Augmenting Features via Contrastive Learning-based Generative Model for Long-Tailed Classification

Minho Park Hyung-Il Kim Hwa Jeon Song
Dong-oh Kang
Electronics and Telecommunications Research Institute (ETRI)
South Korea
{roger618, hikim, songhj, dongoh}@etri.re.kr

Abstract

Thanks to the advances in deep learning-based computer vision, image classification has shown great achievements. However, it has faced a heavy class imbalance issue which is one of the characteristics of real-world datasets. The severe class imbalance makes the classifier easily biased toward majority classes and overfitting to minority classes. To address this issue, supplementing minority classes with artificially generated samples has proven effective. In addition, contrastive learning has been introduced to improve image classification performance recently. Motivated by recent works, we propose feature augmentation via a contrastive learning-based generative model for long-tailed classification. Specifically, features are augmented using the feature dictionary obtained by real samples and the generated convex weights, which are used for learning an image classification model. Here, the model for the feature augmentation is trained based on generative adversarial learning and contrastive learning in an end-to-end manner. The generative adversarial learning helps to generate real-like features, and the contrastive learning improves the feature's discrimination power. Through extensive experiments with various long-tailed classification datasets, we verify the effectiveness of the proposed method.

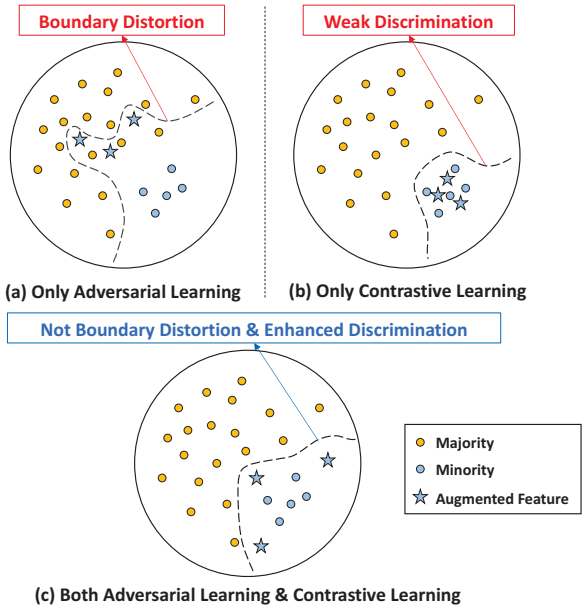

Figure 1. Illustration of embedding space depending on the learning type for feature augmentation: (a) adversarial learning, (b) contrastive learning, and (c) proposed method. In the case of (a), generated features cannot be clustered to each class well, i.e., boundary distortion. In the case of (b), diverse features cannot be generated, i.e., weak discrimination. However, the proposed method can augment diverse features without the loss of feature discrimination power.

1. Introduction

The development of deep learning and the ease of using large-scale datasets have led to significant progress in image classification [38, 10, 30]. However, since it is not easy to collect datasets with equally distributed samples per class, most large-scale datasets have class imbalance problems [30, 43]. A model trained based on these datasets can learn the representation for the majority classes (i.e., head classes) with a sufficient number of images, but features for images from the minority classes (i.e., tail classes) could not be effectively learned. This situation causes the classifier and the model to be trained biased toward the head classes, resulting in degradation of overall performance.

Several studies have been conducted to deal with these problems, which are mainly divided into two ways: re-weighting [9, 40, 29, 3, 51] and re-sampling [2, 4, 5, 24, 32, 35, 19]. The re-weighting is to adjust the weights of loss used in learning. This method aids learning for mi-

nority classes by adjusting the weights for a data sample or a specific class. As an improved version for dealing with the class imbalance problem of cross-entropy loss, a focal loss [29] is proposed, where it assigns greater weight to difficult or easily misclassified cases. In [3], label-distribution-aware margin (LDAM) loss is proposed, which is motivated by minimizing margin-based generalization bounds. As the re-sampling is to handle the amount of data samples directly, there are two major sampling methods: undersampling and oversampling [4]. The undersampling refers to erasing samples from the majority classes with enough training samples. However, overall performance degradation is inevitable because it causes the loss of data samples. In the oversampling, additional samples via generation are supplemented for minority classes. Since a model can be learned with enough samples for minority as well as majority classes, the oversampling strategy is preferred. Recent works [32, 34] show that generative adversarial networks (GANs) [15, 33] are effectively used in oversampling minority classes. A generative model generates artificial data and is used to augment data of minority classes. However, using GANs for oversampling can lead to boundary distortion [41], which ultimately results in decreased performance on the head classes. Therefore, the generative model should be encouraged to generate samples that do not cause boundary distortion while being included within the clusters of the minority class.

More recently, supervised contrastive learning that effectively leverages label information for better image representation has much attention, which has been adopted in the class imbalance problem [23, 45, 28]. Unlike conventional contrastive learning strategy that utilizes only the augmented version of an image as a positive sample, supervised contrastive learning takes advantage of all images in the same class as positive samples as well as the augmented version. According to [45], the experimental results show that the supervised contrastive learning-based method is effective with improved performance in the long-tailed classification problem. Nevertheless, solely training a generative model via contrastive learning results in limited diversity in the distribution of the generated features.

In this paper, we propose Feature Augmentation methodology using Contrastive learning-based Generative model (FACoG). FACoG synthesizes artificial features based on minority samples from the training dataset. The generated features are then trained to mimic the distribution of real features via generative adversarial learning while being clustered in the feature representation space through supervised contrastive learning. If the generated features are trained through generative adversarial learning and supervised contrastive learning in an end-to-end manner, the issues of boundary distortion and weak discrimination described earlier can be avoided. Figure 1 visually demonstrates these advantages. We newly design an augmented supervised contrastive loss in order to embed generated features into the real feature representation space. Also, we propose a two-phase training scheme to prevent the learning instability caused by adversarial learning in the generative model.

We experiment with the proposed model on various imbalanced benchmark datasets, including MNIST [27], Fashion-MNIST [48], CIFAR-10/100 [25] and ImageNet [10]. Experimental results show that the proposed model outperforms state-of-the-art methods on various long-tailed datasets. The main contributions of this paper are summarized as follows:

- We propose a novel feature augmenting method, FACoG based on generative adversarial learning and supervised contrastive learning for long-tailed classification.
- To stably train the feature generation model, we propose a two-phase training strategy. Also, we propose the augmented supervised contrastive loss to enable effective contrastive learning between the augmented features and real features.
- We analyze the effectiveness of our proposed model and the individual impact of its components on several long-tailed datasets, including large-scale datasets.

2. Related Work

2.1. Long-tailed Classification

According to the advances in deep learning-based applications, the demand for large-scale datasets is increasing. However, due to the expensive data acquisition process and the cost of labeling, the dataset easily has a class imbalance problem [30]. It makes representation learning for the minority classes difficult, causing the classifier to be biased toward the head classes. To address the class imbalance problem, there are two typical approaches to avoiding the class-imbalance problem: *re-weighting* and *re-sampling*.

Concerning the *re-weighting*, there have been many studies for modifying a loss function to mitigate the negative impact of minority classes on learning caused by the data imbalance [9, 40, 29, 3]. Since minority classes have a small number of data included in the training, the minority classes are intentionally weighted more in loss, allowing learning to be performed intensively. It artificially increases the proportion of losses to minority classes, leading to balanced learning between the majority and minority classes.

The *re-sampling* approach adjusts the class frequency on the long-tailed dataset to reduce the gap between the majority and minority classes, thereby mitigating the class imbalance [2, 4, 5, 24, 32, 19, 35]. There are two methods:

an undersampling method that intentionally eliminates data from head classes to match minority classes frequency, and an oversampling method that artificially generates data from minority classes to be comparable to head classes [14, 2]. Although the undersampling method is simple and easy to apply, it has a disadvantage in that the overall performance decreases because of intentional data loss. Therefore, recent studies have adopted the oversampling method [32, 34, 24]. In [32], authors propose a network that uses the adversarial training scheme to generate features depending on the class frequency. A major advantage is that it generates samples from the given training data without requiring additional data supplementation. However, the proposed generative model can only be applied to small-sized images and datasets, and the training process is unstable. In [19], the plug-in approach, sample-adaptive feature augmentation (SAFA) is proposed. SAFA proposes a method to extract the diversity and transferability of majority classes and apply them to the features of minority classes for oversampling. The advantage of plug-and-play approach is that it can be easily applied to improve the classifier, but the disadvantage is that the performance improvement is not significant. In this paper, to address the limitations of recent oversampling techniques, we propose a feature dictionary construction and a two-phase training scheme.

2.2. Contrastive Learning

Self-supervised learning has been actively studied to reduce the labeling cost of data or to extract the general-purpose features of images [44, 12, 13, 49]. It is known that general-purpose features obtained in this way greatly help improve the performance of various downstream tasks (e.g., image classification, detection, segmentation, etc.). In particular, contrastive learning-based self-supervised learning [16, 6] has shown remarkable performance improvements in the downstream tasks. Basically, contrastive learning keeps positive samples close to each other in an embedding space and negative samples away from each other. Here, positive samples refer to the augmented versions of the image and negative samples are all the remaining samples. The work in [6] presents a simple framework for contrastive learning of visual representations. In [16], Momentum Contrast (MoCo) is proposed, which involves using a contrastive loss to build large and consistent dictionaries for unsupervised learning. Recently, supervised contrastive learning that leverages label information for better image representation has shown significant performance improvement in image classification [23, 45, 28]. To construct positive samples, the authors in [23] utilize not only the augmented version of the image but also all other images of the class to which it belongs. Based on this method, the proposed model in [45] shows performance improvement in the long-tailed image classification by simultaneously performing supervised contrastive learning and classification using the curriculum learning [50].

The supervised contrastive loss is a loss function used to learn clustering of features extracted from real images in the feature representation space. To embed the generated features into the pre-clustered space, a modification of the loss function is required. We propose an augmented supervised contrastive loss to achieve this goal while maintaining the learning property of the supervised contrastive loss.

3. Method

3.1. Overview of the Proposed Method

In this section, we describe the proposed feature augmentation via contrastive learning-based generative model for long-tailed classification. The proposed FACoG generates convex weights that control the combination of image representations extracted from training images, thus generating augmented features to supplement samples for the minority classes. And, we devise an end-to-end training strategy to enhance the discrimination of generated features based on supervised contrastive learning and generative adversarial learning. Specifically, the proposed FACoG model is trained in two phases: 1) learning for constructing feature dictionary, and 2) learning for generative oversampling.

In the first phase, three modules (i.e., backbone, classifier, and projection networks) are trained by supervised contrastive learning. Then, the features for the training dataset are extracted with the fixed backbone network, which are stored as a dictionary form (i.e., feature dictionary). Note that the feature dictionary is a set of features for each class used to generate new features by convex weights. Through the classifier and projection networks, feature discrimination is enhanced by simultaneously learning class information and negative samples. In the second phase, the class selection is performed for each training iteration, which is for ensuring that minority classes are selected frequently. Then, features corresponding to the selected class in the feature dictionary are sampled. With sampled features, two modules, generator and discriminator, are trained by generative adversarial learning. Our generator generates convex weights for interpolating sampled features. A new feature is generated through the convex combination between the sampled features and the generated convex weights. The generated feature follows the distribution of the real sample through generative adversarial learning with the discriminator. Also, we propose an augmented supervised contrastive loss so that augmented features can be clustered into embedding space for each class.

3.2. Learning for Constructing Feature Dictionary

As mentioned above, three modules are trained to construct a feature dictionary. As a backbone network (f_{back}),

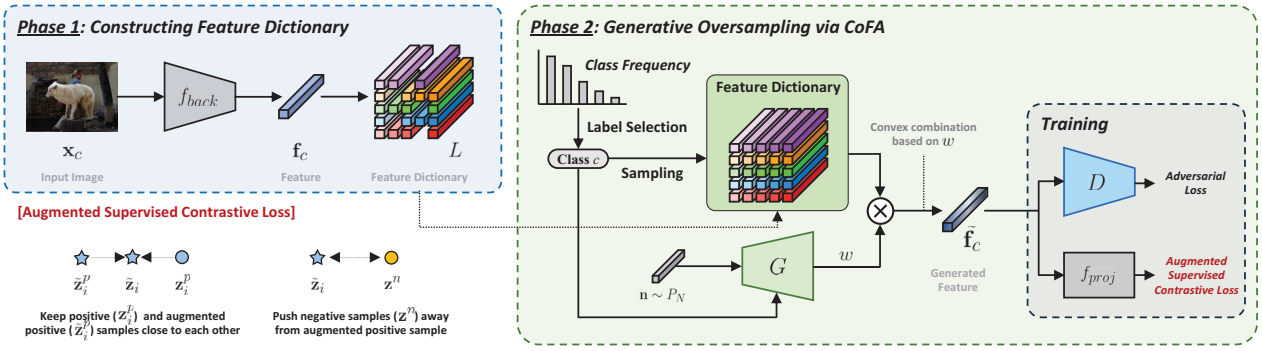

Figure 2. The overall architecture of the proposed learning process of FACoG. In the first phase, feature dictionary (L) is constructed via the backbone network (f_{back}). In the second phase, the class to be generated as a feature is selected through the label selection process of selecting the class with a probability that depends on the class frequency. The generator (G) that receives the selected label and gaussian random noise as input generates weights. For the selected label, the features are sampled in the dictionary and multiplied by convex weights. The parameters of generator are optimized by $\mathcal{L}_{adv}$ and $\mathcal{L}_{asc}$.

we adopt the ResNet [17], which is used for extracting a feature vector $\mathbf{f}_c = f_{back}(\mathbf{x}_c) \in \mathbb{R}^{D_E}$ from an input image $\mathbf{x}_c$ belonging to a class c in the entire class set $\mathcal{C}$. Then, the classifier $f_{cls} : \mathbb{R}^{D_E} \rightarrow \mathbb{R}^c$ with two fully connected layers followed by nonlinear activation returns the class information from the feature vector $\mathbf{f}_c$. And, the projection network $f_{proj} : \mathbb{R}^{D_E} \rightarrow \mathbb{R}^{D_P}$ transforms $\mathbf{f}_c$ to another feature vector $\mathbf{z}_c \in \mathbb{R}^{D_P}$ for a contrastive learning. Note that the structure of the projection network is set identically to that of the classifier except for the output dimension. The three modules are trained using cross-entropy loss $\mathcal{L}_{ce}$ and supervised contrastive loss $\mathcal{L}_{sc}$ The detail of the objective function of phase 1 is described in Section 3.4.

After training the three modules with $\mathcal{L}_{ph1}$, the parameters of f_{back} are frozen. Features $\mathbf{f}_c$ of input images for the c-th class are extracted by f_{back} to construct a feature dictionary, $L_c = [\mathbf{f}_{c,1}; \mathbf{f}_{c,2}; \cdots; \mathbf{f}_{c,N_s}] \in \mathbb{R}^{D_E \times N_s}$, where N_s is the number of features to be stored in the full dictionary L for the c-th class. The full dictionary L for all classes included in $\mathcal{C}$ is represented as $L = [L_1^\top; L_2^\top; \cdots; L_C^\top]$ by concatenating L_c's.

3.3. Learning for Generative Oversampling

Based on the constructed feature dictionary L, we augment features $\tilde{\mathbf{f}}_c$ for the minority classes (i.e., generative oversampling). To learn discriminative features that are similar to real features, we adopt generative adversarial learning and supervised contrastive learning. For generative adversarial learning, generator G for generating a convex weight w and discriminator D for discriminating whether features are from real distribution are devised. G and D consist of three fully-connected layers and the output of G goes through softmax to use as the weight of the convex combination. Note that the weight generator instead of generating features itself is considered for stable learning motivated by the previous work [32]. Specifically, our generator generates w with a random noise $\mathbf{n}$ and desired label c to be augmented, $w_c = G(\mathbf{n}|c)$, where the noise follows a standard normal distribution $\mathbf{n} \sim P_N$. The augmented feature for the desired class c can be represented as the convex combination with the features obtained by sampling N_s samples in L, $\tilde{\mathbf{f}}_c = w_c \cdot \mathbf{f}_c = \sum_{i=1}^{N_s} w_{c,i}\mathbf{f}_{c,i}$. $\tilde{\mathbf{f}}_c$ is learned to imitate the real feature through generative adversarial learning with D by exploiting generative adversarial loss $\mathcal{L}_{adv}$. In addition, $\tilde{\mathbf{f}}_c$ should not only be similar to the features for real samples but also preserve the discrimination power. To this end, we propose augmented supervised contrastive loss $\mathcal{L}_{asc}$, which differs from L_{sc} in that it considers positive samples for both real and augmented images.

We train the model for generative oversampling based on a total loss $\mathcal{L}_{ph2} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{asc}\mathcal{L}_{asc}$, where λ_{adv} and λ_{asc} are hyperparameters for controlling two loss functions, respectively. During training, $\mathbf{f}$ are randomly sampled and used within the dictionary for each iteration. Please refer to the detail of $\mathcal{L}_{ph2}$ in Section 3.4.

3.4. Training Details

Objective Functions In the first phase, $\mathcal{L}_{ce}$ refers to a conventional cross-entropy loss. And, we define $\mathcal{L}_{sc}$ as the extension of the unsupervised contrastive loss [6] like:

$$\mathcal{L}_{sc} = -\sum_{i \in I} \frac{1}{|\{\mathbf{z}_i^p\}|} \sum_{\mathbf{z}_j \in \{\mathbf{z}_i^p\}} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j/\tau)}{\sum_{\mathbf{z}_k, k \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_k/\tau)}, \tag{1}$$

where I and $|\cdot|$ denote the multi-viewed batch and the cardinality of a set, respectively. For the readability of the equation, we omit the notation for class c (i.e. $\mathbf{z}_{c,i} \rightarrow \mathbf{z}_i$) and denote the set of same class of $\mathbf{z}_i$ (i.e., positive samples) as $\{\mathbf{z}_i^p\}$. τ is a temperature parameter with a value greater than 0. The three modules (f_{back}, f_{cls}, and f_{prokj}) are trained as

a total loss $\mathcal{L}_{ph1}$ computed by the weighted sum of cross-entropy loss $\mathcal{L}_{ce}$ and supervised contrastive loss $\mathcal{L}_{sc}$ with hyperparameters λ_{ce} and λ_{sc}. The objective function of Phase 1 can be written as:

$$\mathcal{L}_{ph1} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{sc}\mathcal{L}_{sc}. \tag{2}$$

In the second phase, we compute $\mathcal{L}_{adv}$ by the least square formulation of generative adversarial loss to prevent gradient loss problems [31] as:

$$\mathcal{L}_{adv} = \min_G \max_D \mathbb{E}_{\mathbf{f}_c \sim p_c^d}[(1 - D(\mathbf{f}_c))^2] + \mathbb{E}_{\tilde{\mathbf{f}}_c \sim p_c^g}[(D(\tilde{\mathbf{f}}_c))^2], \tag{3}$$

where p_c^d and p_c^g are the conditional probability distribution of the real and generated features with class c, respectively. The proposed $\mathcal{L}_{asc}$ for training G is defined as follows:

$$\mathcal{L}_{asc} = -\sum_{i \in I} \frac{1}{|F(i)|} \sum_{\mathbf{z}_j \in F(i)} \log \frac{\exp(\tilde{\mathbf{z}}_i \cdot \mathbf{z}_j/\tau)}{\sum_{\mathbf{z}_k, k \neq i} \exp(\tilde{\mathbf{z}}_i \cdot \mathbf{z}_k/\tau)}, \tag{4}$$

where $\tilde{\mathbf{z}}_i$ is the output of f_{proj} with the input of $\tilde{\mathbf{f}}_i$, $\tilde{\mathbf{z}}_i = f_{proj}(\tilde{\mathbf{f}}_i)$. $\tilde{\mathbf{z}}_i^p$ indicates the positive sample corresponding to $\tilde{\mathbf{z}}_i$. $F(i) \equiv \{\mathbf{z}_i^p\} \cup \{\tilde{\mathbf{z}}_i^p\}$ is the set of positive samples of $\mathbf{z}_i$ and $\tilde{\mathbf{z}}_i$. The overall objective function of phase 2 ($\mathcal{L}_{ph2}$) can be written as:

$$\mathcal{L}_{ph2} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{asc}\mathcal{L}_{asc}. \tag{5}$$

In training, the parameter update for G and D, and the fine-tuning of the parameter of f_{cls} are alternately performed. To fine-tune f_{cls}, we calculate the cross-entropy loss for the augmented features and update the parameters of f_{cls} accordingly.

Label Selection When learning G, features for each class are chosen with the same probability because it must be able to infer convex weights well for all classes. On the other hand, because the oversampling is required only for minority classes in f_{cls} learning, more supplementation is required in inverse proportion to the class frequency. Therefore, the class with a small number of data should be selected frequently . To this end, the label selection process proceeds differently between G learning and f_{cls} learning. In this paper, we adopt the way that calculates the probability of selecting a label suggested in [32].

4. Experiments

4.1. Datasets

In this paper, we conducted extensive experiments on various benchmark datasets: MNIST [27], Fashion-MNIST [48], CIFAR-10/100 [25], and ImageNet [10]. Following the experimental setting in [30], we constructed long-tailed version of each dataset (i.e., MNIST-LT, Fashion-MNIST-LT, CIFAR-10/100-LT and ImageNet-LT).

MNIST-LT and Fashion-MNIST-LT Since the original MNIST and Fashion-MNIST datasets are balanced datasets, we constructed imbalanced datasets by sampling data by class according to the method in [32]. For comparison, we measured the performance with Average Class Specific Accuracy (ACSA) [47, 20] and Geometric Mean (GM) [26, 1], the metric used in works in [32] under the same experimental setting.

CIFAR-10-LT and CIFAR-100-LT As the original CIFAR-10 and 100 datasets are also balanced datasets, these consist of 50,000 training images and 10,000 validation images. And, each class contains the same number of images. As shown in [3, 9], we constructed the training set by reducing the number of images per class exponentially by the desired imbalance ratio, where the validation set remained. The imbalance ratio refers to the difference in the ratio of the number of data between the class with the most data and the class with the least data. In this paper, this value was set to 10, 50, and 100. For comparison, we measured the overall top-1 accuracy.

ImageNet-LT The long-tailed version of ImageNet dataset in [30] consists of 115.8K images for 1,000 classes and the imbalance ratio was set to 256. Following [30], we measured each top-1 accuracy by constructing three subsets according to the number of samples within the dataset: Many (number of samples more than 100 samples), Medium (number of samples not less than 20 and not more than 100), and Few (number of samples no more than 20).

4.2. Implementation Details

In all experiments, we used ResNet as the backbone, but we used different architectures for each dataset. We used ResNet-32 for MNIST-LT, Fashion-MNIST-LT and CIFAR-10/100-LT datasets. On the other hand, ResNeXt-50-32x4d was used for ImageNet-LT dataset. Dimensions of feature vector (D_E) were set to 512 and 2048 for ResNet-32 and ResNeXt-50-32x4d, respectively. Also, dimensions of projected feature (D_P) were set to 128 and 1024 for ResNet-32 and ResNeXt-50-32x4d, respectively. Implementation and learning of the proposed model used PyTorch [36] and SGD with a moment of 0.9 and weight decay of 1×10^{-4} as optimizers. The learning rates of phase 1 and 2 were set to 0.5 and 0.005, respectively. Also, the batch sizes of phase 1 and 2 were set to 256 and 128. The reason for reducing the batch size in phase 2 is for reducing memory overhead to sample several features in the dictionary and to perform a loss computation in parallel. In addition, experimentally setting it to 128 shows the best results. The sum of epochs in phase 1 and phase 2 was 200 for CIFAR-10/100-LT, and each epoch was heuristically selected. For ImageNet-LT, MNIST-LT and Fashion-MNIST-LT, overall epochs were set to 100. To control the impact between the $\mathcal{L}_{ce}$, $\mathcal{L}_{sc}$, $\mathcal{L}_{adv}$, and $\mathcal{L}_{asc}$, we set the value of λ as follows: $\lambda_{ce} = 1.0$, $\lambda_{sc} = 0.1$,

Table 1. Classification performance for the MNIST-LT and Fashion-MNIST-LT datasets.

Datasets	MNIST-LT		Fashion-MNIST-LT	
	ACSA	GM	ACSA	GM
CE	0.88	0.87	0.82	0.80
SMOTE [4]	0.88	0.87	0.82	0.80
DOS [32]	-	-	0.82	0.81
cGAN [37]	0.88	0.87	0.81	0.78
GAMO w/o D [32]	0.87	0.86	0.81	0.80
GAMO [32]	0.89	0.88	0.82	0.80
Hybrid-SC [45]	0.93	0.92	0.76	0.72
FACoG	**0.95**	**0.95**	**0.83**	**0.82**

$\lambda_{adv} = 1.0$, and $\lambda_{asc} = 0.1$. The temperature parameter τ is set to 0.1. Following [8], we used AutoAugment [7] and Cutout [11] for training of f_{cls} and SimAugment for training of f_{proj} for CIFAR-10/100-LT dataset. We traind the proposed model on a server with Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz, 512 GB memory, and NVIDIA A100 tensor core GPU.

4.3. Comparisons with State-of-the-art methods

In this section, we compare the proposed method to the state-of-the-art methods: re-weighting [29, 9, 3, 21, 39, 28, 18], re-sampling [24, 35, 19], contrastive learning [45], and others related to long-tailed classification [30, 22, 46]. Additionally, we set a baseline method that trains the backbone and classifier using cross-entropy loss (denoted as CE). For comparison, we conducted experiments with MNIST-LT, Fashion-MNIST-LT, CIFAR-10/100-LT, and ImageNet-LT datasets.

MNIST-LT and Fashion-MNIST-LT In this experiment, the recent oversampling and supervised contrastive learning-based methods are compared. Table 1 shows the top-1 accuracy (%) for MNIST-LT and Fashion-MNIST-LT datasets. The baseline shows the lowest performance because it does not consider the effect of the long-tailed datasets. As shown in Table 1, the proposed model outperforms state-of-the-art methods for all experiments.

CIFAR-10-LT and CIFAR-100-LT Table 2 shows the classification performance for CIFAR-10-LT and CIFAR-100-LT datasets. As one of the representative re-weighting methods, the Focal loss [29] shows a slightly higher performance than CE. We also experiment on models that combine several re-weighting (e.g., LDAM, KCL) and re-sampling (e.g., M2m, Hybrid-SC, TSC, RIDE+CMO, LDAM+SAFA) methods to consider class imbalance when sampling learning batches. Table 2 shows the top-1 accuracy (%) of state-of-the-art methods and the proposed model for CIFAR-10-LT and CIFAR-100-LT datasets with various imbalance ratios. It shows that the proposed model has higher performance than the existing model. The proposed model has an average performance improvement of 2-3% over the latest model performance.

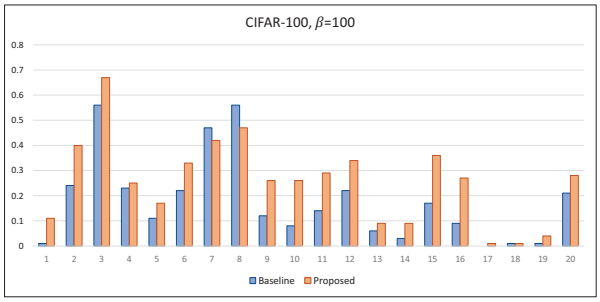

Figure 3. Top-1 accuracy on tail-classes in CIFAR-100-LT dataset with the imbalance ratio of 100 ($\beta = 100$). Here, the tail-classes refer to classes that correspond to the bottom 20% of the number of data in CIFAR-100-LT. The blue and orange bars indicate the results of Ours trained until phase 1 and phase 2, respectively. It can be seen that the performance has improved through generative oversampling in most tail-classes.

ImageNet-LT To evaluate the proposed method in a large-scale dataset, we used the ImageNet-LT dataset. As shown in Table 3, FACoG outperformed SOTA model on top-1 accuracy for overall datasets. While there was a slight 0.2% performance drop compared to RIDE (4 experts) for the "Many" subset, there were notable improvements of 1.4% and 2.9% for the "Medium" and "Few" subsets, respectively. Particularly, a significant performance improvement was observed for the 'Few' subset, which corresponds to the tail class.

4.4. Effectiveness of FACoG on Tail-classes

In this paper, we proposed the FACoG to improve performance in long-tailed classification by generating artificial samples and complementing data for minority classes. To investigate the effect of the proposed method for tail-classes, we conducted the experiment for validating top-1 accuracy for each tail-class with a lower 20% number of data in the CIFAR-100-LT dataset. The baseline and the proposed model are models without and with the classifier fine-tuning process through FACoG, respectively. Figure 3 shows the performance of tail-classes as a histogram when the imbalance ratio of the CIFAR-100-LT dataset was set to 100. Blue and orange bar means the results of the proposed model trained until phase 1 and phase 2, respectively. As shown in Fig. 3, the proposed model in most tail-classes showed higher performance than the baseline model. When the performance for tail-classes was averaged, the baseline model showed 17.7%, the proposed model showed 25.6%, and the average performance improvement was 7.9%. Through this experiment, it can be seen that the performance improvement of minority classes

Table 2. Classification performance for the CIFAR-10-LT and CIFAR-100-LT. The highest and second-highest results are marked in bold.

Datasets	**CIFAR-10-LT**			**CIFAR-100-LT**		
Imbalance Ratio (β)	100	50	10	100	50	10
CE	70.36	74.81	86.39	38.32	43.85	55.71
Focal loss [29]	70.38	76.72	86.66	38.41	44.32	55.78
CB-Focal [9]	74.57	79.27	87.10	39.60	45.17	57.99
CE-DRW [3]	76.34	79.97	87.56	41.51	45.29	58.12
LDAM [3]	73.40	76.80	87.00	39.60	45.00	56.90
M2m-LDAM [24]	79.10	-	87.50	43.50	-	57.60
KCL [21]	77.60	81.70	88.00	42.80	46.30	57.60
Hybrid-SC [45]	**81.40**	**85.36**	**91.12**	46.72	51.87	**63.05**
TSC [28]	79.70	82.90	88.70	43.80	47.40	59.00
RIDE+CMO [35]	-	-	-	**50.00**	**53.00**	60.20
LDAM+SAFA [19]	80.48	83.57	88.94	46.04	50.02	59.11
FACoG	**83.90**	**87.86**	**92.22**	**51.34**	**56.64**	**68.11**

Table 3. Classification performance for the ImageNet-LT dataset. The highest results are marked in bold.

Method	**Many**	**Medium**	**Few**	**All**
CE	65.9	37.5	7.7	44.4
Focal [29]	63.3	37.4	7.7	43.2
OLTR [30]	52.1	39.7	20.3	41.2
τ-norm [22]	59.1	46.9	30.7	49.4
Balanced Softmax [39]	62.2	48.8	29.8	51.4
LWS [22]	60.2	47.2	30.3	49.9
LADE [18]	62.3	49.3	31.2	51.9
RIDE (4 experts) [46]	**68.2**	53.8	36.0	56.8
LDAM-SAFA [19]	63.8	49.9	33.4	53.1
FACoG	68.0	**55.2**	**38.9**	**57.8**

through generative oversampling was well achieved.

4.5. Ablation Studies

To verify the effectiveness of the proposed method, we conducted the ablation studies: 1) effect of loss functions used for training the proposed method, and 2) effect of additional training strategies.

$\mathcal{L}_{adv}$ and $\mathcal{L}_{asc}$ Table 4 shows the top-1 accuracy obtained by varying the loss type. The baseline is the model that has been trained up to phase 1 named as "Phase1". When $\mathcal{L}_{adv}$ and $\mathcal{L}_{asc}$ are used individually, the performance improvement is insignificant or rather reduced. In the case of $\mathcal{L}_{adv}$, there is little performance improvement because the augmentation feature constitutes different distributions from the real feature. In the case of $\mathcal{L}_{asc}$, the augmented feature imitates the real feature well, so the performance has improved to some extent, but it is not large. On the other hand, models using both $\mathcal{L}_{adv}$ and $\mathcal{L}_{asc}$ effectively improve performance.

Table 4. Performance for analyzing the contribution of $\mathcal{L}_{adv}$ and $\mathcal{L}_{asc}$ for the proposed method on CIFAR-100-LT datasets.

	CIFAR-100-LT		
Imbalance Ratio (β)	100	50	10
Phase1	48.35	53.00	67.39
Phase1+$\mathcal{L}_{asc}$	49.03	53.28	66.76
Phase1+$\mathcal{L}_{adv}$	49.73	54.21	67.25
Phase1+$\mathcal{L}_{adv}$+$\mathcal{L}_{asc}$	**49.85**	**55.24**	**68.17**

Re-weight and Data Augmentation Table 5 reports top-1 accuracy depending on whether the re-weighting method and data augmentation strategy are used. "Rew." refers the model that has been trained with Deferred Re-Weighting (DRW) [3] method. "Aug." means the model that uses AutoAugment [7] and Cutout [11] for data augmentation strategies. There is little difference in performance for models learned simply by the DRW method, but when used with FACoG, it helps to improve performance. In addition, data augmentation strategies are very helpful in improving overall performance, especially for models learned with FACoG, which significantly increase their accuracy. When real features have various distributions through various data augmentation methods, more diverse augmentation features are generated through FACoG, enabling effective minority oversampling. Finally, when both DRW and augmentation techniques are used, the greatest performance improvement is shown.

4.6. Visualization Results

To analyze the effect of the proposed feature augmentation qualitatively, we visualize the feature space for the Fashion-MNIST-LT dataset using t-SNE [42]. In Fig. 4, the left figure shows the embedding space for the real fea-

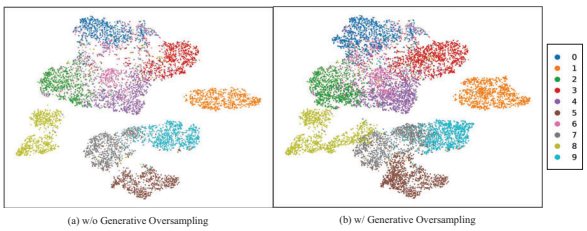

Figure 4. t-SNE visualization results of the proposed model without and with generative oversampling in Fashion-MNIST-LT dataset. (a) presents real features from the proposed model without generative oversampling. (b) shows real and augmented features by the proposed model with generative oversampling.

Table 5. Performance for analyzing the contribution of re-weight, augmentation strategies and FACoG. "Rew." refers the model that has been trained with re-weighting method. "Aug." is the model that uses AutoAugment and CutMix for data augmentation strategies.

Rew.	Aug.	FACoG	CIFAR-100-LT		
			100	50	10
-	-	-	48.08	53.77	65.45
✓	-	-	47.95	53.98	65.85
-	✓	-	48.31	54.38	66.22
-	-	✓	49.07	55.19	62.93
✓	-	✓	48.90	55.20	65.20
-	✓	✓	49.85	55.24	**68.17**
✓	✓	✓	**51.34**	**56.64**	68.11

tures without the proposed feature augmentation. And, the real and augmented features by the proposed method are visualized in the right figure. Through the visualization results, we observe that the proposed method generates well-distributed features within each class cluster. In other words, this result shows that the classification performance can be improved by the proposed method as the diversity is improved without loss of discrimination power.

4.7. Effect of N_s

In this section, we analyze the trend of classification performance changes for the hyperparameter N_s. As introduced in Section 3.2, N_s is the number of reference features used for the feature augmentation from real samples. Since the generator generates features within the convex hull formed by the real data of the selected class, training becomes difficult when many samples are included. In addition, the training time increases in proportion to the number of real samples. Therefore, choosing this hyperparameter properly is one of important training heuristics. Figure 5 shows the classification performance (top-1 accuracy) and

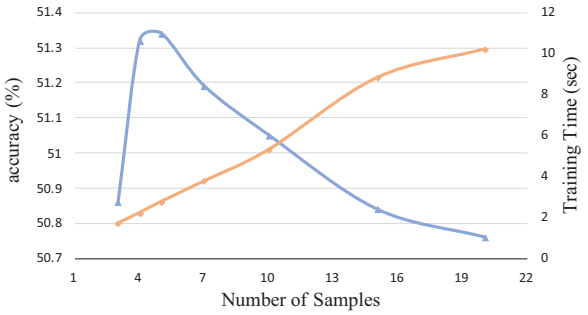

Figure 5. Results of top-1 accuracy and training time obtained by varying N_s on CIFAR-100-LT dataset with 100 imbalance ratio. The values on the x-axis refer to N_s. The left and right sides of the y-axis refer to top-1 accuracy and training time, respectively.

training time (sec) for CIFAR-100-LT ($\beta = 100$) depending on N_s. As expected, the training time increases linearly with the number of real samples. However, we observe that it is desirable to select the hyperparameter ranging from 4 to 6 that shows high performance. From these experimental results, we set N_s as 5.

5. Conclusion

In this paper, we proposed a novel Feature Augmentation methodology using Contrastive learning-based Generative model named as FACoG. The proposed method supplements the number of samples for minority classes by features through generative adversarial learning. At the same time, it prevents boundary distortion by ensuring that the generated features cluster within the feature representation space of the minority class through supervised contrastive learning. To effectively perform supervised contrastive learning on the generated features and features extracted from the training data, we propose the augmented supervised contrastive loss. To ensure the stable training of FACoG, we divided the training process into two phases. First, we built a feature dictionary through pretraining of the backbone network and then trained a generative model based on it. We demonstrate the effectiveness of FACoG through extensive experiments on various long-tailed classification benchmarks.

Acknowledgement

This work was partly supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (No.23ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System, 70%) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00907, Development of AI Bots Collaboration Platform and Self-organizing AI, 30%).

References

[1] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

[2] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[5] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

[8] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021.

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[13] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014.

[14] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8, 2003.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[18] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021.

[19] Yan Hong, Jianfu Zhang, Zhongyi Sun, and Ke Yan. Safa: Sample-adaptive feature augmentation for long-tailed image classification. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 587–603. Springer, 2022.

[20] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[21] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020.

[22] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[24] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13896–13905, 2020.

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[26] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, page 179. Nashville, USA, 1997.

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6918–6928, 2022.

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[30] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[31] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[32] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1695–1704, 2019.

[33] Minho Park, Hak Gu Kim, and Yong Man Ro. Generative guiding block: Synthesizing realistic looking variants capable of even large change demands. In *IEEE International Conference on Image Processing*, pages 4444–4448. IEEE, 2019.

[34] Minho Park, Hwa Jeon Song, and Dong-Oh Kang. Imbalanced classification via feature dictionary-based minority oversampling. *IEEE Access*, 10:34236–34245, 2022.

[35] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6887–6896, 2022.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[38] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *arXiv preprint arXiv:1705.08045*, 2017.

[39] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

[40] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018.

[41] Shibani Santurkar, Ludwig Schmidt, and Aleksander Madry. A classification-based study of covariate shift in gan distributions. In *International Conference on Machine Learning*, pages 4480–4489. PMLR, 2018.

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[43] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[44] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[45] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 943–952, 2021.

[46] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

[47] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017.

[48] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[49] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

[50] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020.

[51] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2022.