Generalized Parametric Contrastive Learning

Jiequan Cui, Student Member, IEEE, Zhisheng Zhong, Student Member, IEEE, Zhuotao Tian, Student Member, IEEE, Shu Liu, Member, IEEE, Bei Yu, Member, IEEE, and Jiaya Jia, Fellow, IEEE

Abstract—In this paper, we propose the Generalized Parametric Contrastive Learning (GPaCo/PaCo) which works well on both imbalanced and balanced data. Based on theoretical analysis, we observe supervised contrastive loss tends to bias on high-frequency classes and thus increases the difficulty of imbalanced learning. We introduce a set of parametric class-wise learnable centers to rebalance from an optimization perspective. Further, we analyze our GPaCo/PaCo loss under a balanced setting. Our analysis demonstrates that GPaCo/PaCo can adaptively enhance the intensity of pushing samples of the same class close as more samples are pulled together with their corresponding centers and benefit hard example learning. Experiments on long-tailed benchmarks manifest the new state-of-the-art for long-tailed recognition. On full ImageNet, models from CNNs to vision transformers trained with GPaCo loss show better generalization performance and stronger robustness compared with MAE models. Moreover, GPaCo can be applied to semantic segmentation task and obvious improvements are observed on 4 most popular benchmarks.

Index Terms—Representation learning, contrastive learning, OOD robustness, long-tailed recognition, semantic segmentation.

I. INTRODUCTION

CONVOLUTIONAL neural networks (CNNs) have achieved great success in various tasks, including image classification [1, 2], object detection [3, 4] and semantic segmentation [5]. Especially, with the rise of neural network search [6], [7], [8], [9], [10], performance of CNNs have further taken a big step. However, The huge progress highly depends on large-scale and high-quality datasets, such as ImageNet [11], MS COCO [12] and Places [13]. But when deal with real-world applications, generally we face the long-tailed distribution problem – a few classes contain many instances, while most classes contain only a few instances. Learning in such an imbalanced setting is challenging as the low-frequency classes can be easily overwhelmed by high-frequency ones. Without considering this situation, CNNs will suffer from significant performance degradation.

Contrastive learning [15], [16], [17], [18], [19] is a major research topic due to its success in self-supervised representation learning. Khosla [20] extend non-parametric contrastive loss into non-parametric supervised contrastive loss by leveraging label information, which trains representation in the first stage and learns the linear classifier with the fixed backbone in the second stage. Though supervised contrastive learning works well in a balanced setting, for imbalanced datasets, our theoretical analysis shows that high-frequency classes will have a higher lower bound of loss and contribute much higher importance than low-frequency classes when equipping it in training. This phenomenon leads to model bias on high-frequency classes and increases the difficulty of imbalanced learning. As shown in Fig. 2, when the model is trained with supervised contrastive loss on ImageNet-LT, the gradient norm varying from the most frequent class to the least one is rather steep. In particular, the gradient norm dramatically decreases for the top 200 most frequent classes.

Previous work [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33] explored to rebalance in traditional supervised cross-entropy learning. In this paper, we tackle the above mentioned imbalance issue in supervised contrastive learning and make use of contrastive learning for long-tailed recognition.

To rebalance in supervised contrastive learning, we introduce a set of parametric class-wise learnable centers into supervised contrastive learning. We name our algorithm Parametric Contrastive Learning (PaCo) shown in Fig. 3(a). With such a simple and yet effective operation, we theoretically prove that the optimal values for the probability that two samples are a true positive pair (belonging to the same class), varying from the most frequent class to the least frequent class, are more balanced. Thus their lower bound of loss values are better organized. This phenomenon means the model takes more care of low-frequency classes, making the PaCo loss benefit imbalanced learning. Fig. 2 shows that, with our PaCo loss in training, gradient norm varying from the most frequent class to the least one are moderated better than supervised contrastive learning, which matches our analysis.

Further, we analyze the PaCo loss under a balanced setting. Our analysis demonstrates that with more samples clustered around their corresponding centers in training, the PaCo loss...
GPaCo demonstrates impressive performance on imbalanced and balanced data. (a) shows that GPaCo achieves state-of-the-art on long-tailed classification. GPaCo benefits semantic segmentation task and significant improvement is observed on ADE20K in (b). In (c), top-1 accuracy are reported for full ImageNet, ImageNet-R, and ImageNet-S. mCE and rel. mCE are used for ImageNet-C. Compared with MAE models, GPaCo enjoys better generalization ability and stronger robustness on ImageNet and its variants.

Fig. 2. Rebalance in contrastive learning. We collect the average L2 norm of the gradient of weights in the last classifier layer on ImageNet-LT. Category indices are sorted by their image counts. The gradient norm varying from the most frequent class to the least one is steep for supervised contrastive learning [20]. In particular, the gradient norm dramatically decreases for the top 200 most frequent classes. Trained with PaCo, the gradient norm is better balanced.

increases the intensity of pushing samples of the same class close, which benefits hard examples learning.

MoCo [16] enables small mini-batch training for self-supervised contrastive learning with a momentum encoder and a queue. Chen et al. [34] claims that it is the “stop-gradient” but not the momentum encoder that is necessary to avoid collapsing solutions for self-supervised learning. We examine the effects of each component in PaCo including augmentation strategy, loss function (relation to multi-task), momentum encoder and queue size. Interestingly, the simplified Generalized Parametric Contrastive Learning (GPaCo) by removing the momentum encoder even boosts model performance and robustness, indicating the unnecessary of momentum encoder in PaCo framework.

Finally, we conduct experiments on imbalanced data including long-tailed version of CIFAR [23], [35], ImageNet [36], Places [36] and iNaturalist 2018 [37]. Experimental results show that we create a new record for long-tailed recognition. On balanced data, We testify the effectiveness of GPaCo on full ImageNet [11] and CIFAR [38]. Compared with MAE models, we achieve better generalization ability and stronger robustness. With GPaCo loss for semantic segmentation, obvious improvements are obtained on popular benchmarks including ADE20K [39], COCO-Stuff [40], PASCAL Context [41] and Cityscapes [42]. GPaCo demonstrates its great generality with various tasks as shown in Fig. 1. Our key contributions are as follows.

- We identify the shortcoming of supervised contrastive learning under an imbalanced setting – it tends to bias high-frequency classes.
- We extend supervised contrastive loss to the PaCo loss, which is more friendly to imbalanced learning, by introducing a set of parametric class-wise learnable centers.
- We examine the necessary of components in PaCo and simplify PaCo to GPaCo, observing that the momentum encoder can hurt model performance and should be removed.
- GPaCo can benefit the training on imbalanced data and balanced data varying from CNNs to vision transformers. Experiments on tasks e.g., long-tailed classification, full ImageNet/CIFAR classification, out-of-distribution robustness, and semantic segmentation demonstrate the generality of GPaCo.

II. RELATED WORK

Re-Sampling/Re-Weighting: The most classical way to deal with long-tailed datasets is to over-sample low-frequency class images [21], [26], [43], [44] or under-sample high-frequency class images [21], [24], [45]. However, Oversampling can suffer from heavy over-fitting to low-frequency classes especially on small datasets. For under-sampling, discarding a large portion of high-frequency class data inevitably causes degradation of the generalization ability of CNNs. Re-weighting [22], [46], [47],
FIG. 3. Framework of Generalized Parametric Contrastive Learning (GPaCo/PaCo) and $L_{extra}$ curve. (a) illustrates the details of GPaCo/PaCo. PaCo incorporates a momentum encoder, whereas GPaCo eliminates it. We introduce a set of parametric class-wise learnable centers for rebalancing in contrastive learning. More analysis is in Section III-C for PaCo and Section III-F for GPaCo. (b) plots the curve of $L_{extra}$ instantiated with full ImageNet. More analysis on balanced data is in Section III-D.

[48], [49], [50] the loss functions is an alternative way to re-balance by either enlarging weights on more challenging and sparse classes or randomly ignoring gradients from high-frequency classes [51]. However, with large-scale data, re-weighting makes CNNs difficult to optimize during training [22], [46].

One/Two-Stage Methods: Since deferred re-weighting and re-sampling were proposed by Cao [35], Kang [28] and Zhou [52] observed re-weighting or re-sampling strategies could benefit classifier learning while hurting representation learning. Kang [28] proposed to decompose representation and classifier learning. It first trains the CNNs with uniform sampling, and then fine-tune the classifier with class-balanced sampling while keeping parameters of representation learning fixed. Zhou [52] proposed one cumulative learning strategy, with which they bridge representation learning and classifier re-balancing.

The two-stage design is not for end-to-end frameworks. Tang [31] analyzed the reason from the perspective of causal graph and concluded that the bad momentum causal effects played a vital role. Cui [30] proposed residual learning mechanism to address this issue.

Non-Parametric Contrastive Loss: Contrastive learning [15], [16], [17], [18], [19] is a framework that learns similar/dissimilar representations from data that are organized into similar/dissimilar pairs. An effective contrastive loss function, called InfoNCE [53], is

$$L_{q,k^+,k^-} = -\log \frac{\exp(q \cdot k^+ / \tau)}{\exp(q \cdot k^+ / \tau) + \sum_k \exp(q \cdot k^- / \tau)},$$

where $q$ is a query representation, $k^+$ is for the positive (similar) key sample, and $\{k^-\}$ denotes negative (dissimilar) key samples. $\tau$ is a temperature hyper-parameter. In the instance discrimination pretext task [54] for self-supervised learning, a query and a key form a positive pair if they are data-augmented versions of the same image. It forms a negative pair otherwise.

Traditional cross-entropy with linear fc layer weight $w$ and true label $y$ among $n$ classes is expressed as

$$L_{cross-entropy} = -\log \frac{\exp(q \cdot w_y)}{\sum_{i=1}^{n} \exp(q \cdot w_i)}.$$
In implementation, the loss is usually scaled by \( \frac{1}{|P(i)|} \) and a temperature \( \tau \) is applied like in (1). Different from self-supervised contrastive loss, which treats query and key as a positive pair if they are the data-augmented version of the same image, supervised contrastive loss treats them as one positive pair if they belong to the same class.

**B. Theoretical Motivation**

**Analysis of Supervised Contrastive Learning:** Khosla [20] introduced supervised contrastive learning to encourage more compact representation. We observe that it is not directly applicable to long-tailed recognition. As shown in Table I, the performance significantly decreases compared with traditional supervised cross-entropy. From an optimization point of view, supervised contrastive loss concentrates more on high-frequency classes than low-frequency ones, which is unfriendly for imbalanced learning.

**Remark 1:** (Optimal value for supervised contrastive learning) When supervised contrastive loss converges, the optimal value for the probability that two samples are a true positive pair with label \( y \) is \( \frac{1}{K_y} \), where, \( q(y) \) is the frequency of class \( y \) over the whole dataset, \( queue \) is the momentum queue in MoCo [16], [17] and \( K_y \approx length(queue) \cdot q(y) \).

**Interpretation:** As indicated by Remark 1, high-frequency classes have a higher lower bound of loss value and contribute much more importance than low-frequency classes in training. Thus the training process can be dominated by high-frequency classes. To handle this issue, we introduce a set of parametric class-wise learnable centers for rebalancing in contrastive learning.

**C. Rebalance in Contrastive Learning**

As described in Fig. 3(a), we introduce a set of parametric class-wise learnable centers \( C = \{ c_1, c_2, \ldots, c_n \} \) into the original supervised contrastive learning, and called this new form **Parametric Contrastive Learning (PaCo).**

Correspondingly, the loss function is changed to

\[
\mathcal{L}_i = \sum_{z_+ \in P(i) \cup \{ c_y \}} -w(z_+) \log \frac{\exp(z_+ \cdot T(x_i))}{\sum_{z_k \in A(i) \cup C} \exp(z_k \cdot T(x_i))},
\]

where

\[
w(z_+) = \begin{cases} 
\alpha, & z_+ \in P(i) \\
1.0, & z_+ \in \{ c_y \} 
\end{cases}
\]

and

\[
z \cdot T(x_i) = \begin{cases} 
z \cdot G(x_i), & z \in A(i) \\
z \cdot F(x_i), & z \in C.
\end{cases}
\]

Following Chen [17], the transform \( G(\cdot) \) is a two-layer MLP while \( F(\cdot) \) is the identity mapping, i.e., \( F(x) = x \). \( \alpha \) is one hyper-parameter in \((0,1)\). \( P(i) \) and \( A(i) \) are the same with supervised contrastive learning in (3). In implementation, the loss is scaled by \( \frac{1}{\sum_{z_+ \in P(i) \cup \{ c_y \}} w(z_+)} \) and a temperature \( \tau \) is applied like in (3).

**Remark 2:** (Optimal value for parametric contrastive learning) When parametric contrastive loss converges, the optimal value for the frequency of class \( y \) over the whole dataset, \( q(y) \) is the frequency of class \( y \) and the optimal value for the probability that a sample is closest to its corresponding center \( c_y \) among \( C \) is \( \frac{1}{1 + \alpha \cdot K_y} \), where \( q(y) \) is the frequency of class \( y \).

**Interpretation:** Suppose the most frequent class \( y_h \) has \( K_{y_h} \approx q(y_h) \cdot length(queue) \) and the least frequent class \( y_l \) has \( K_{y_l} \approx q(y_l) \cdot length(queue) \). As indicated by Remarks 2 and 1, the optimal value for the probability that two samples are a true positive pair varying from the most frequent class to the least one is rebalanced from

\[
\frac{1}{K_{y_h}} \rightarrow \frac{1}{K_{y_l}} \approx \frac{1}{\alpha + K_{y_h}}
\]

The smaller \( \alpha \), the more uniform the optimal value from the most frequent class to the least one is, friendly to low-frequency classes learning.

However, when \( \alpha \) decreases, the intensity of contrast among samples will be weaker, the intensity of contrast between samples and centers will be stronger. The whole loss becomes closer to supervised cross-entropy. To make good use of contrastive learning and rebalance at the same time, we observe that \( \alpha=0.05 \) is a reasonable choice.

**D. PaCo Under Balanced Setting**

For balanced datasets, all classes have the same frequency, i.e., \( q_i = q(y_i) \) and \( K_i = K_{y_i} \) for any class \( y_i \) and class \( y_j \). In this case, PaCo reduces to an improved version of multi-task with weighted sum of supervised cross-entropy and supervised contrastive loss. The connection between PaCo and multi-task is

\[
ExpSum = \sum_{c_k \in C} \exp(c_k \cdot F(x_i)) + \sum_{z_k \in A(i)} \exp(z_k \cdot G(x_i)).
\]

We also write the PaCo loss as

\[
\mathcal{L}_i = \sum_{z_+ \in P(i) \cup \{ c_y \}} -w(z_+) \log \frac{\exp(z_+ \cdot T(x_i))}{\sum_{z_k \in A(i) \cup C} \exp(z_k \cdot T(x_i))},
\]
\[ \mathcal{L}_{extra} = -\log(P_{sup}) - \alpha K^* \log(1 - P_{sup}). \]  

(6)

Here, we take full ImageNet as an example, i.e., \( q^* = 0.001 \), length(queue) = 8192, \( \alpha = 0.05 \), \( \alpha K^* = 0.41 \). Then the function curve for \( \mathcal{L}_{extra} \) is shown in Fig. 3(b). With \( P_{sup} \) increases from 0 to 1, the function value decreases until \( P_{sup} = 0.71 \) and then goes up, which implies \( \mathcal{L}_{extra} \) obtains the smallest loss value when \( P_{sup} = 0.71 \). Note that, when the whole PaCo loss in (4) achieves the optimal solution, \( P_{sup} = 0.71 \) still establishes as demonstrated by Remark 2. With \( P_{sup} \) increases in the training, we analyze how it affects the intensity of supervised contrastive loss and supervised cross-entropy loss in the following.

**Adaptive Weighting Between \( \mathcal{L}_{sup} \) and \( \mathcal{L}_{supcon} \):** Note that the optimal value for the probability that two samples are a true positive pair with label \( y \) is 0.035 as indicated by Remark 2. We suppose \( p_t, p_h \in (0, 0.71) \) and \( p_t < p_h \). To achieve the optimal value, when \( P_{sup} = p_t \), the supervised contrastive loss value \( \mathcal{L}_{supcon} \) must decrease as in (7).

\[
\mathcal{L}_{supcon} = -\sum_{z \in P(i)} \log \frac{\exp(z \cdot \mathbf{G}(x_i))}{\sum_{z_k \in A(i)} \exp(z_k \cdot \mathbf{G}(x_i))} = -\sum_{z \in P(i)} \log \frac{\exp(z \cdot \mathbf{G}(x_i))}{\sum_{z_k \in A(i)} \exp(z_k \cdot \mathbf{G}(x_i))} = -K^* \log \frac{0.035}{1 - p_t} \]  

(7)

Here \( P_{sup} \) increases from \( p_t \) to \( p_h \), \( \mathcal{L}_{supcon} \) must decrease to a much smaller loss value to achieve the optimal solution, which implies the need to make two different class samples much more discriminative, i.e., increasing inter-class margins, and thus the intensity of supervised contrastive loss will enlarge.

An intuition is that as \( P_{sup} \) increases, more samples are pulled together with their corresponding centers. Along with stronger intensity of supervised contrastive loss at that time, it is more likely to push hard examples close to those samples that are already around right centers.

**E. Center Learning Rebalance**

PaCo balances the contrastive learning (for moderating contrast among samples). However the center learning also needs to be balanced, which has been explored in [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [55]. We incorporate Balanced Softmax [27] into the center learning. Then the PaCo loss is changed from (4) to what follows:

\[
\mathcal{L}_i = \sum_{z \in P(i) \cup \{c_k\}} -w(z) \log \frac{\psi(z, T(x_i))}{\sum_{z_k \in A(i) \cup C} \psi(z_k, T(x_i))},
\]

where

\[
\psi(z_k, T(x_i)) = \begin{cases} 
\exp(z_k \cdot \mathbf{F}(x_i)) & z_k \in A(i); \\
\exp(z_k \cdot \mathbf{F}(x_i)) \cdot q(y_k) & z_k \in C.
\end{cases}
\]

We emphasize that Balanced Softmax is only a practical remedy for center learning rebalance. The theoretical analysis remains as a future work.

**F. Generalized Parametric Contrastive Learning**

MoCo [16] and MoCo v2 [17] use a momentum encoder and a queue to allow small mini-batch training and achieve even better performance than SimCLR [15] which requires large mini-batch training, while Chen et al. [34] claims that it is "stop-gradient" rather than momentum encoder that avoid collapsing solutions in self-supervised learning.

We delve deeper to explore how each component in PaCo affects performance by examining the augmentation strategy, formulation of GPaCo/PaCo, the momentum encoder, and queue size in Section IV-E. Interestingly, we observe that the derived Generalized Parametric Contrastive Learning (GPaCo) by removing the momentum encoder boosts performance on imbalanced and balanced data again, indicating that the strong regularization effects from the momentum encoder can be harmful to the model performance in supervised learning.

In addition to long-tailed recognition, GPaCo models enjoy better generalization ability across CNNs (ResNets) to vision transformers (ViT models) on full ImageNet and CIFAR, which is demonstrated in Sections IV-B and IV-C. Besides image classification, GPaCo shows its great generality on other tasks, e.g., out-of-distribution robustness, and semantic segmentation.

To demonstrate the advantages of GPaCo on out-of-distribution data, we load MAE [14] pre-trained weights and then fine-tune on full ImageNet with same training strategy as in [14]. Compared with MAE baselines, GPaCo models achieve much stronger robustness across different benchmarks, which is discussed in Section IV-C.

For semantic segmentation task, GPaCo treats each pixel as an example. However, huge number of pixels make it impractical to directly apply GPaCo to pixel classification. To reduce computational cost and GPU memory, we randomly sample 8,192 pixels from down-sampled pixel features. These selected pixel features go through a 3-layer mlp transform and then are fed into GPaCo, which is adopted as an auxiliary loss for training.
IV. EXPERIMENTS

A. Long-Tailed Recognition

We follow the common evaluation protocol \cite{28, 30, 36} in long-tailed recognition — that is, training models on the long-tailed source label distribution and evaluating their performance on the uniform target label distribution. We conduct experiments on long-tailed version of CIFAR-100 \cite{23, 35}, Places \cite{36}, ImageNet \cite{36} and iNaturalist 2018 \cite{37} datasets.

CIFAR-100-LT Datasets: We use the long-tailed version of CIFAR datasets with the same setting as those used in \cite{23, 35, 52}. They control the degrees of data imbalance with an imbalance factor $\beta$, $\beta = \frac{N_{\text{max}}}{N_{\text{min}}}$ where $N_{\text{max}}$ and $N_{\text{min}}$ are the numbers of training samples for the most and least frequent classes respectively. Following \cite{52}, we conduct experiments with imbalance factors 100, 50, and 10.

ImageNet-LT and Places-LT: ImageNet-LT and Places-LT were proposed in \cite{36}. ImageNet-LT is a long-tailed version of ImageNet dataset \cite{11} by sampling a subset following the Pareto distribution with power value $\alpha=6$. It contains 115.8K images from 1,000 categories, with class cardinality ranging from 5 to 1,280. Places-LT is a long-tailed version of the large-scale scene classification dataset Places \cite{13}. It consists of 184.5K images from 365 categories with class cardinality ranging from 5 to 4,980.

iNaturalist 2018: The iNaturalist 2018 \cite{37} is one species classification dataset, which is on a large scale and suffers from extremely imbalanced label distribution. It is composed of 437.5K images from 8,142 categories. In addition to the extreme imbalance, the iNaturalist 2018 dataset also confronts the fine-grained problem \cite{57}.

Implementation Details: For image classification on ImageNet-LT, we used ResNet-50, ResNetXt-50-32x4d, and ResNetXt-101-32x4d as our backbones for experiments. For iNaturalist 2018, we conduct experiments with ResNet-50 and ResNet-152. All models were trained using SGD optimizer with momentum $\mu = 0.9$.

Contrastive learning benefits from longer training compared with traditional supervised learning with cross-entropy as Chen \cite{15} concluded, which is also validated by previous work of \cite{15, 16, 17, 18, 19}. MoCo \cite{16, 17}, BYOL \cite{18} and SWAV \cite{19} train 800 epochs for model convergence. Supervised contrastive learning \cite{20} trains 350 epochs for feature learning and another 350 epochs for classifier learning.

Following MoCo \cite{16, 17}, when we train models with GPaCo/PaCo, the learning rate decays by a cosine scheduler from 0.04/0.02 to 0 with batch size 128 on 4 GPUs in 400 epochs. The temperature is set to 0.2. $\alpha$ is 0.05. For a fair comparison, we re-implement baselines with the same training time and RandAugment \cite{56} for recent state-of-the-arts of Decouple \cite{28}, Balanced Softmax \cite{27} and RIDE \cite{29}. Especially, for RIDE, based on model ensemble, we compare with it under comparable inference latency in Fig. 1(a).

For Places-LT, following previous setting \cite{30, 36}, we choose ResNet-152 as the backbone network, pre-train it on the full ImageNet-2012 dataset (provided by torchvision), and finely tune it for 30 epochs on Places-LT. The learning rate decays by a cosine scheduler from 0.02 to 0 with batch size 128. The temperature is set to 0.2. $\alpha$ is 0.02/0.05 for GPaCo and PaCo. For CIFAR-100-LT, we strictly follow the setting of \cite{27} for fair comparison. A smaller temperature of 0.07 and $\alpha = 0.01$ are adopted.

Comparison on ImageNet-LT: Table II shows extensive experimental results for comparison with recent SOTA methods. We observe that Balanced Softmax \cite{27} still achieves comparable results with Decouple \cite{28} across various backbones under such strong training setting on ImageNet-LT, consistent with what is claimed in the original paper. For RIDE that is based on model ensemble, we analyze the real inference speed by calculating inference time with a batch of 64 images on Nvidia GeForce 2080Ti GPU.

We observe RIDEResNet with 3 experts even has higher inference latency than a standard ResNetXt-50-32x4d (15.3ms versus 13.1ms); RIDEResNetXt with 3 experts yields higher inference latency than a standard ResNetXt-101-32x4d (26ms versus 25ms). This result is in accordance with the conclusion that network fragmentation reduces the degree of parallelism and thus decreases efficiency in \cite{9, 58}. For fair comparison, we do not apply knowledge distillation tricks for all these methods. As shown in Fig. 1(a) and Table II, under comparable inference latency, GPaCo/PaCo significantly surpasses these baselines.

Comparison on Places-LT: The experimental results on Places-LT are summarized in Table III. Due to the architecture change of RIDE, it is not applicable to load the publicly pre-trained model on full ImageNet, while GPaCo/PaCo is more flexible because the network architecture is the same as those of \cite{30, 36}. Under fair training setting by finely tuning 30 epochs without RandAugment, PaCo surpasses SOTA Balanced Softmax by 2.6%. GPaCo even achieves 41.7% after removing the momentum encoder. An interesting observation is that RandAugment has little effect on the Places-LT dataset. A similar phenomenon can be observed on the iNaturalist 2018 dataset.

| Method     | ResNet-50 | ResNetXt-50 | ResNetXt-101 |
|------------|-----------|-------------|--------------|
| CE(baseline) | 41.6      | 44.4        | 44.8         |
| Decouple-cRT | 47.3      | 49.6        | 49.4         |
| Decouple-$\tau$-norm | 46.7      | 49.4        | 49.6         |
| De-confound-TDE | 51.7      | 51.8        | 53.3         |
| ResLT       | -         | 52.9        | 54.1         |
| MsiLAS      | 52.7      | -           | -            |
| Decouple-$\tau$-norm $^\dagger$ | 54.5      | 56.0        | 57.9         |
| Balanced Softmax $^\dagger$ | 55.0      | 56.2        | 58.0         |
| PaCo        | 57.0      | 58.2        | 60.0         |
| GPaCo$^\dagger$ | 58.5      | 58.9        | 60.8         |

$^\dagger$ Denotes models are trained with random augment \cite{56} in 400 epochs. More comparisons with ride \cite{29} are in Fig. 1.
TABLE III
PERFORMANCE ON PLACES-LT [36], STARTING FROM AN IMAGENET PRE-TRAINED RESNET-152 PROVIDED BY TORCHVISION

| Method        | Many | Medium | Few | All |
|---------------|------|--------|-----|-----|
| CE (baseline) | 45.7 | 27.3   | 8.2 | 30.2|
| OLTR          | 44.7 | 37.0   | 25.3| 35.9|
| Decouple-$\tau$-norm | 37.8 | 40.7   | 31.8| 37.9|
| Balanced Softmax | 42.0 | 39.3   | 30.5| 38.6|
| RealTF        | 39.8 | 43.6   | 31.4| 39.8|
| MiSLAS        | 39.6 | 43.3   | 36.1| 40.4|
| RIDE (2 experts) | -    | -      | -   | -   |

*“+” denotes the model trained with randaugment [56].

More evaluation numbers are in the supplementary file, available online. They can be intuitively understood since RandAugment is designed for ImageNet classification, which inspires us to explore general augmentations across different domains.

**Comparison on iNaturalist 2018**: Table VI lists experimental results on iNaturalist 2018. Under fair training setting, PaCo consistently surpasses recent SOTA methods of Decouple, Balanced Softmax and RIDE. Our method is 1.4% higher than Balanced Softmax. We also apply PaCo on large ResNet-152 architecture. And the performance boosts to 75.3% top-1 accuracy. Surprisingly, With GPaCo, a ResNet-50 model achieves 75.4% top-1 accuracy, implying that the momentum encoder for PaCo can hurt optimization in training. Note that we only transfer the hyper-parameters of PaCo on ImageNet-LT to iNaturalist 2018 without any change. Tuning hyper-parameters for GPaCo/PaCo will bring further improvement.

**Comparison on CIFAR-100-LT**: The experimental results on CIFAR-100-LT are listed in Table VIII. For the CIFAR-100-LT dataset, we mainly compare with the SOTA method Balanced Softmax [27] with the same training setting where Cutout [60] and AutoAugment [61] are used in training. As shown in Table VIII, PaCo consistently outperforms Balanced Softmax across different imbalance factors with such a strong setting. Specifically, PaCo surpasses Balanced Softmax by 1.2%, 1.8% and 1.2% under imbalance factor 100, 50 and 10 respectively, which testify the effectiveness of our PaCo method. GPaCo models show better performance than PaCo models.

**B. Full ImageNet and CIFAR Recognition**

As analyzed in Section III-D, for balanced datasets, PaCo reduces to an improved version of multi-task learning, which adaptively adjusts the intensity of supervised cross-entropy loss and supervised contrastive loss. To verify the effectiveness of PaCo under this balanced setting, we conduct experiments on full ImageNet and full CIFAR. They are indicative to compare GPaCo/PaCo with supervised contrastive learning (Supcon) [20]. Note that, under full ImageNet and CIFAR, we remove the rebalance in center learning, i.e., Balanced Softmax, for fair comparisons.

**Full ImageNet**: In the implementation, we transfer hyper-parameters of GPaCo/PaCo on ImageNet-LT to full ImageNet without modification. SGD optimizer with momentum $\mu = 0.9$ is used. $\alpha = 0.05$, temperature is 0.2 and queue size is 8,192. For multi-task training, the supervised contrastive loss is an additional regularization and the loss weight is extensively explored from 0.1 to 1.0 in Section IV-E. The same data augmentation strategy is applied as in GPaCo/PaCo, which is discussed in Section IV-E.

The experimental results are summarized in Table V. With SimAugment, our ResNet-50 PaCo model achieves 78.7% top-1 accuracy, which outperforms supervised contrastive learning model by 0.8%. Equipped with strong augmentation, i.e., RandAugment [56], the performance further improves to 79.3%. ResNet-101/200 trained with PaCo consistently surpass supervised contrastive learning. Replacing PaCo with GPaCo, The ResNet-50 model boosts to 79.7% top-1 accuracy.

**Full CIFAR-100**: For CIFAR implementation, we follow supervised contrastive learning and train ResNet-50 with only the SimAugment. Compared with full ImageNet, we adopt a smaller temperature of 0.07, $\alpha = 0.01$ and batch size 256 with learning rate 0.1. As shown in Table IX, on CIFAR-100, PaCo outperforms supervised contrastive learning by 2.6%, which validates the advantages of PaCo. Again, the GPaCo model achieves 80.3% top-1 accuracy, significantly surpassing supervised contrastive learning by 3.8%. Note that, following [9], we use a weight-decay of 5e-4.

**C. Out-of-Distribution Robustness**

He et al. [14] justifies that masked autoencoders are scalable vision learners. With such a simple pre-training strategy that masking random patches of the input image and reconstructing the missing pixels, the good representation is learned and well transferred to downstream tasks, i.e., full ImageNet classification, object detection, semantic segmentation, and robustness on out-of-distribution data. In this section, we verify that GPaCo can enhance model generalization ability and robustness on full ImageNet and its variants when compared with MAE [14] models.

**Robustness Evaluation**: We extensively evaluate the model performance on out-of-distribution robustness using following benchmarks: 1) ImageNet-C [62], with various common image corruptions; 2) ImageNet-R [63], which contains natural renditions of ImageNet object classes with different textures and local image statistics; 3) ImageNet-Sketch [64], which includes sketch images of the same ImageNet classes collected online.

**Comparison With MAE Models [14]**: We summarize experimental results in Table IV. Top-1 accuracy is reported on full ImageNet, ImageNet-R, and ImageNet-S. Mean Corruption Error (mCE) and Relative Mean Corruption Error (rel. mCE) [62] are used for ImageNet-C. mCE is to measure absolute robustness to corruptions while rel. mCE is a better metric when we compare models with different top-1 accuracy.

Compared with MAE models, we achieve better performance on full ImageNet, surpassing them by 0.4 and 0.3 respectively for ViT-B [59] and ViT-L [59]. On ImageNet-C, ImageNet-R, and ImageNet-S, GPaCo models usually outperforms MAE models.
TABLE IV
EVALUATION WITH ViT MODELS [59] ON FULL IMAGE NET AND OUT-OF-DISTRIBUTION DATA. GPaCo MODELS ENJOY BETTER GeneralIZATION ABILITY AND STRONGER ROBUSTNESS COMPARED WITH MAE MODELS. WE USE THEIR OFFICIAL OPEN-SOURCE CODE FOR REPRODUCING RESULTS OF MAE MODELS WITH 8 Nvidia GeForce RTX3090 GPUs. TOP-1 ACCURACY ARE REPORTED FOR FULL IMAGE NET, IMAGE NET-R, AND IMAGE NET-S. mCE AND rel. mCE ARE REPORTED FOR IMAGE NET-C.

| Method | Model | Full ImageNet (**) | ImageNet-C (mCE ↓) | ImageNet-C (rel. mCE ↓) | ImageNet-R (**) | ImageNet-S (**) |
|--------|-------|---------------------|---------------------|-------------------------|-----------------|----------------|
|        |       | w / Mixup and CutMix|                     |                         |                 |                |
| MAE    | ViT-L | 85.7                | 32.4                | 41.4                    | 60.3            | 45.5           |
| MAE    | ViT-B | 83.6                | 39.1                | 49.9                    | 49.9            | 36.1           |
|        |       | w/o Mixup and CutMix|                     |                         |                 |                |
| SupCon | ResNet-50 | 78.7               | 67.2                | 94.6                    | 64.4            | 41.1           |
| GPaCo  | ResNet-50 | 79.7               | 50.9                |                         |                 |                |
| MAE    | ViT-L | 85.2                | 36.4                | 46.3                    | 55.3            | 42.6           |
| MAE    | ViT-B | 82.5                | 44.9                | 56.9                    | 44.9            | 32.9           |
| GPaCo  | ViT-L | 84.0                | 37.2                | 47.3                    | 51.7            | 39.4           |
| GPaCo  | ViT-B | 86.0                | 30.7                | 39.0                    | 60.3            | 48.3           |

D. Semantic Segmentation
In this section, we transfer GPaCo to the downstream task, i.e., semantic segmentation. In implementation, we use GPaCo as an auxiliary loss. Specifically, for each image, we randomly

by 1.7 ~ 3.3, demonstrating much stronger robustness on out-of-distribution data.
sample 8,192 pixel features on the down-sampled feature map and feed them into GPaCo loss for training optimization.

Datasets: ADE20K [39] contains 22K densely annotated images with 150 fine-grained semantic concepts. The training and validation sets consist of 20K and 2K images, respectively. COCO-Stuff [40] includes 10K images from the COCO training set. The training and validation sets consist of 9K and 1K images, respectively. It covers 171 classes. For PASCAL Context [41], the subset of 59 frequent classes is used following previous work. Cityscapes [42] consists of 19 classes, covering street scenes from 50 different cities.

Training and Evaluation: We implement our GPaCo in mmseg codebase [65]. On ADE20K [39], we follow [66] to experiment with Swin transformers and UperNet [67]. On COCO-Stuff [40], Pascal Context [41] and Cityscapes [42], we use ResNet-50/101 and deeplabv3+. Default training hyper-parameters in baselines are adopted in the GPaCo model training phase, i.e., the standard random scale jittering between 0.5 and 2.0, random horizontal flipping, random cropping, as well as random color jittering. We report both single scale (s.s.) and multi-scale (m.s.) evaluation results on validation data. For multi-scale inference, scales of 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 are used.

Results: Experimental results are summarized in Table VII. On ADE20K, obvious improvements are observed from Swin-T to Swin-L models when equipped with GPaCo, outperforming baselines by 1.0, 1.5, 0.8 mIoU respectively. On COCO-Stuff and PASCAL Context, GPaCo models significantly surpass baselines by around 1.1 ~ 1.7 mIoU.

E. Ablation Study

Data Augmentation Strategy for PaCo & GPaCo: Data augmentation is the key for success of contrastive learning as indicated by Chen [15]. For PaCo & GPaCo, we also conduct ablation studies for different augmentation strategies. Several observations are intriguingly different from those of [68]. We experiment with the following ways of data augmentation.

- SimAug: an augmentation policy [16], [17] that applies random flips and color jitters followed by Gaussian blur.
- RandAug [56]: A two stage augmentation policy that uses random parameters in place of parameters tuned by AutoAugment. The random parameters do not need to be tuned and hence reduces the search space.
- RandAugStack: RandAug followed by Gaussian blur.

For the common random resized crop used along with the above three strategies, work of [68] explains that the optimal hyper-parameter for random resized crop is (0.2, 1) in self-supervised contrastive learning. This setting is also adopted by other work of [15], [16], [17], [18], [19]. However, in this paper, we observe severe performance degradation on ImageNet-LT with ResNet-50 (55.0% versus 52.2%) for PaCo when we change the hyper-parameter from (0.08, 1) to (0.2, 1). This is because PaCo involves center learning while other self-supervised frameworks only apply non-parametric contrastive loss as described in Section II. Note that the same phenomenon is also observed on traditional supervised learning with cross-entropy loss.

Another observation is that GPaCo can make better use of strong augmentations compared with PaCo. The work of [69] demonstrates that directly applying strong data augmentation in MoCo [16], [17] does not work well. Here we observe a similar conclusion with RandAug [56] for PaCo. However, after removing the momentum encoder, GPaCo can achieve better performance with stronger augmentation as shown in Table X.

Multi-Task Learning: Re-weighting is a classical method for dealing with imbalanced data. Here we directly apply the re-weighting method of Cui [23] in contrastive learning to compare with PaCo. Moreover, Balanced softmax (BalSfx) [27], as one state-of-the-art method for traditional cross-entropy in long-tailed recognition, is also applied to contrastive learning re-balancing. The experimental results are summarized in Table XI. It is obvious PaCo significantly surpasses the two baselines.

PaCo/GPaCo balances the contrastive learning (for moderating contrast among samples). However the center learning also needs to be balanced, which has been explored in [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [55]. To compare with state-of-the-art methods in long-tailed recognition, we incorporate Balanced Softmax (BalSfx) [27] into the center learning. As shown in Table XII, after re-balance in center learning, GPaCo boosts performance to 58.9%, surpassing baselines.

We also verify the advantage of GPaCo over multi-task learning on balanced data — full ImageNet and full CIFAR-100. As

| Method      | CIFAR-100 Top-1 Acc | CIFAR-100 Top-1 Acc |
|-------------|---------------------|---------------------|
| CE (baseline) | 77.9                | 78.0                |
| multi-task   | 76.5                | 79.1                |
| Supcon       | 78.0                | 80.3                |

| Method | View-1 | View-2 | Top-1 Acc |
|--------|--------|--------|-----------|
| PaCo   | SimAug | SimAug | 55.0      |
| PaCo   | RandAug| SimAug | 57.0      |
| PaCo   | RandAug| RandAug| 56.5      |
| GPaCo  | RandAug| SimAug | 57.9      |
| GPaCo  | RandAug| RandAugStack| 58.5 |

| Method | Top-1 Acc |
|--------|-----------|
| CE     | 48.4      |
| multi-task (CE+Re-weighting) | 49.0 |
| multi-task (CE+BalSfx) | 48.6 |
| PaCo   | 51.0      |
shown in Table XIII, extensive values for the weight between cross-entropy and supervised contrastive learning have been explored. GPaCo obviously achieve much higher performance.

**GPaCo With Queue Length:** The queue is designed to enlarge the number of samples for GPaCo/PaCo loss. We examine the effects of different queue size on model performance. Empirical study on ImageNet-LT with ResNeXt-50 conducted. As shown in Fig. 4, the larger queue size usually leads to better performance. However, when the queue size increase from 4,096 to 8,192, the performance gains become much smaller.

---

### TABLE XII
**Comparison With Multi-Task Re-Weighting Baselines That Perform Center Learning Rebalance on ImageNet-LT**

| Method | Backbone | Weight | Top 1 Acc |
|--------|----------|--------|-----------|
| multi-task (Balsf + rw) | ResNeXt-50 | 0.05 | 57.0 |
| multi-task (Balsf + rw) | ResNeXt-50 | 0.10 | 57.1 |
| multi-task (Balsf + rw) | ResNeXt-50 | 0.20 | 57.1 |
| multi-task (Balsf + rw) | ResNeXt-50 | 0.30 | 57.0 |
| multi-task (Balsf + rw) | ResNeXt-50 | 0.50 | 57.2 |
| multi-task (Balsf + rw) | ResNeXt-50 | 0.80 | 57.2 |
| multi-task (Balsf + rw) | ResNeXt-50 | 1.00 | 56.9 |
| PaCo | ResNeXt-50 | - | 58.2 |

| Method | Backbone | Weight | Top 1 Acc |
|--------|----------|--------|-----------|
| multi-task* (Balsf + rw) | ResNeXt-50 | 0.50 | 58.3 |
| GPaCo | ResNeXt-50 | 0.80 | 58.2 |
| multi-task* (Balsf + rw) | ResNeXt-50 | 0.50 | 57.8 |
| GPaCo | ResNeXt-50 | - | 58.9 |
| multi-task* (Balsf + rw) | ResNeXt-101 | 0.50 | 60.0 |
| GPaCo | ResNeXt-101 | - | 60.8 |

Models are all trained with randomaugment in 400 epochs.

### TABLE XIII
**Comparison With Multi-Task Baselines on Balanced Data**

| Method | Backbone | Weight | Top 1 Acc |
|--------|----------|--------|-----------|
| multi-task (CE + Supcon) | ResNet-50 | 0.50 | 78.0 |
| PaCo | ResNet-50 | - | 79.1 |
| multi-task* (CE + Supcon) | ResNet-50 | 0.10 | 79.0 |
| multi-task* (CE + Supcon) | ResNet-50 | 0.30 | 79.1 |
| multi-task* (CE + Supcon) | ResNet-50 | 0.50 | 79.1 |
| multi-task* (CE + Supcon) | ResNet-50 | 0.80 | 78.9 |
| GPaCo | ResNet-50 | 1.00 | 79.0 |

**Full CIFAR-100 Dataset**

| Method | Backbone | Weight | Top 1 Acc |
|--------|----------|--------|-----------|
| multi-task (CE + Supcon) | ResNet-50 | 0.50 | 78.0 |
| GPaCo | ResNet-50 | - | 80.3 |

**Full ImageNet Dataset**

| Method | Backbone | Weight | Top 1 Acc |
|--------|----------|--------|-----------|
| multi-task (CE + Supcon) | ViT-B | 0.50 | 83.4 |
| GPaCo | ViT-B | - | 84.0 |

**Fig. 4.** The larger queue size, the better performance. Experimental results with ResNeXt-50 on ImageNet-LT are plotted.

---

### TABLE XIV
**Ablation for the Necessary of Two-View Training on Full ImageNet With ViT-B**

| Method | Full ImageNet(×) | ImageNet-C (rel. mCE ↓) |
|--------|-----------------|------------------------|
| single-view | 82.5 | 49.1 |
| two-view | 84.0 | 47.3 |

### TABLE XV
**Ablation for Model Ensemble on ImageNet-LT**

| Method | Single Model | Few |
|--------|--------------|-----|
| ResLT*(ResNeXt-50) | 68.7 | 59.5 |
| RIDE (ResNeXt-50) | 67.3 | 57.1 |
| GPaCo (ResNeXt-50) | 67.4 | 57.1 |
| GPaCo (ResNeXt-101) | 68.7 | 59.5 |

**Model Ensemble**

| Method | Single Model | Few |
|--------|--------------|-----|
| ResLT* | 64.0 | 56.6 |
| RIDE | 71.8 | 53.9 |
| GPaCo (ResNeXt-50 2-experts) | 69.9 | 59.5 |
| GPaCo + (ResNeXt-101 2-experts) | 70.9 | 62.0 |

**“*” Represents the ensemble model of a ResNeXt-50 and a ResNeXt-101. “**” Represents the results are from their original paper [70].

---

**V. CONCLUSION**

In this paper, we have proposed the Generalized Parametric Contrastive Learning (GPaCo/PaCo), which can deal with both imbalanced and balanced data well. It is based on the theoretical analysis of supervised contrastive learning and rebalance from the convergent optimal values. On balanced data, our analysis of PaCo demonstrates that it can adaptively enhance the intensity of pushing two samples of the same class close as more samples are pulled together with their corresponding centers, which can potentially benefit hard examples learning in training.

We conduct experiments on various benchmarks of CIFAR-LT, ImageNet-LT, Places-LT, and iNaturalist 2018. The experimental results show that we create a new state-of-the-art for long-tailed recognition. With balanced data, experimental necessary of two views in GPaCo: In training of GPaCo, a positive pair consists of two images belonging to the same class. We explore whether it is necessary to generate two views as in the self-supervised learning [15], [16], [17]. As shown in Table XIV, performance and robustness significantly drops with single view training, which implies the vast importance of two views training in GPaCo.

**Model Ensemble:** Models trained on imbalanced data can easily suffer from over-fitting issue on low-frequency classes, thus leading to high variance of model predictions. RIDE proposes an improved version of model ensemble strategy and has verified the advantages of model ensemble for long-tailed recognition. However, GPaCo/PaCo improves single model performance with better feature representation. We go deeper to explore the model ensemble strategy on GPaCo models. As shown in Table XV, the improvements coming from model ensemble is much smaller for GPaCo models when compared with RIDE and ResLT (improvements of ResLT versus RIDE versus GPaCo are 4.7% versus 4.6% versus 2.4%), demonstrating that GPaCo training can reduce prediction variance of trained models.

---

**Necessary of Two Views in GPaCo:** In training of GPaCo, a positive pair consists of two images belonging to the same class. We explore whether it is necessary to generate two views as in the self-supervised learning [15], [16], [17]. As shown in Table XIV, performance and robustness significantly drops with single view training, which implies the vast importance of two views training in GPaCo.

---

**Model Ensemble:** Models trained on imbalanced data can easily suffer from over-fitting issue on low-frequency classes, thus leading to high variance of model predictions. RIDE proposes an improved version of model ensemble strategy and has verified the advantages of model ensemble for long-tailed recognition. However, GPaCo/PaCo improves single model performance with better feature representation. We go deeper to explore the model ensemble strategy on GPaCo models. As shown in Table XV, the improvements coming from model ensemble is much smaller for GPaCo models when compared with RIDE and ResLT (improvements of ResLT versus RIDE versus GPaCo are 4.7% versus 4.6% versus 2.4%), demonstrating that GPaCo training can reduce prediction variance of trained models.

---

In this paper, we have proposed the Generalized Parametric Contrastive Learning (GPaCo/PaCo), which can deal with both imbalanced and balanced data well. It is based on the theoretical analysis of supervised contrastive learning and rebalance from the convergent optimal values. On balanced data, our analysis of PaCo demonstrates that it can adaptively enhance the intensity of pushing two samples of the same class close as more samples are pulled together with their corresponding centers, which can potentially benefit hard examples learning in training.

We conduct experiments on various benchmarks of CIFAR-LT, ImageNet-LT, Places-LT, and iNaturalist 2018. The experimental results show that we create a new state-of-the-art for long-tailed recognition. With balanced data, experimental...
results on full ImageNet and CIFAR show that GPaCo models have better generalization ability and stronger robustness. Transferring GPaCo to semantic segmentation, obvious improvements are obtained.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[2] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Representations, 2015.

[3] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 936–944.

[4] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8759–8768.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6230–6239.

[6] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning transferable architectures for scalable image recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8697–8710.

[7] H. Liu, K. Simonyan, and Y. Yang, “DARTS: Differentiable architecture search,” in Proc. Int. Conf. Learn. Representations, 2019.

[8] M. Tan et al., “MnasNet: Platform-aware neural architecture search for mobile,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2815–2823.

[9] J. Cui, P. Chen, R. Li, S. Liu, X. Shen, and J. Jia, “Fast and practical neural architecture search,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6508–6517.

[10] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, “Once-for-all: Train one network and specialize it for efficient deployment,” in Proc. Int. Conf. Learn. Representations, 2020.

[11] O. V. Akesovskiy, “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 2, pp. 211–252, 2015.

[12] T. Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[13] 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.

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 16 000–16 009.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. 37th Int. Conf. Mach. Learn., 2020, pp. 1597–1607.

[16] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 9726–9735.

[17] X. Chen, H. Fan, R. B. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” 2020, arXiv: 2003.04297.

[18] J. Grill et al., “Bootstrap your own latent - A new approach to self-supervised learning,” in Proc. 34th Int. Conf. Neural Inf. Process. Syst., H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020, Art. no. 1706.

[19] M. Caron, I. Misra, J. Mairal, P. Goyal, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” in Proc. 34th Int. Conf. Neural Inf. Process. Syst., H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020, Art. no. 851.

[20] P. Khosla et al., “Supervised contrastive learning,” in Proc. 34th Int. Conf. Neural Inf. Process. Syst., H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020, Art. no. 1567.

[21] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” Neural Netw., vol. 106, pp. 249–259, 2018.

[22] C. Huang, Y. Li, C. C. Loy, and X. Tang, “Learning deep representation for imbalanced classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 5375–5384.

[23] Y. Cui, M. Jia, T. Lin, Y. Song, and S. J. Belongie, “Class-balanced loss based on effective number of samples,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 9260–9269.

[24] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[52] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, “BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition,” 2019, arXiv: 1912.02413.

[53] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2018, arXiv: 1807.03748.

[54] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3733–3742.

[55] R. Duggal et al., “ELF: An early-exiting framework for long-tailed classification,” 2020, arXiv: 2006.11979.

[56] 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, 2020, pp. 702–703.

[57] X.-S. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, “Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 6116–6125, Dec. 2019.

[58] N. Ma, X. Zhang, H. Zheng, and J. Sun, “ShuffleNet V2: Practical guidelines for efficient CNN architecture design,” in Proc. Eur. Conf. Comput. Vis. V., Ferrarri, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., 2018, pp. 122–138.

[59] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, arXiv: 1011.1928.

[60] T. Devries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” 2017, arXiv: 1708.04552.

[61] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, “AutoAugment: Learning augmentation strategies from data,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 113–123.

[62] D. Hendrycks and T. G. Dettlerich, “Benchmarking neural network robustness to common corruptions and surface variations,” 2018, arXiv: 1807.01697.

[63] D. Hendrycks et al., “The many faces of robustness: A critical analysis of out-of-distribution generalization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 8340–8349.

[64] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, “Learning robust global representations by penalizing local predictive power,” in Proc. Adv. Neural Inf. Process. Syst., 2019, Art. no. 943.

[65] M. Contributors, “MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark,” 2020, [Online]. Available: https://github.com/open-mmlab/mms segmentation.

[66] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 10 012–10 022.

[67] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 418–434.

[68] W. Ouyang, X. Wang, C. Zhang, and X. Yang, “Factors in finetuning deep model for object detection with long-tail distribution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 864–873.

[69] X. Wang and G-J. Qi, “Contrastive learning with stronger augmentations,” 2021, [Online]. Available: https://openreview.net/forum?id=KJSC_Asn14.

[70] 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.

Zhisheng Zhong (Student Member, IEEE) received the BEng degree in communication engineering from the Beijing University of Posts and Telecommunications (BUPT), in 2016, and the master's degree in computer science from Peking University (PKU), in 2019. Now he is working toward the PhD degree with the Department of Computer Science Engineering (CSE), The Chinese University of Hong Kong (CUHK). He serves as a reviewer for NeurIPS, CVPR, ICCV, ICLR and etc. His research interests lie in deep learning and computer vision.

Zhuotao Tian (Student Member, IEEE) received the BEng degree (Honors) from the Harbin Institute of Technology (HIT), in 2018, and the PhD degree from the Chinese University of Hong Kong (CUHK), in 2022, under the supervision of Prof. Jiaya Jia and Prof. Bei Yu. He is currently a researcher with SmartMore. He serves as a reviewer for top venues including IEEE Transactions on Pattern Analysis and Machine Intelligence, CVPR, ICCV, ECCV, ICMIL, ICLR, NeurIPS, AAAI, International Journal of Computer Vision and IEEE Transactions on Image Processing. His research interests include broad topics related to general 2D-3D semantic segmentation, few-shot learning, and semi-supervised learning.

Shu Liu (Member, IEEE) received the BS degree from the Huazhong University of Science and Technology, and the PhD degree from the Chinese University of Hong Kong. Now serves as co-founder and technical head in SmartMore. He was the winner of 2017 COCO Instance Segmentation Competition and received the Outstanding Reviewer of ICCV, in 2019. He continuously served as a reviewer for IEEE Transactions on Pattern Analysis and Machine Intelligence, CVPR, ICCV, NeurIPS, ICLR and etc. His research interests lie in deep learning and computer vision.

Bei Yu (Member, IEEE) received the PhD degree from The University of Texas at Austin, in 2014. He is currently an Associate Professor with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. He has served as TPC chair of ACM/IEEE Workshop on Machine Learning for CAD, and in many journal editorial boards and conference committees. He is editor of IEEE TCCPS Newsletter. He received nine Best Paper Awards from DATE 2022, ICCAD 2021 & 2013, ASPDAC 2021 & 2012, ICTAI 2019, Integration, the VLSI Journal, in 2018, ISPD 2017, SPIE Advanced Lithography Conference 2016, and six ICCAD/ISPD contest awards.

Jiequan Cui (Student Member, IEEE) received the BEng degree from Computer Science Department, ShanDong University (SDU), in 2018, and the PhD degree from The Chinese University of Hong Kong (CUHK), in 2022. He is currently a research fellow with the School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU). He serves as a reviewer for IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, IEEE Transactions on Image Processing, CVPR, ICCV, ECCV, ICLR, NeurIPS, and ICML. His research interests include model generalization and robustness, autonomous learning, adversarial robustness, neural architecture search, and image segmentation.