IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 April 2023, accepted 9 May 2023, date of publication 17 May 2023, date of current version 25 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3277204

RESEARCH ARTICLE

Long-Tailed Visual Recognition via Improved Cross-Window Self-Attention and TrivialAugment

YING SONG[1,2,3], MENGXING LI[1,2], AND BO WANG[4]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination, Beijing Information Science and Technology University, Beijing 100101, China
[2]Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, China
[3]State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100086, China
[4]Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450002, China

Corresponding author: Ying Song (songying@bistu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872043; and in part by the State Key Laboratory of Computer Architecture, Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), under Grant CARCHA202103.

ABSTRACT In the real world, large-scale image data sets usually present long-tailed distribution. When traditional visual recognition methods are applied to long-tail image data sets, problems such as model failure and sudden decline in recognition accuracy occur. While, when deep learning models encounter long-tailed datasets, they tend to perform poorly. In order to mitigate the impact of these problems, we propose CWTA (Long-tailed Visual Recognition via improved Cross-Window Self-Attention and TrivialAugment). CWTA uses CNN to better capture the local features of the image, uses the Cross-Window Self-Attention mechanism to dynamically adjust the perception domain to better deal with image noise, and uses TrivialAugment to enhance the diversity of a few types of data samples, thus improving the recognition accuracy of long-tailed distributed images. The experimental results show that the proposed CWTA performs best in the classification accuracy of different categories on different long-tailed datasets. We also compared CWTA with other long-tailed recognition algorithms (such as OLTR, LWS, ResLT, PaCo, and BALLAD), and the CWTA is the best when ResNet-50 as the Backbone. On the CIFAR100-LT, ImageNet-LT, and Places-LT datasets, the acc of all categories of CWTA is 12.9%, 0.4%, and 1.3% higher than that of BALLAD, respectively. For F_1-Score on CIFAR100-LT, ImageNet-LT, and Places-LT datasets, CWTA is 6.6%, 2.2%, and 1.5% higher than BALLAD, respectively.

INDEX TERMS Long-tailed recognition, self-attention, vision transformer, CNN, TrivialAugment.

I. INTRODUCTION

In various fields and problems, from natural science to social science, data usually present long-tailed distribution, that is a small number of categories occupy the majority of samples, while most categories only occupy a small number of samples, as shown in Figure 1. Classification and recognition systems that directly use long-tailed data for training tend to over-fit the head data, thus ignoring the tail category in the prediction. How to effectively use the unbalanced long-tailed data to train a balanced classifier is our concern. References [1], [2] From the perspective of industry demand, this research will also greatly improve the speed of data collection and significantly reduce the collection cost. The existing deep learning model [1], [2], [3], [4], [5] tends to support the head classes when processing long-tailed distribution datasets, and the accuracy of tail class classification is low. The essence of difficulty in learning the long-tailed recognition method is mainly divided into the following three aspects: First, there are too few tail samples in the long-tailed distribution data set, and the imbalance ratio of the whole data set is too high. Second, the loss of depth model is dominated by the head class, which makes the separated hyperplane seriously deviate from the tail class. Third, the tail class data is too small, resulting in too low intra-class diversity.

The basic methods of long-tailed recognition include re-sampling, re-weighting, and transfer learning · Re-sampling [6] reduces the over-fitting of head data to a certain extent.

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

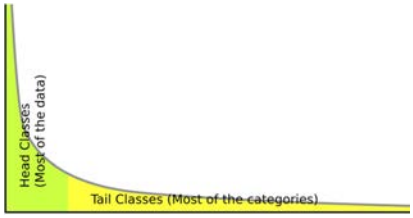

FIGURE 1. Example of long-tailed distribution dataset.

However, because a small amount of data onto the tail class is often repeatedly learned, lacks sufficient sample differences, and is not robust enough, and a large amount of data onto sufficient differences in the head is often not fully learned, resampling is not a truly perfect solution. The re-weighting [6] is mainly reflected off the classification loss, but it does not take account the fact that a large number of tailed categories has only a few samples, resulting in relatively low recognition accuracy of tail classes. The transfer learning model [6] is to learn general knowledge from the public classes of the head, and then migrate to several sample classes of the tail. However, this method usually requires the design of additional complex modules.

CNN (Convolutional Neural Network) [7], [8], [9], [10], [11] have made extraordinary contributions in the field of computer vision and have achieved good results in almost all CV (Computer Vision) tasks. Vision Transformer related models [12], [13], [14], [15] has made great achievements in computer vision tasks such as image classification, object detection, and semantic segmentation tasks.

Aiming at the low accuracy of tail class recognition of long-tailed distribution image classification, inspired by CNN and Vision Transformer, we propose an innovative method CWTA (Long-Tailed Visual Recognition via improved Cross-Window Self-Attention and TrivialAugment) to introduce Vision Transformer to long-tailed distribution datasets. CWTA is a new hybrid network structure model based on a Cross-Window Self-Attention mechanism, Convolutional Neural Network (CNN), and TrivialAugment (TA), which is proposed to the basis of CWCT [16]. CWTA mainly uses the advantages of Transformer to capture the global features of images, CNN to model local features, and TrivialAugment to enhance the features of tail classes. Compared with previous data enhancement strategies, TA is parameterless, and its search efficiency and data enhancement are superior to previous data enhancement strategies. Subsequent comparative experiments show that this method is effective when applied to long-tailed image classification. Based on CWCT, CWTA improves the characteristics of the long-tailed distribution dataset and introduces TrivialAugment to improve the classification accuracy of the minor classes.

Cross-Window

CNN

TrivialAugment

The motivation of this research is to conduct extensive experimentation using convolutional neural networks, Cross-Window Self-Attention mechanism, and TrivialAugment to Improve the overall accuracy of long-tail recognition and the accuracy few categories.

The primary contributions of our proposed study are:

- Based on CWCT, we propose a method CWTA. CWTA introduces a vision transformer from image classification to long-tailed image classification to improve the overall classification accuracy and the classification accuracy of minority classes.
- CWTA uses the advantages of Transformer to capture the global features of images, CNN to model local features, and TrivialAugment to enhance the features of tail classes.
- The TrivialAugment strategy is utilized to generate better outcomes of minority classes, and effectively avoid the influence of over-fitting on classification.

The remainder of the paper is divided and organized as follows: The second part of this paper mainly introduces the related work in four aspects: data resampling, data reweighting, migration learning, and data enhancement. The third part describes the design and implementation of the CWTA in detail. The fourth part analyzes the effect of CWTA through comparison. The fifth part summarizes the work of this paper.

II. RELATED WORK

The problem of long-tailed recognition mainly involves the problem of data imbalance and the problem of few samples because the head category is easy to lead the loss of model training and the tail category sample diversity is insufficient. The relevant research work is described in the following four aspects, namely re-sampling, reweighting, transfer learning and data augment.

A. RE-SAMPLING

Data re-sampling is a preprocessing technology to solve the problem of unbalanced classification of data onto the perspective of the data itself. In the past, a large number of sampling technologies have been proposed to different perspectives, mainly including oversampling for a few classes and under-sampling for most classes. In 2019, Zhou et al. [1] proposed a bilateral branch network model based on curriculum learning, which uses long-tailed datasets for training while re-sampling balanced datasets for training. It is pointed out that the learning process of decoupled feature networks and classifiers is conducive to the learning of long-tailed datasets. Kang et al. [17].deeply explored the decoupling problem between feature networks and classifiers in the long-tailed recognition process, and discussed a variety of sampling methods, including class balanced sampling method [18] sampling based on the inverse frequency distribution of class samples, and step-by-step balanced sampling method. In 2020, Ren et al. [19] proposed a balanced meta-softmax sampling method based on meta-learning to estimate the optimal sampling rate of different categories in long-tailed learning. Specifically, the proposed meta-learning method is a two-level optimization strategy, which learns the best sample distribution parameters by optimizing the model classification performance on the balanced meta-verification

set. In the Feature augmentation and sampling adaption (FASA) by Zang et al. [20] in 2021, it was pointed out that the model classification loss metric was used to adjust the feature sampling rate of different classes on the verification set of equilibrium elements. This allows for more sampling of underrepresented tail classes.

However, in the re-sampling method, the head class is under-sampled and some samples are discarded, which will reduce the effect of the model in the head class. Oversampling tail classes and repeatedly sampling some samples will lead to overfitting on tail classes. It may lead to overfitting of a few categories.

B. RE-WEIGHTING

The re-weighting method is mainly used to re-balance various types [21], [22] by adjusting the loss values of different types in the training process. In 2020, Tan et al. [3] proposed a new loss function, Equalization Loss. In the paper, they believed that positive samples of each category could be regarded as negative samples of other categories. Too large a negative sample gradient of the tail categories led to their poor learning effect. An ignorance strategy can be introduced to alleviate the problem of too large a negative sample gradient in the learning process. All kinds of loss function can mitigate the impact of unbalanced data distribution on model learning to a certain extent, and improve the recognition rate of minority samples, including the recognition rate of tail category samples in long-tailed distribution data. In 2021, the LADE [23] proposed by Hong et al. introduced label distribution disentangling loss, disentangled the learned model from the long-tailed training distribution, and then adapted the model to any test class distribution when the test tag frequency is available.

The disadvantage of such methods as reweighting is that they sacrifice the performance of the head class to improve the performance of the tail class. Although the model performance on all classes is improved, this cannot change the problem of data scarcity, especially the extreme scarcity of tail class data.

C. TRANSFER LEARNING

Transfer learning [24], [25] usually enhances target domain model training by seeking source domains (e.g., datasets, tasks, or classes) to transfer information, representation, and knowledge. In 2020, online feature authentication (OFA) [24] proposed by Chu et al. used class activation maps [26] to decouple sample features into class-specific and class-diagnostic features. On this basis, OFA expanded the tail class by combining the class-specific feature of the tail classes sample with the class-diagnostic feature of the head classes sample. After that, more enhanced and original features were used to fine-tune the model classifier and re-balanced the sampler to promote better long-tailed learning performance. In 2021, the Rare class sample generator (RSG) proposed by Wang et al. [25] mentioned that in the long-tailed problem, the feature space of the tail class is much smaller than that of the head class. To solve this problem, RSG proposes to generate new tail class samples to expand the feature space of the tail classes and ''push'' the decision boundary. For this reason, RSG dynamically estimates a set of feature centers of each class and uses the feature displacement between the head class sample features and the nearest intra-class feature center to enhance the features of each tail sample. In order to further maximize the feature distance and increase the diversity of tail sample features generated, RSG introduces a maximized vector loss to force the direction of feature displacement and the direction of sample features to become ''co liner.''

Due to the introduction of additional knowledge, the method based on migration learning can improve the performance of tail classes without sacrificing the performance of the head. Considering that the lack of sufficient tail samples is one of the key problems of long-tailed learning, this kind of method is worth further exploring.

D. DATA AUGMENTATION

Data Augmentation aims to use a series of data enhancement technologies to enhance the size and quality of data sets for model training [27]. Non-transfer authentication seeks to improve or design traditional data augmentation methods to solve the long-tailed problem. In 2021, Zhong et al. proposed MiSLAS [28] to study data mixup in long-tailed learning and found that data mixup is helpful to correct the overconfidence model. In the uncoupled training scheme, mixup has a positive impact on representation learning, while it has a negative or negligible impact on classification learning. Based on these observations, MiSLAS proposes to use data mixup in the befouled scheme to enhance presentation learning. In addition, Remix [29] also used the data mixup method to learn from the Standing Committee and introduced a re-balanced mixup method to enhance the tail class. In 2021, Li et al. [30] proposed that Meta semantic annotation (MetaSAug) use a variant of implicit semantic data annotation (ISDA) [31] to enhance the tail classes. Specifically, ISDA obtains the semantic direction by estimating the class conditional statistics (i.e. the covariance matrix of the sample features) of the sample features and generates a diverse enhancement samples by translating the sample features with different semantic meanings directions. However, due to insufficient tail samples, the covariance matrix of the tail cannot be estimated. To solve this problem, MetaSAug explores meta-learning to guide the learning of covariance matrices for each class. In this way, the covariance matrix of the tail class can be estimated more accurately, thus generating more tail class feature information.

Data enhancement is a relatively basic technology that can be used for various long-tailed problems, which makes this method more practical than other methods in practical applications. However, simply using the existing class-agnostic enhancement technology to improve long-tailed learning is unfavorable, because there are more samples in the head

classes, and the enhancement degree will become greater, which may further aggravate the imbalance.

Overall, both re-sampling and re-weighting methods essentially utilize known dataset distributions to brute force the data distribution during the learning process, i.e., reverse weighting, which strengthens the learning of tail categories and counteracts the long tail effect. Transfer learning is the process of learning general knowledge from common classes at the top, and then transferring it to classes with fewer samples at the end. Data augmentation refers to the method of adding small changes to existing data or creating composite data from existing data to increase the amount of minority class data.

The CWTA differs from traditional long-tailed image classification methods by introducing the self-attention mechanism from Vision Transformer in the field of image classification. It utilizes the advantages of CNN and Cross-Window Self-Attention to efficiently model the global and local features of the image, and TrivialAugment to enhance the diversity of minority class samples.

III. THE DESIGN OF CWTA

This section mainly introduces the design idea of the CWTA method. First, the CWTA method is introduced, and then the CWTA Block and TrivialAugment in the CWTA method are introduced in detail.

A. CWTA

CWTA method firstly preprocesses the data, and then inputs the data into the training model, which can facilitate the data enhancement of training data, thus solving the problem of over-fitting in the training process. The input picture can also meet the data size requirements of the input layer of the network. The Figure 2 shows the overall structure of CWTA training long-tailed datasets. For most classes, data preprocessing mainly involves random position clipping, random horizontal flipping, random up and down inversion, etc. For a few categories, it mainly involves random position clipping, random horizontal flipping, random up and down inversion, and TA data enhancement, as shown in Figure 3. After pretreatment, a stem structure similar to CWCT [16] is adopted for most categories and a few categories, which adopts 3 × 3 Conv, step size 2, output channel 32. Then do it Twice 3 × 3 Conv, the step size is 1, so that the model can get better convergence. Following the design of CNN's classical network structure ResNet, the model is divided into four stages to generate feature maps of different scales.

Each stage is a block aggregation layer composed of convolution and layer normalization, which can reduce the resolution of the feature map and double the projection dimension of the feature map. In each stage, several CWTA Blocks are stacked in turn for feature transformation, while maintaining the same resolution of the input. For example, the Figure 2 shows that "Stage 2" of the CWTA model contains three CWTA Blocks. Each CWTA Block contains LPU, Cross-Window Self-Attention, and IRFFN respectively. The pseudocode of the CWTA algorithm is shown in Table 1.

TABLE 1. CWTA algorithm procedure.

Algorithm 1 CWTA Pseudocode

Input: $I = \{images\}$, $I = \{labels\}$, $A = \{a_1, a_2, \ldots, a_n\}$
Output: $model_{weight}$

```
a = random(A)
m_strength = I, I ∈ {1, 2, ..., 30}
input, model = sampler(I, l)
for all sample ∈ {input, label} do
    if rate_label < num_label / total_num then
        sample = a (sample, m_strength)
    end if
end for
2 * 2 Conv
for i = 1 to 3 do
    3 * 3 Conv
    GELU BN
end for
for Stage = 1 to 4 do
    2 * 2 Conv
    Local Perception Unit
    Layer Norm
    Cross-Window
    LayerNorm
    IRFFN
end for
```

TABLE 2. Parameters of each layer of CWTA.

Output size	Layer name	CWTA
112 × 112	Stem	3 × 3, 32, stride 2
56 × 56	Patch Aggr.	2 × 2, 64, stride 2
Stage 1	LPU Cross-Window IRFFN	$\begin{bmatrix} 3\times3, 64 \\ H_1 = 1, k_1 = 8 \\ R_1 = 4 \end{bmatrix} \times 3$
28 × 28	Patch Aggr.	2 × 2, 128, stride 2
Stage 2	LPU Cross-Window IRFFN	$\begin{bmatrix} 3\times3, 128 \\ H_2 = 2, k_2 = 4 \\ R_2 = 4 \end{bmatrix} \times 3$
14 × 14	Patch Aggr.	2 × 2, 256, stride 2
Stage 3	LPU Cross-Window IRFFN	$\begin{bmatrix} 3\times3, 256 \\ H_3 = 4, k_3 = 2 \\ R_3 = 4 \end{bmatrix} \times 16$
7 × 7	Patch Aggr.	2 × 2, 512, stride 2
Stage 4	LPU Cross-Window IRFFN	$\begin{bmatrix} 3\times3, 512 \\ H_4 = 8, k_4 = 1 \\ R_4 = 4 \end{bmatrix} \times 3$
1 × 1	FC	1 × 1, 1280
1 × 1	Classifier	1 × 1, 1000
#Params		25.14M
#FLOPs		4.04B

As shown in Table 2, the hyper-parameters of CWTA in different stages and layers are added. H_i and k_i are the number of heads and reduction rates in Cross-Window Self-Attention of stage i, respectively. R_i denote the expansion ratio in IRFFN of stage i. Patch Aggr. is the short for patch aggregation layer.

B. CWTA BLOCK

As shown in Figure 2, the CWTA Block includes a Local Perception Unit (LPU) [32], a Cross-Window Self-Attention mechanism, and a reverse residual feed forward network (IRFFN) [32]. These three parts will be described in detail.

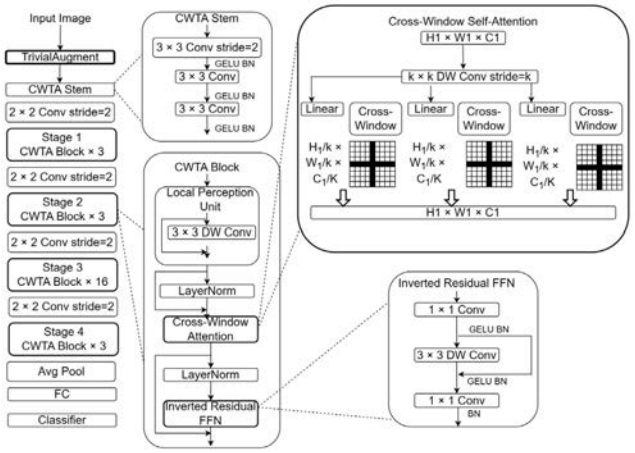

FIGURE 2. CWTA architecture for long-tailed datasets.

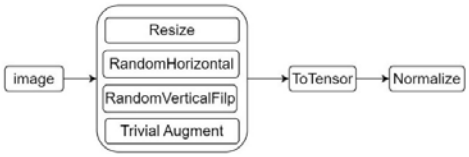

FIGURE 3. TrivialAugment preprocessing process for the few categories.

1) LOCAL PERCEPTION UNIT (LPU)

In visual tasks, it is necessary to design data enhancement methods based on translation invariance, such as rotation and displacement. However, in the transformer, absolute position coding is generally used to represent features. Because this method adds a unique position coding to each patch, it is easy to break translation invariance. In addition, Vision Transformer ignores the local relationship and structure information inside patches. In order to overcome these limitations, we, like CMT, use the Local Perception Unit (LPU) to extract local feature information. LPU is defined as follows Formula (1).

$$LPU\,(X) = DWConv\,(X) + X \tag{1}$$

where $X \in R^{H\times W\times d}$, $H \times W$ is the resolution of the current stage input image, and d is the feature dimension, $DWConv\,(\cdot)$ represents Depth Wise Convolution.

2) CROSS-WINDOW SELF-ATTENTION

二次复杂度

CSWin [33] mentioned that although the original Full-Attention has strong long-distance pixel modeling capability, its computational complexity is square with the size of the map.

In order to solve this problem, the existing research work suggests that self-attention should be carried out in the local attention window, and a halo or moving window should be used to expand the receptive field. However, more feature stacking means more computing overhead. In order to expand the attention area and obtain the global receptive field more effectively, the Cross-Window Self-Attention mechanism is proposed by improving to cross-shaped windows [33] proposed by CSWin. In order to expand the receptive field of attention and more effectively extract global feature information from images, the Cross-Window Self-Attention mechanism first performs linear operations, and then captures features in parallel in the horizontal and vertical directions, forming a cross-window. The Cross-Window Self-Attention mechanism is to divide multi heads into two parts, with half of the heads capturing horizontal attention and the other half capturing vertical attention, that is, the global feature information of an image can be obtained in a larger range, saving model run time and computational complexity.

To reduce the computational overhead, we used k depth convolution and step k for convolution operation, as shown in Figure 2. In this paper, linear and relative position code B [14] are added at the beginning of each self-attention header. Where $B \in R^{n\times \frac{n}{k^2}}$, and it is randomly initialized and learnable. According to the Multi-Head self-attention mechanism, the input characteristics $B \in R^{(H\times W)\times C}$ are first linearly projected to k heads. Each head is then locally self-attention in horizontal or vertical stripes. For horizontal stripes, self-attention X is evenly divided into equal width sw Non-overlapping horizontal stripes of $\left[X^1, \cdots, X^M\right]$. And each stripe contains $(sw \times W)$ tokens. Where, sw is the stripe width, which can be adjusted to balance the learning ability and computational complexity. Assume that the projection query, key, and value of k^{th} head are all dimension-s d_k, then the output of the horizontal stripe self-attention of the k^{th} head can be defined as Formulas (2 - 4).

$$X = \left[X^1, X^2, \cdots, X^M\right], \quad \left(X^i \in R^{(sw\times W)\times C}, M = H/sw\right) \tag{2}$$

$$Y_k^i = Attention\left(X^i W_k^Q, X^i W_k^K, X^i W_k^V\right), \quad (\mathrm{i} = 1, \cdots, M) \tag{3}$$

$$H \sim Attention_k\,(X) = B\left[Y_k^1, Y_k^2, \cdots, Y_k^M\right] \tag{4}$$

where $W_k^Q \in R^{C\times d_k}$, $W_k^K \in R^{C\times d_k}$, $W_k^V \in R^{C\times d_k}$ respectively the query, key, and worth projection matrix of k^{th} head, d_k set to C/K. Self-attention of vertical stripes can be deduced similarly, and the output of k^{th} head is expressed as $V \sim Attention_k\,(X)$.

The k heads are divided into two groups. Each group has $k/2$ heads, and K is usually an even number. The first group of heads executes self-attention of horizontal stripes, and the second group of heads executes self-attention of vertical stripes. Finally, on the output dimension, two different sets of outputs are combined. The Cross-Window Self-Attention Mechanism is shown in formula (5).

$$\begin{aligned} CrossWindow &\sim Attention\,(X) \\ &= Concat\,(head_1, \cdots, head_K)\,W^O B \\ head_K &= H \sim Attention_k\,(X), \quad k = 1, \cdots, K/2, \\ head_K &= V \sim Attention_k\,(X), \\ &\qquad k = K/2+1, \cdots, K. \end{aligned} \tag{5}$$

跨窗口自注意机制是将多个头部分成两部分：一半的头部捕捉水平注意，另一半捕捉垂直注意

Among them, $W^O \in R^{C\times C}$ is a common projection matrix used to project self-attention results in the target output dimension. The default setting of the target output dimension is C. As mentioned above, a key point of Cross-Window Self-Attention design is to perform the Liner operation before self-attention, add the relative position deviation B, and in the next part, divide the multi-head into two groups according to the horizontal and vertical directions to enter the characteristic receptive field at different angles.

3) INVERTED RESIDUAL FEED-FORWARD NETWORK (IRFFN)

The original FFN proposed in Vit consists of two linear layers, which are separated by GELU activation. The first floor will enlarge the size by 4 times, and the second floor will reduce the size by the same proportion. The calculation method of FFN is shown in Formula (6).

$$FFN\ (X) = GELU\ (XW_1 + b_1)\ W_2 + b_2 \tag{6}$$

where $W_1 \in R^{d\times 4d}$, $W_2 \in R^{4d\times d}$ represent the weights of two linear layers respectively. b_1 and b_2 are the deviation item. This paper uses IRFFN proposed in CMT, which is composed of an extension layer, Depth Wise Convolution layer, and a projection layer. IRFFN achieves better performance by changing the location of the short-net connection. The calculation method of IRFFN is shown in Formulas (7) and (8).

$$IRFFN(X) = Conv(F(Conv(X))) \tag{7}$$

$$F(X) = DWConv(X) + X \tag{8}$$

The active layer is omitted in Formula (8). At the beginning of Attention, Depth Wise Convolution is used to extract local information, which only adds a little computational overhead. By adding a shortcut, it provides a residual operation similar to ResNet to accelerate the propagation between network levels. Based on the above three parts, the CWTA block can be expressed as follows.

In Which X_i' an X_i'' represents the output characteristics of the LPU and cross of block i respectively, as shown in Formula (9).

$$\begin{aligned} X_i' &= LPU\ (X_{i-1}) \\ X_i'' &= Cross\left(LN\left(X_i'\right)\right) + X_i' \\ X_i &= IRFFN\left(LN\left(X_i''\right)\right) + X_i'' \end{aligned} \tag{9}$$

Standardized operations are performed through *LN*. In actual use, multiple blocks with different synchronization lengths can be stacked according to the picture resolution to ensure a balance between performance and speed. Through the standardization operation of *LN*, in practical use, multiple block stacking operations with different step sizes can be carried out according to the picture resolution to ensure the balance between performance and speed.

C. TRIVIALAUGMENT (TA)

TrivialAuigment (Tuning-free Yet State-of-the-Art Data Augmentation, TA) [34] is a simplest baseline and outperforms previous methods. TrivialAugment is parameter-free and only applies a single augmentation to each image. TrivialAugment uses augmentation similar to the previous methods [35], [36]. TA is defined as a function a that maps image x and discrete intensity parameter m to the enhanced image. Not all augmentation uses the intensity parameter, but most enhancements use it to define the intensity of image distortion.

TABLE 3. TrivialAugment algorithm enhancement procedure.

Algorithm 1 TrivialAugment Procedure
1: procedure TA(x: $image$)
2: Sample an augmentation a from A
3: Sample a strength m from $\{0,1,\cdots,30\}$
4: Return $a(x,m)$
5: end procedure

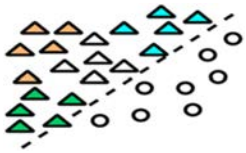

FIGURE 4. Visualization of two classes.

TrivialAugment works as follows. TrivialAugment takes the image x and a set of enhancement A as input. Then, it simply uniformly and randomly samples an enhancement from A, applies the enhancement to the given image x with intensity m, uniformly and randomly samples from the possible intensity set $\{0,\ldots,30\}$, and then returns the enhanced image. We summarize this very simple and parameter less process as pseudocode in Algorithm 1, as shown in Table 3. TA is not a special case of RandAugment (RA), because RA uses a fixed optimization intensity for all images, and TA re samples this intensity for each image.

Although the previous method used multiple subsequent enhancements, TA only applies one enhancement to each image. This allows the distribution of TA-enhanced data sets to be seen as the average of the $|A|$ data distribution generated by each expansion applied to the complete dataset. In Figure 4, we visualize this concept as a deterministic enhancement without strength parameters. Different from previous research work, we do not generate complex distributions from random combinations of augmented methods, but simply represent augmented data distributions applied to a given dataset.

IV. EXPERIMENTS

In order to verify the effect of CWTA method, we carried out long-tailed recognition contrast experiments in CIFAR100-LT, Places-LT, and ImageNet-LT. All experiments were trained on a device with 4 NVIDIA V100 16GB GPU.

A. DATASETS

1) CIFAR100-LT

CIFAR100-LT [37] is the benchmark data set for long-tailed identification. Like the original CIFAR dataset [38], the long-tailed version of the dataset also contains the same

FIGURE 5. Some image sample pictures (from left to right are: CIFAR100-LT, ImageNet-LT, Places-LT).

category. CIFAR100-LT via exponential function $n = n_t \times \mu^t$ is created by reducing the number of training samples for each class, where t is the class index, $u \in (0, 1)$, and n_t is the original number of training images. The test set remains unchanged. The imbalance factors of the long-tailed CIFAR dataset is defined as the number of training samples of the largest class divided by the number of training samples of the smallest class, ranging from 10 to 200.

2) ImageNet-LT

The ImageNet-LT [39] is derived from the original ImageNet [40]. By extracting a subset from 1000 categories according to the Pareto distribution, the maximum number of images in each category are 1280, and the minimum number of images in each category are 5.

3) PLACES-LT

Places [41] is a large-scale scene-centered dataset, and Places-LT is a long-tailed subset of Places that follows the Pareto distribution [34]. Places-LT contains a total of 62500 images from 365 categories, and the cardinality of categories ranges from 5 to 4980.

Figure 5 shows some image sample pictures in CIFAR100-LT, ImageNet-LT, and Places-LT dataset in the training model.

B. EXPERIMENTAL RESULTS

We conducted comparative experimental analysis on CIFAR100-LT, ImageNet-LT, and Places-LT datasets, comparing the classification accuracy. Next, we will introduce the experimental results on each dataset separately. Because some categories in the long-tailed distribution dataset may have only a few pictures, and some may have thousands of pictures, the commonly used Acc cannot effectively express the performance of the model, so the classification accuracy of different categories is given in the following paper.

- All: acc of all categories
- Many-shot: acc of categories with more than 100 pictures
- Medium-shot: acc of categories with the number of pictures between 20 and 100
- Few-shot: acc of categories with less than 20 pictures.

In addition, we also used F_1-Score to compare the performance of CWTA and other long-tailed classification models. Among them, some papers do not provide relevant parameters and cannot calculate the F_1-Score of their methods. For those models that cannot calculate F_1-Score, we use "-" instead of F_1-Score in the experimental results. The calculation formula for F_1-Score is shown in Formula (10). The P represents Precision, and the R represents Recall.

$$F_1 = \frac{2 * PR}{P + R} \tag{10}$$

TABLE 4. Comparison of accuracy on CIFAR100-LT.

Method	Backbone	Many	Medium	Few	All	F_1
OLTR[39]	ResNet-32	61.8	41.4	17.6	41.2	52.3
LDAM[42]	ResNet-32	61.5	41.7	20.2	42.0	52.9
cRT[17]	ResNet-32	64.0	44.8	18.1	43.3	51.9
RIDE[43]	ResNet-32	69.3	49.3	26.0	49.1	57.3
TADE[44]	ResNet-32	65.4	49.3	29.3	49.8	58.8
BALLAD[45]	ResNet-50	62.4	52.3	38.2	51.6	62.1
CMT[32]	ResNet-50	71.6	60.3	55.5	61.2	63.5
CWCT[16]	ResNet-50	77.4	61.1	56.3	63.5	67.9
CWTA	**ResNet-50**	**75.3**	**63.7**	**58.7**	**64.5**	**68.7**

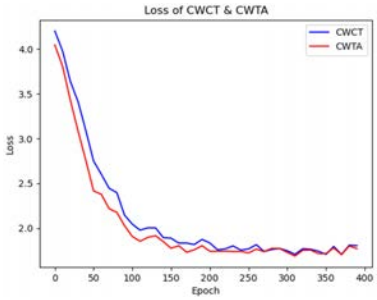

FIGURE 6. The change of CWCT and CWTA loss with Epoch on CIFAR100-LT.

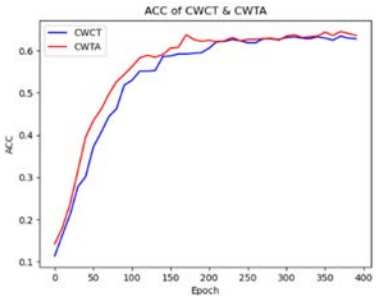

FIGURE 7. The accuracy of CWCT and CWTA on CIFAR100-LT varies with Epoch.

1) EXPERIMENTAL RESULTS ON CIFAR100-LT

Compared with the traditional long-tailed image classification algorithm, CWTA has improved its accuracy on CIFAR100-LT. The accuracy of CWTA in the few categories is 20.5% higher than that of BALLAD. The F_1-Score of CWTA is 6.6% higher than that of BALLAD. It can be analyzed from Table 4 that for the CIFAR100-LT dataset, when ResNet-50 is used as the Backbone, CWTA achieved the best results.

CWTA combines the advantages of CNN and Transformer, uses a Cross-Window Self-Attention mechanism for feature learning and introduces TrivialAugment for tail classes. Therefore, compared with the traditional CNN, the features extracted by CWTA are more balanced, and the classification accuracy is higher on the long-tailed dataset. Figure 6 shows the loss distribution of CWCT and CWTA in the training process. It can be found that CWTA introduced with TA can converge more quickly in the training process. It can be seen from Figure 7 that the final accuracy rate after CWTA training is higher.

TABLE 5. Comparison of accuracy on ImageNet-LT.

Method	Backbone	Many	Medium	Few	All	F_1
OLTR[39]	ResNeXt-50	43.2	35.1	18.5	35.6	47.6
τ -normalized[17]	ResNet-50	56.6	44.2	27.4	46.7	-
	ResNet-101	59.4	47.0	30.6	49.6	-
	ResNet-152	59.6	47.5	32.2	50.1	-
	ResNeXt-50	59.1	46.9	30.7	49.4	-
	ResNeXt-101	59.1	47.0	31.7	49.6	-
cRT[17]	ResNeXt-50	61.8	46.2	27.4	49.6	53.7
LWS[17]	ResNeXt-152	62.2	50.1	35.8	52.8	-
	ResNet-50	57.1	45.2	29.3	47.7	-
	ResNet-101	60.1	47.6	31.2	50.2	-
	ResNet-152	60.6	47.8	31.4	50.5	-
	ResNeXt-50	60.2	47.2	30.3	49.9	50.6
	ResNeXt-101	60.5	47.2	31.2	50.1	-
ResLT[46]	ResNeXt-152	63.5	50.4	34.2	53.3	-
	ResNeXt-50	63.0	50.5	35.5	52.9	55.2
Balanced	ResNeXt-101	63.3	53.3	40.3	55.1	-
Softmax[19]	ResNet-50	66.7	52.9	33.0	55.0	-
	ResNeXt-50	67.7	53.8	34.2	56.2	-
PaCo[47]	ResNeXt-101	69.2	55.8	36.3	58.0	-
	ResNet-50	65.0	55.7	38.2	57.0	62.3
	ResNeXt-50	67.5	56.9	36.7	58.2	-
BALLAD[45]	ResNeXt-101	68.2	58.7	41.0	60.0	-
	ResNet-50	71.0	66.3	59.5	67.2	66.0
CMT[33]	ResNet-50	49.6	44.3	41.2	45.6	47.7
CWCT[16]	ResNet-50	66.7	56.9	53.3	58.5	53.9
CWTA	**ResNet-50**	**72.3**	**66.5**	**58.7**	**67.6**	**68.2**

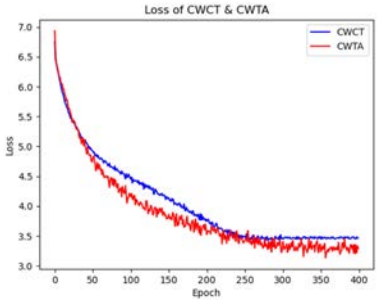

FIGURE 8. The change of loss of CWCT and CWTA on ImageNet-LT with Epoch.

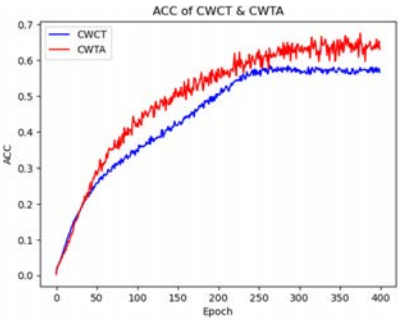

FIGURE 9. The accuracy of CWCT and CWTA on ImageNet-LT varies with Epoch.

2) EXPERIMENTAL RESULTS ON ImageNet-LT

Table 5 shows the result of the ImageNet-LT dataset, when ResNet-50 is used as the Backbone, the accuracy of CWTA in all categories is 0.4% higher than that of BALLAD. The F_1-Score of CWTA is 2.2% higher than that of BALLAD. Figure 8 shows the loss distribution of CWCT and CWTA during training. It can be found that CWTA introduced with TA can converge faster in the training process.

It can be seen from Figure 9 that the final accuracy rate after CWTA training is higher. Compared with the traditional long-tailed image classification algorithm, both CWCT and CWTA improve classification accuracy. CWTA combines CNN and vision transformer, and uses a Cross-Window Self-Attention mechanism for feature learning. Therefore, the features extracted by CWTA are more balanced and have better performance on the long-tailed datasets. After introducing TA for tail classes, CWTA performs better on the long-tailed datasets than CWCT.

TABLE 6. Comparison of accuracy on Places-LT.

Method	Backbone	Many	Medium	Few	All	F_1
OLTR[39]	ResNet-152	44.7	37.0	25.3	35.9	46.4
cRT[17]	ResNet-152	42	37.6	24.9	36.7	45.5
LWS[17]	ResNet-152	40.6	39.1	28.6	37.6	46.2
ResLT[46]	ResNet-152	39.8	43.6	31.4	39.8	51.2
PaCo[47]	ResNet-50	37.5	47.2	33.9	41.2	52.3
BALLAD[45]	ResNet-50	46.7	48.0	42.7	46.5	56.8
	ResNet-101	48.0	48.6	46.0	47.9	-
	ViT-B/16	49.3	50.2	48.4	49.5	-
CMT[32]	ResNet-50	52.0	25.4	20.7	32.5	49.8
CWCT[16]	ResNet-50	53.5	32.5	25.7	42.6	53.4
CWTA	**ResNet-50**	**54.1**	**47.5**	**40.9**	**47.8**	**58.3**

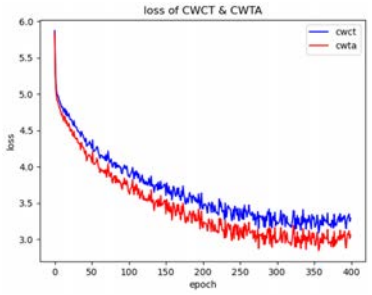

FIGURE 10. The changes of CWCT and CWTA loss with Epoch on Places-LT.

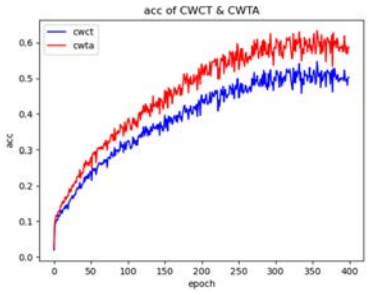

FIGURE 11. The accuracy of CWCT and CWTA on Places-LT varies with Epoch.

3) EXPERIMENTAL RESULTS ON PLACES-LT

Table 6 shows that for the Places-LT dataset when ResNet-50 is used as the Backbone, the accuracy of CWTA in all categories is 1.3% higher than that of BALLAD. The F_1-Score of CWTA is 1.5% higher than that of BALLAD.

Through the comparative experiment in Places-LT, we can draw the following conclusions: 1) CWCT combines CNN and vision transformer and uses a Cross-Window Self-Attention mechanism for feature learning. The extracted features are more balanced, and have better performance on long-tailed data than traditional CNN; 2) After the introduction of the TA layer, compared with CWCT, CWTA performs more significantly on the long-tailed data level. Figure 10 shows the loss distribution of CWCT and CWTA in the training process. It can be found that the CWTA introduced with TA can converge more quickly in the training process. It can be seen from Figure 11 that the final accuracy rate after CWTA training is higher; 3) Compared with the traditional long-tailed image classification algorithm, both CWCT and CWTA have improved accuracy.

TABLE 7. Results of ablation experiments on Cifar100-lt, ImageNet-LT, and Places-LT.

		CMT	CWCT	CMTA	CWTA
CIFAR100-LT	Many	71.6	77.4	74.3	**75.3**
	Medium	60.3	61.0	62.4	**63.7**
	Few	55.5	56.3	57.2	**58.7**
	All	61.2	63.4	63.9	**64.5**
ImageNet-LT	Many	49.6	66.7	51.3	**72.3**
	Medium	44.3	56.9	47.4	**66.5**
	Few	41.2	53.3	43.6	**58.6**
	All	45.6	58.5	47.2	**67.5**
Places-LT	Many	52.1	53.5	53.0	**54.1**
	Medium	25.4	32.4	30.3	**47.5**
	Few	20.7	25.6	23.9	**40.8**
	All	32.5	42.4	35.3	**47.8**

C. EXPERIMENT SUMMARY

Through comparative experiments on three long-tailed datasets, ImageNet-LT, CIFAR100-LT, and Places-LT, we can see that CWTA performs well in long-tailed image classification. It is worth mentioning that under the condition of using the same Backbone, CWTA improves the accuracy more significantly than the BALLAD method, which performs better in the field of long-tailed image classification at present. To sum up, the advantages of Transformer are used to capture the global features of the image, CNN is used to model the local features, and TrivialAugment is enhanced for tail categories. This method is effective for long-tailed image classification.

D. ABLATION STUDY

As shown in Table 7, we conducted an ablation study on CIFAR100-LT, ImageNet-LT, and Places-LT datasets respectively, and proved the role of Cross-Window Self- Attention mechanism and TrivialAugment on CWTA. CWCT is the research content of our previous article, which adds the innovative cross-window self-attention mechanism based on CMT. CMTA is to add innovative TrivialAugment based on CMT. The original CMT did not consider the characteristics of the long-tailed datasets, and the model was biased toward the head classes. CWCT has improved the classification accuracy of head classes and tail classes because it has extracted more image features using the cross-window selfattention mechanism, but it still favors the head classes. CMTA increased the intra-class diversity of tail classes data through TrivialAugment and improved the classification accuracy of tail classes, but the overall classification accuracy still lags behind that of CWTA. CWTA uses the cross-window self-attention mechanism and TrivialAugment at the same time, and the classification accuracy is significantly improved in all classes and tail classes.

V. CONCLUSION

In this paper, we propose a new long-tailed image classification method called CWTA. The core design of CWTA is to use the advantages of cross-window self-attention mechanism to capture the global features of images, use CNN to model local features, and enhance TrivialAugment data for tail categories. The effectiveness of the proposed CWTA method is verified by experiments on three commonly used long-tailed datasets.

Although the design of CWTA combines the advantages of CNN and Transformer, and enhances TA data for tail classes, it performs well in long-tail image classification. The essence of this method is still based on the data enhancement strategy, and its effect depends on the selection of enhancement methods. The final actual performance needs to be further improved.

Due to the complexity of Self-Attention Mechanism, CWTA cannot be executed as efficiently as pure convolutional neural network models in real-world industrial deployment scenarios. We will continue to optimize in the next work.

REFERENCES

[1] B. Zhou, Q. Cui, X. Wei, and Z. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9716–9725.

[2] M. A. Jamal, M. Brown, M. Yang, L. Wang, and B. Gong, "Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7607–7616.

[3] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, and J. Yan, "Equalization loss for long-tailed object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11659–11668.

[4] F. Saeed, A. Paul, M. J. Ahmed, M. J. J. Gul, W. Hong, and H. Seo, "Intelligent implementation of residential demand response using multiagent system and deep neural networks," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 22, p. e6168, Nov. 2021.

[5] F. Saeed, A. Paul, W. H. Hong, and H. Seo, "Machine learning based approach for multimedia surveillance during fire emergencies," *Multimedia Tools Appl.*, vol. 79, pp. 16201–16217, Jun. 2020.

[6] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," 2021, *arXiv:2110.04596*.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[8] F. Saeed, A. Paul, and H. Seo, "A hybrid channel-communication-enabled CNN-LSTM model for electricity load forecasting," *Energies*, vol. 15, no. 6, p. 2263, Mar. 2022.

[9] R. Girshick, "R-CNN: Region-based convolutional neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2014.

[10] H. A. Shah, F. Saeed, S. Yun, J. Park, A. Paul, and J. Kang, "A robust approach for brain tumor detection in magnetic resonance images using finetuned EfficientNet," *IEEE Access*, vol. 10, pp. 65426–65438, 2022.

[11] F. Saeed, M. J. Ahmed, M. J. Gul, K. J. Hong, A. Paul, and M. S. Kavitha, "A robust approach for industrial small-object detection using an improved faster regional convolutional neural network," *Sci. Rep.*, vol. 11, no. 1, p. 23390, Dec. 2021.

[12] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[13] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[15] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "UniFormer: Unifying convolution and self-attention for visual recognition," 2022, *arXiv:2201.09450*.

[16] M. Li, Y. Song, and B. Wang, "CWCT: An effective vision transformer using improved cross-window self-attention and CNN," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces Abstr. Workshops (VRW)*, Mar. 2022, pp. 149–154.

[17] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," 2019, *arXiv:1910.09217*.

[18] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 467–482.
[19] J. Ren, C. Yu, S. Sheng, X. Ma, H. Zhao, S. Yi, and H. Li, "Balanced meta-softmax for long-tailed visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4175–4186.
[20] Y. Zang, C. Huang, and C. Change Loy, "FASA: Feature augmentation and sampling adaptation for long-tailed instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3437–3446.
[21] Y. Zhang, P. Zhao, J. Cao, W. Ma, J. Huang, Q. Wu, and M. Tan, "Online adaptive asymmetric active learning for budgeted imbalanced data," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018.
[22] Y. Zhang, P. Zhao, S. Niu, Q. Wu, J. Cao, J. Huang, and M. Tan, "Online adaptive asymmetric active learning with limited budgets," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2680–2692, Jun. 2021.
[23] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang, "Disentangling label distribution for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6622–6632.
[24] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 694–710.
[25] J. Wang, T. Lukasiewicz, X. Hu, J. Cai, and Z. Xu, "RSG: A simple but effective module for learning imbalanced datasets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3783–3792.
[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[27] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.
[28] Z. Zhong, J. Cui, S. Liu, and J. Jia, "Improving calibration for long-tailed recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16484–16493.
[29] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan, "Remix: Rebalanced mixup," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 95–110.
[30] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "MetaSAug: Meta semantic augmentation for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5208–5217.
[31] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu, "Implicit semantic data augmentation for deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
[32] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12165–12175.
[33] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12114–12124.
[34] S. G. Müller and F. Hutter, "TrivialAugment: Tuning-free yet state-of-the-art data augmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 754–762.
[35] T. C. LingChen, A. Khonsari, A. Lashkari, M. R. Nazari, J. S. Sambee, and M. A. Nascimento, "UniformAugment: A search-free probabilistic data augmentation approach," 2020, *arXiv:2003.14348*.
[36] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3008–3017.
[37] Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9260–9269.
[38] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Tech. Rep., 2009, p. 7.
[39] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2532–2541.
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[41] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[42] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
[43] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu, "Long-tailed recognition by routing diverse distribution-aware experts," 2020, *arXiv:2010.01809*.
[44] Y. Zhang, B. Hooi, L. Hong, and J. Feng, "Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition," 2021, *arXiv:2107.09249*.
[45] T. Ma, S. Geng, M. Wang, J. Shao, J. Lu, H. Li, P. Gao, and Y. Qiao, "A simple long-tailed recognition baseline via vision-language model," 2021, *arXiv:2111.14745*.
[46] J. Cui, S. Liu, Z. Tian, Z. Zhong, and J. Jia, "ResLT: Residual learning for long-tailed recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3695–3706, Mar. 2023.
[47] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia, "Parametric contrastive learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 715–724.

YING SONG received the Ph.D. degree in computer engineering from the Institute of Computing Technology (ICT), Chinese Academy of Sciences. She is currently an Associate Professor with the Computer School, Beijing Information Science and Technology University. Her work has covered topics, such as performance modeling, resource management, cloud computing, and big data computing platform. She has been authored or coauthored more than 30 publications in these areas, since 2007. Her main research interests include computer architecture, parallel and distributed computing, and virtualization technology. She served for various academic conferences.

MENGXING LI received the bachelor's degree in software engineering from Beibaoding University, in 2019. She is currently pursuing the master's degree with the School of Computer Science, Beijing University of Information Science and Technology, China. Her research interest includes distributed storage.

BO WANG received the B.S. degree in computer science from Northeast Forest University (NEFU), Harbin, China, in 2010, and the Ph.D. degree in computer science from Xi'an Jiaotong University (XJTU), Xi'an, China, in 2017. He was a Guest Student with the State Key Laboratory of Computer Architecture, Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), from 2012 to 2016. He is currently a Lecturer with the Software Engineering College, Zhengzhou University of Light Industry (ZZULI). He has published more than ten research articles in these areas. His research interests include distributed systems, cloud computing, edge computing, resource management, and task scheduling. He served for various academic journals and conferences.