proposed in this paper on weather-image-classification tasks, relevant evaluation metrics commonly used in image-classification tasks [30] were used: the accuracy(), precision rate(), recall rate() and F1 score. The calculation processes are shown in Equations (10)–(13).

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{10}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{13}$$

4.2. Parameter Setting

The module parameters and training process parameter settings are closely related to the performance results. After extensive experimental tuning, the specific parameter settings were as follows: this paper used the basic version of the VIT pre-trained model, the layer size was 12, the input dimension size was 224 × 224, the chunk size was 16 × 16, and the embedding dimensions were 768. The number of heads of the multi-headed attention

mechanism was eight. The input dimension size of the convolutional self-attention module was 128, and the convolutional core scale was 3×3.

The input dimension size of the Atrous self-attentive module was 256. The loss function was a multiclassification cross-entropy loss function, the batch size was 32, and the initial learning rate size was 1×10^{-5}. The number of training rounds was eight, and the optimizer RAdam [31] was used to adaptively adjust the learning rate to make the model jump out of local optima in the training process, so as to accelerate the model convergence and improve the training effect.

4.3. Analysis of the Experimental Results

To verify the effectiveness of the model in this paper on the weather-image-type-recognition task, the recent deep-learning algorithms AlexNet-feature fusion [9], VGG16-TL [11], MeteCNN [10], VGG16-GNet-SNN [12] and L-CT (lightweight convolutional transformer) were used as experimental comparisons, and relevant ablation experiments were set up to verify the magnitude of each module's contribution to the overall weather-image-classification performance of the model. To reduce the interference of random factors on the model performance and improve the reproducibility of the experimental results [32], all random number seeds were fixed in this paper to ensure consistent results of the parameter matrix initialization. An average of 10 experimental results was used for the final performance results.

4.3.1. Comparison with Other Deep-Learning Methods

Figures 7 and 8 show the trends of the F1 score and loss value, respectively, for each experimental model on the combined dataset.

From the results in Table 2, it can be seen that the model VIT-DA in this paper achieved the highest performance and improved the evaluation index F1 scores by 4.04%, 3.39%, 1.61%, 2.05%, 1.5% and 1.78%, respectively, over other the excellent deep-learning models AlexNet-feature fusion, VGG16-TL, MeteCNN, VGG16-GNet-SNN and L-CT on the merged dataset. This demonstrates the effectiveness of the pre-trained model VIT combined with dual enhanced attention to improve the weather-image-classification performance with the presence of diverse weather phenomena in the dataset, thus, covering the vast majority of weather conditions.

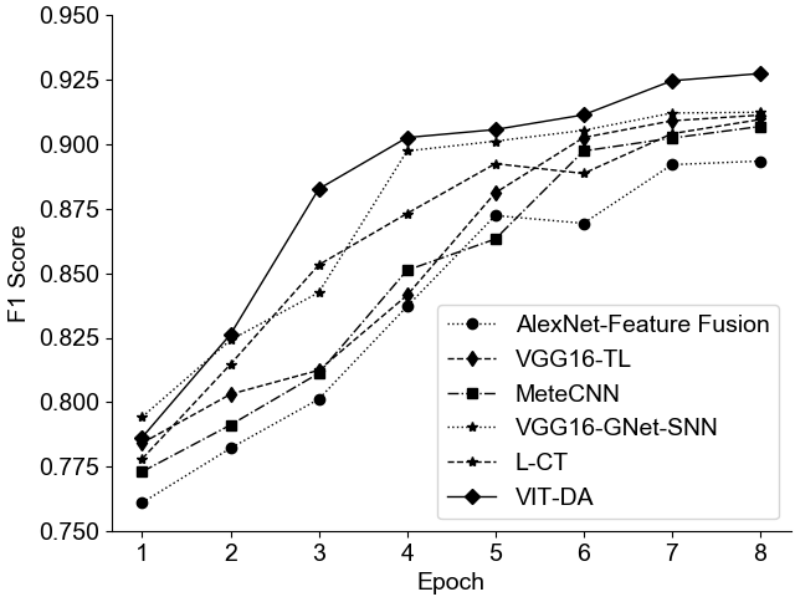

Figure 7. Trends of the F1 scores of different models.

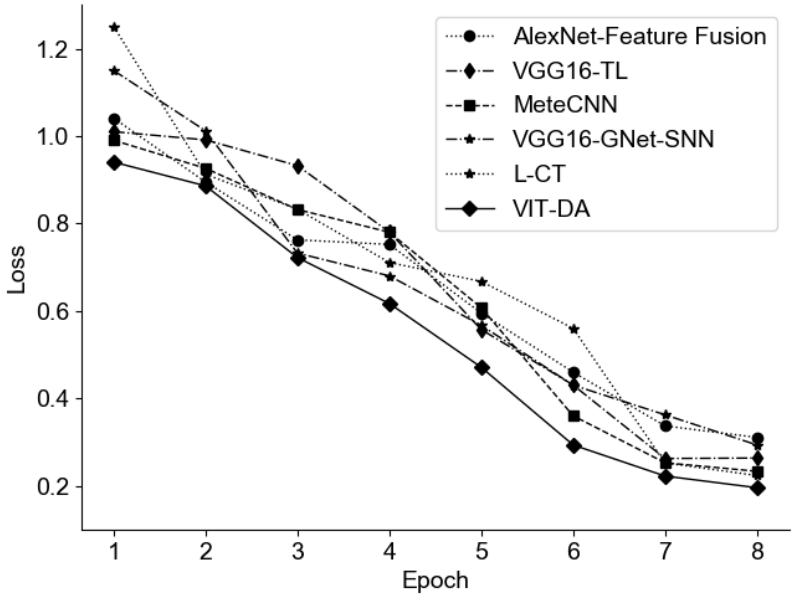

Figure 8. Trends of the loss values of different models.

Table 2. Experimental results of the weather-image-classification model.

Dataset	Model	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)
MWD	AlexNet-Feature Fusion	95.48	95.62	95.34	95.47
	VGG16-TL	95.62	95.77	95.43	95.59
	MeteCNN	96.33	96.50	96.14	96.32
	VGG16-GNet-SNN	96.48	96.74	96.23	96.47
	L-CT	95.94	96.06	95.82	95.93
	VIT-DA	97.48	97.52	97.45	97.47
WEAPD	AlexNet-Feature Fusion	85.44	85.61	85.27	85.43
	VGG16-TL	86.14	86.28	86.01	86.14
	MeteCNN	85.79	85.85	85.73	85.77
	VGG16-GNet-SNN	86.41	86.53	86.27	86.39
	L-CT	86.56	86.60	86.47	86.53
	VIT-DA	87.70	87.84	87.57	87.69
MWD & WEAPD	AlexNet-Feature Fusion	89.35	89.46	89.24	89.34
	VGG16-TL	91.13	91.20	91.07	91.12
	MeteCNN	90.68	90.75	90.62	90.68
	VGG16-GNet-SNN	91.24	91.27	91.20	91.23
	L-CT	90.95	91.04	90.87	90.95
	VIT-DA	92.74	92.87	92.61	92.73

The F1 values of the VIT-DA model of this paper in the MWD, WEAPD and combined datasets were 97.47%, 87.69% and 92.73% respectively, with the best model performance in the MWD dataset. This is due to the fact that there are only six weather phenomena in MWD, and the number of samples in each category is more balanced while there are clear differences between each weather category. The WEAPD dataset, on the other hand, has similar weather types, all of which are severe weather, resulting in a low recognition accuracy. The dataset after combining MWD and WEAPD had a larger number of categories, and some similar weather was difficult to distinguish, which made the recognition less effective than for MWD.

4.3.2. Module Ablation Experiment

To reflect the magnitude of the contribution of each module in the model to the overall classification performance, ablation experiments were set up. Model VIT indicates that the image global feature vector (CLS) output from the final layer was directly used and

input to the fully connected layer for classification; model VIT-CSA indicates that the image classes were output from the fully connected layer by using the CSA module alone for secondary feature learning; and model VIT-ASA indicates that only the ASA module was subsequently used for feature extraction.

From the results in Table 3, it can be seen that using only the pre-trained model VIT for weather-image classification achieved the lowest performance in all three datasets. In contrast, the models VIT-CSA and VIT-ASA (using only the CSA and ASA modules for secondary feature learning in the combined dataset) decreased by 0.69% and 1.28% compared to the F1 score of the model VIT-DA metric in this paper. This proves that both the CSA and ASA modules had a positive impact on the weather-image-classification task and also demonstrates the effectiveness of the dual enhanced-attention module in being able to simultaneously capture both low-level and high-level semantic feature representations of images.

Table 3. Module ablation experiment results.

Dataset	Model	F1 (%)
MWD	VIT	94.61
	VIT-CSA	96.65
	VIT-ASA	96.29
	VIT-DA	97.47
WEAPD	VIT	83.41
	VIT-CSA	85.79
	VIT-ASA	85.43
	VIT-DA	87.69
MWD & WEAPD	VIT	89.41
	VIT-CSA	92.04
	VIT-ASA	91.45
	VIT-DA	92.73

4.3.3. Performance Comparison of the Pre-Trained Models

To verify the good performance of the pre-trained model VIT on weather-image classification, the pre-trained models ResNet50, VGG16, AlexNet and GoogLeNet, which are excellent performers in image classification with a convolutional neural network as the main component module, were used for experimental comparison and tested on the merged weather-image dataset. The experimental models were loaded with pre-trained weight parameters, and fine-tuning operations were performed on the experimental dataset. The related results are shown in Table 4.

Table 4. Results of the pre-trained model performance.

Dataset	Model	F1 (%)
MWD & WEAPD	VGG-16	87.24
	AlexNet	86.78
	GoogLeNet	88.26
	ResNet50	88.53
	VIT	89.41

From the experimental results in Table 4, it can be seen that the image classification pre-trained model VIT used in this paper achieved the highest F1 score, thereby, proving the superior performance of deep-image global feature extraction using the image-chunking technique and based on a transformer encoder, which could more accurately identify different types of weather phenomena and outperformed other pre-trained models with a convolutional neural network as the main backbone module. This also shows that the transformer encoder with the multi-headed self-attentive mechanism module as the

core had better feature-extraction capability compared with the traditional convolutional operation, and the model VIT application was more effective.

4.3.4. Optimizer Performance Comparison

To verify the effectiveness of the model in this paper using the optimizer RAdam on the weather-image-classification task, the classical optimizers with excellent performance, Adam, AdamW, AMSGrad, Lookahead [33] and SGD, were used as comparisons for experiments on the merged weather-image-classification dataset and to ensure consistent parameter settings except for the optimizer. The trends of the model F1 scores with training rounds with different optimizers are shown in Figure 9.

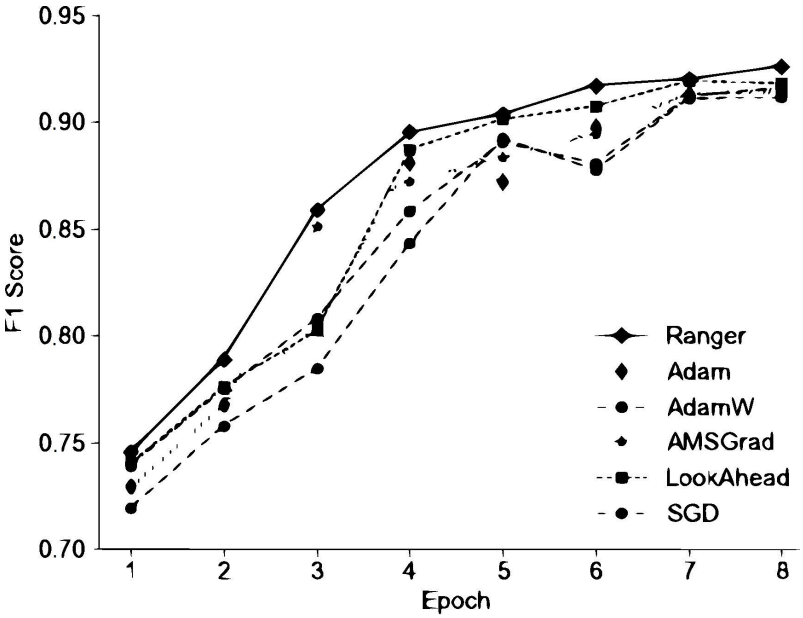

Figure 9. Performance comparison of different optimisers.

From the model performance results in Figure 9, it can be seen that, when the optimizer was RAdam, the F1 score of the model in this paper achieved 92.73%, and the model weather-image-type-recognition effect was better than with the other optimizers. The fluctuations in the process of the value rise were smaller and had better stability. This is because RAdam uses the dynamic rectifier strategy to adjust the adaptive momentum of Adam and, at the same time, adopts the warm-up strategy at the initial start-up stage to ensure early stability and an effective breakthrough during the training process, thus, improving the overall training effect of the model.

5. Conclusions

For the weather-image-classification task, in this paper, we proposed a weather-image-classification model combining a VIT and a dual enhanced-attention module, which extracts the basic semantic feature representation of weather images using the pre-trained model VIT and captures the low-level and high-level deep-level features through the dual enhanced-attention module.

Experimental validation was performed on the multi-classification weather-phenomena datasets MWD and WEAPD, and the two were combined to obtain a more comprehensive weather-phenomena dataset. The experimental results show that the VIT-DA model in this paper achieved the best performance on all three datasets, thereby, proving that it can effectively improve the classification accuracy of weather images and recognise most weather types. The pre-trained model VIT outperformed other models with conventional CNN-based backbone networks in weather-image-classification performance, and the de-

gree of contribution of each module to the overall performance was also verified through ablation experiments.

However, the model in this paper required high computational resources and performance, the total number of parameters was large, and recognition errors occurred for some very similar weather phenomena. Next, we will consider further compressing the model parameter size to reduce the time cost of model training and improve the inference speed, and we will introduce a block-scoring strategy for feature selection for the misclassification problem of similar weather phenomena to distinguish approximate weather types.

Author Contributions: Methodology, J.L.; Formal analysis, J.L.; Resources, X.L.; Data curation, J.L.; Writing—original draft, J.L.; Supervision, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

References

1. Zhang, Y.; Sun, J.; Chen, M.; Wang, Q.; Ma, R. Multi-Weather Classification Using Evolutionary Algorithm on EfficientNet. In Proceedings of the 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Kassel, Germany, 22–26 March 2021.
2. Zhao, B.; Hua, L.; Li, X.; Lu, X.; Wang, Z. Weather Recognition via Classification Labels and Weather-Cue Maps. *Pattern Recognit.* **2019**, *95*, 272–284. [CrossRef]
3. Wang, F.; Zhang, Z.; Liu, C.; Yu, Y.; Pang, S.; Duic, N.; Shafie-Khah, M.; Catalao, J.P.S. Generative Adversarial Networks and Convolutional Neural Networks Based Weather Classification Model for Day Ahead Short-Term Photovoltaic Power Forecasting. *Energy Convers. Manag.* **2019**, *181*, 443–462. [CrossRef]
4. Zou, J.; Li, W.; Chen, C.; Du, Q. Scene Classification Using Local and Global Features with Collaborative Representation Fusion. *Inf. Sci.* **2016**, *348*, 209–226. [CrossRef]
5. Adhikari, A.; Choudhuri, A.R.; Ghosh, D.; Chattopadhyay, N.; Chakraborty, R. Breast Cancer Histopathological Image Classification Using Convolutional Neural Networks. In Proceedings of the International Conference on Innovations in Software Architecture and Computational Systems, Kolkata, India, 2–3 April 2021; pp. 183–195.
6. Zhang, Z.; Ma, H. Multi-Class Weather Classification on Single Images. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 4396–4400.
7. Zhu, Z.; Zhuo, L.; Qu, P.; Zhou, K.; Zhang, J. Extreme Weather Recognition Using Convolutional Neural Networks. In Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM), San Jose, CA, USA, 11–13 December 2016; pp. 621–625.
8. Li, Z.; Jin, Y.; Li, Y.; Lin, Z.; Wang, S. Imbalanced Adversarial Learning for Weather Image Generation and Classification. In Proceedings of the 2018 14th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 12–16 August 2018; pp. 1093–1097.
9. Guo, C.; Hu, Y.W.; Liu, P.; Yang, J. Feature fusion-based outdoor weather image classification. *J. Comput. Appl.* **2020**, *40*, 1023–1029.
10. Xiao, H.; Zhang, F.; Shen, Z.; Wu, K.; Zhang, J. Classification of Weather Phenomenon From Images by Using Deep Convolutional Neural Network. *Earth Space Sci.* **2021**, *8*, e2020EA001604. [CrossRef]
11. Zuo, J.G.; Liu X.M; Cai, B. Outdoor image weather recognition based on image chunking and feature fusion. *Comput. Sci.* **2022**, *49*, 197–203.
12. Toğaçar, M.; Ergen, B.; Cömert, Z. Detection of Weather Images by Using Spiking Neural Networks of Deep Learning Models. *Neural Comput. Appl.* **2021**, *33*, 6147–6159. [CrossRef]
13. Chen, F.; Zhang, T.; Chen, R.B. Lightweight convolutional Transformer-based image classification method and its application to remote sensing image classification. *J. Electron. Inf.* **2022**, *44*, 1–9.
14. Wang, Y.; Li, Y.X. Research on Multi-Class Weather Classification Algorithm Based on Multi-Model Fusion. In Proceedings of the 2020 IEEE fourth Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020.
15. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
16. Lu, S.; Wang, S.-H.; Zhang, Y.-D. Detection of Abnormal Brain in MRI via Improved AlexNet and ELM Optimized by Chaotic Bat Algorithm. *Neural Comput. Appl.* **2021**, *33*, 10799–10811. [CrossRef]
17. Lu, S.-Y.; Zhang, Z.; Zhang, Y.-D.; Wang, S.-H. CGENet: A Deep Graph Model for COVID-19 Detection Based on Chest CT. *Biology* **2022**, *11*, 33. [CrossRef]

18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Houlsby, N. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. *arXiv* **2021**, arXiv:2010.11929.
19. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Gao, W. Pre-Trained Image Processing Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20 June–25 June 2021; pp. 12299–12310.
20. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11 October 2021; pp. 9992–10002.
21. Jose Valanarasu, J.M.; Yasarla, R.; Patel, V.M. TransWeather: Transformer-Based Restoration of Images Degraded by Adverse Weather Conditions. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 2343–2353.
22. Wu, K.; Peng, H.; Chen, M.; Fu, J.; Chao, H. Rethinking and Improving Relative Position Encoding for Vision Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10033–10041.
23. Lu, Z.; Whalen, I.; Dhebar, Y.; Deb, K.; Goodman, E.D.; Banzhaf, W.; Boddeti, V.N. Multiobjective Evolutionary Design of Deep Convolutional Neural Networks for Image Classification. *IEEE Trans. Evol. Computat*. **2021**, *25*, 277–291. [CrossRef]
24. Pei, Y.; Huang, Y.; Zou, Q.; Zhang, X.; Wang, S. Effects of Image Degradation and Degradation Removal to CNN-Based Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1239–1253. [CrossRef]
25. Chen, X.; Xie, L.; Wu, J.; Tian, Q. Cyclic CNN: Image Classification With Multiscale and Multilocation Contexts. *IEEE Internet Things J.* **2021**, *8*, 7466–7475. [CrossRef]
26. Yang, C.; Wang, Y.; Zhang, J.; Zhang, H.; Wei, Z.; Lin, Z.; Yuille, A. Lite Vision Transformer with Enhanced Self-Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022; pp. 11998–12008
27. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
28. 1. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *Neural Netw.* **2018**, *107*, 3–11. [CrossRef]
29. Lin, D.; Lu, C.; Huang, H.; Jia, J. RSCM: Region Selection and Concurrency Model for Multi-Class Weather Recognition. *IEEE Trans. Image Process.* **2017**, *26*, 4154–4167. [CrossRef]
30. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5966–5978. [CrossRef]
31. Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. On the Variance of the Adaptive Learning Rate and Beyond. In Proceedings of the International Conference on Learning Representations, Rhodes, Greece, 12–18 September 2020.
32. Li, Y.; Wu, X.; Li, C.; Li, X.; Chen, H.; Sun, C.; Rahaman, M.M.; Yao, Y.; Zhang, Y.; Jiang, T. A Hierarchical Conditional Random Field-Based Attention Mechanism Approach for Gastric Histopathology Image Classification. *Appl. Intell.* **2022**, *52*, 9717–9738. [CrossRef]
33. Zhang, M.R.; Lucas, J.; Hinton, G.; Ba, J. Lookahead Optimizer: K Steps Forward, 1 Step Back. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9597–9608.

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.