CUDA: Curriculum of Data Augmentation for Long-tailed Recognition

Anonymous Author(s)
Affiliation
Address
email

Abstract

Class imbalance problems frequently occur in real-world tasks, and conventional deep learning algorithms are well known for performance degradation on imbalanced training datasets. To mitigate this problem, many approaches have aimed to balance among given classes by *re-weighting* or *re-sampling* training samples. These re-balancing methods increase the impact of minority classes and reduce the influence of majority classes on the output of models. Despite extensive recent studies, no deep analysis has been conducted on determination of classes to be augmented and strength of augmentation has been conducted. In this study, we propose a simple and efficient novel curriculum, which is designed to find the appropriate per-class strength of data augmentation, called CUDA: CUrriculum of Data Augmentation for long-tailed recognition. CUDA can simply be integrated into existing long-tailed recognition methods. We persent the results of experiments showing that CUDA effectively achieves better generalization performance compared to the state-of-the-art method on imbalanced datasets such as CIFAR-100-LT.

1 Introduction

Deep neural networks (DNNs) have significantly improved over the past few decades on a wide range of tasks [21, 45, 44]. This effective performance is made possible by come from well-organized datasets such as MNIST [30], CIFAR-10/100 [27], and ImageNet [48]. However, as Van Horn et al. [51] indicated, gathering such balanced datasets is notoriously difficult in real-world applications. In addition, the models perform poorly when trained on an improperly organized dataset, *e.g.,* in cases with class imbalance, because minority samples can be ignored due to their small portion.

The simplest solution to the class imbalance problem is to prevent the model from ignoring minority classes. To improve generalization performance, many studies have aimed to emphasize minority classes or reduce the influence of the majority samples. Reweighting [6, 40] or resampling [4, 52] are two representative methods that have been frequently applied to achieve this goal.

Although these elaborate rebalancing approaches have been adopted in some applications, limited information on minority classes due to fewer samples remains problematic. To this end, we first consider that controlling the strength of class-wise augmentation can provide another dimension to mitigate the class imbalance problem. In this paper, we use the number of augmentation operations and their magnitude to control the extent of the augmentation, which we refer to herein as its strength, *e.g.,* a strength parameter of 2 means that two randomly sampled operations with a pre-defined magnitude index of 2 are used.

Our key finding is that class-wise augmentation improves performance in the non-augmented classes while that for the augmented classes may not be significantly improved, and in some cases, performances may even decrease. As described in Figure 1, regardless of whether a given dataset is

Submitted to the Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

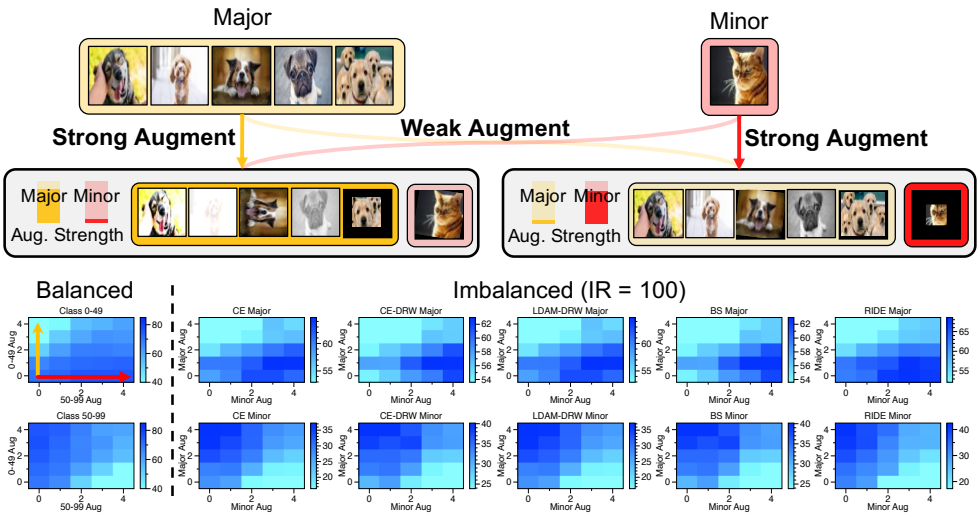


Figure 1: Motivation of CUDA. If one half of the classes (*e.g.,* class indices in 0 to 49) are strongly augmented, the performance of the other half of the classes (*i.e.,* in 50 to 99) increases. This phenomenon is also observed in the imbalanced case in which many more instances are present for the top first half (major) have more samples than the second half classes (minor). Setup and further analysis are described in Appendix B.

class imbalanced, conventional class imbalance methods show similar trends: when only the major classes are strongly augmented (*e.g.,* strength 4), the performance of majority classes decreases, whereas that for the minority classes have better results. This result motivates us to find the proper augmentation strength for each class to improve the performance for other classes while maintaining its own performance.

Contribution. We propose a simple algorithm called CUrriculum of Data Augmentation (CUDA) to find the proper class-wise augmentation strength for long-tailed recognition. The proposed method consists of two modules, which compute a level-of-learning score for each class and leverage the score to determine the augmentation. Therefore, CUDA increases and decreases the augmentation strength of the class that was successfully and wrongly predicted by the trained model.

We empirically examine performance of CUDA on synthetically imbalanced datasets, CIFAR-100-LT [6]. With the high compatibility of CUDA, we apply our framework to various long-tailed recognition methods and achieve better performance compared to the existing long-tailed recognition methods. We describe the further experimental results on ImagNet-LT [39] and iNaturalist 2018 [51] in Appendix D.

2 CUrriculum of Data Augmentation for Long-Tailed Recognition

The core philosophy of CUDA is to *"generate an augmented sample that becomes the most difficult sample without losing its original information."* In this section, we describe design of CUDA in terms of two parts: (1) a method to generate the augmented samples based on the given strength parameter, and (2) a method to measure a Level-of-Learning (LoL) score for each class.

2.1 Problem Formulation of Long-tailed Recognition

Suppose that the training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ is composed of images with size d, $x_i \in \mathbb{R}^d$, and their corresponding labels $y_i \in \{1, ..., C\}$. $\mathcal{D}_c \subset \mathcal{D}$ is a set of class c, *i.e.,* $\mathcal{D}_c = \{(x, y) | y = c, (x, y) \in \mathcal{D}\}$. Without loss of generality, we assume $|\mathcal{D}_1| \geq |\mathcal{D}_2| \geq \cdots \geq |\mathcal{D}_C|$, where $|\mathcal{D}|$ denotes the cardinality of the set $\mathcal{D}$. We denote the $N_{\text{max}} := |\mathcal{D}_1|$ and $N_{\text{min}} := |\mathcal{D}_C|$. LTR algorithms, $\mathcal{A}_{\text{LTR}}(f_\theta, \mathcal{D})$, mainly focus on training the model f_θ with parameter θ when the class distribution of training dataset $\mathcal{P}_{\text{train}}(y)$ and test dataset $\mathcal{P}_{\text{test}}(y)$ are not identical. More precisely, $\mathcal{P}_{\text{train}}(y)$ is highly imbalanced while $\mathcal{P}_{\text{test}}(y)$ is balanced, *i.e.,* uniform distribution.

Algorithm 1: **CU**rriculum of **D**ata **A**ugmentation

Input: LTR algorithm $\mathcal{A}_{\text{LTR}}(f, \mathcal{D})$, training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, train epochs E, aug. probability p_{aug}, threshold γ, number of sample coefficient T.
Output: trained model f_θ
Initialize: $L_c^0 = 0 \; \forall c \in \{1, ..., C\}$
for $e \leq E$ **do**
  Update $L_c^e = V_{\text{LoL}}(\mathcal{D}_c, L_c^{e-1}, f_\theta, \gamma, T) \quad \forall c$ `// Alg. 2`
  Generate $\mathcal{D}_{\text{CUDA}} = \{(\bar{x}_i, y_i) | (x_i, y_i) \in \mathcal{D}\}$ where

$$\bar{x}_i = \begin{cases} \mathcal{O}(x_i, L_{y_i}^e) & \text{with prob. } p_{\text{aug}} \\ x_i & \text{otherwise.} \end{cases}$$

  Run LTR algorithm using $\mathcal{D}_{\text{CUDA}}$, *i.e.,* $\mathcal{A}_{\text{LTR}}(f_\theta, \mathcal{D}_{\text{CUDA}})$.
end

Algorithm 2: V_{LoL}: Update LoL score

Input: $\mathcal{D}_c$, L, f_θ, γ, T
Output: updated L
Initialize: check $= 1$
for $l \leq L$ **do**
  /* $V_{\text{correct}}(\mathcal{D}_c, l, f_\theta, T)$ */
  Sample $\mathcal{D}'_c \subset \mathcal{D}_c$ s.t. $|\mathcal{D}'_c| = T(l+1)$
  Compute $v = \sum_{x \in \mathcal{D}'_c} \mathbb{1}_{\{f(\mathcal{O}(x;l)=c\}}$
  if $v \leq \gamma T(l+1)$ **then**
    check $\leftarrow 0$; **break**
  end
end
if check $= 1$ **then** $L \leftarrow L + 1$
else $L \leftarrow L - 1$

2.2 Curriculum of Data Augmentation

In this section, we describe our proposed DA with strength parameter, and the methods used to measured the LoL score. Then, we integrate the two methods in a single framework to propose `CUDA`.

DA with a strength parameter. Let us assume that there exist pre-defined K augmentation operations. We utilize visual augmentation operations which is indiced as $k \in \{1, \cdots, K\}$, *e.g.,* Gaussian blur, Rotation, Horizontal flip. Each augmentation operation $\mathcal{O}_k^{m_k(s)} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ has its own pre-defined augmentation magnitude function $m_k(s)$ where the strength parameter $s \in \{0, ..., S\}$. These operations are described in detail along with each magnitude functions in Appendix E.

Given an augmentation strength parameter s and an input image x, we model a sequence of augmentation operations $\mathcal{O}(x; s)$ as follows:

$$\mathcal{O}(x; s) = \mathcal{O}_{k_s}^{m_{k_s}(s)} \circ \mathcal{O}_{k_{s-1}}^{m_{k_{s-1}}(s)} \circ \cdots \circ \mathcal{O}_{k_1}^{m_{k_1}(s)}(x), \quad k_i \sim \text{Cat}(K, \mathcal{U}(K)) \quad \forall i = \{1, \ldots, s\},$$

where, $\text{Cat}(\cdot)$ and $\mathcal{U}(\cdot)$ denote categorical and discrete uniform distributions, respectively. The sequential augmentation operation $\mathcal{O}(x; s)$ samples s operations from the categorical distribution when the probability of seeing the operations follows uniform distribution.

Level-of-Learning (LoL). To control the strength of augmentation properly, we check whether the model can correctly predict augmented versions without losing the original information. To enable this, we define the LoL for each class c at epoch e, *i.e.,* L_c^e, which is adaptively updated as the training continues as follows:

$$L_c^e = V_{\text{LoL}}(\mathcal{D}_c, L_c^{e-1}, f_\theta, \gamma, T),$$

where

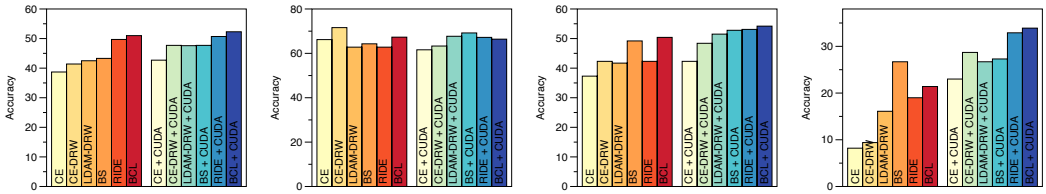
$$V_{\text{LoL}}(\mathcal{D}_c, L_c^{e-1}, f_\theta, \gamma, T) = \begin{cases} L_c^{e-1} + 1 & \text{if } V_{\text{Correct}}(\mathcal{D}_c, l, f_\theta, T) \geq \gamma T(l+1) \quad \forall l \in \{0, ..., L_c^{e-1}\} \\ L_c^{e-1} - 1 & \text{otherwise} \end{cases}.$$

Here, $\gamma \in [0, 1]$ is threshold hyperparameter, T is coefficient of the number of samples used to updating LoL. V_{correct} is a function which outputs the number of correctly predicted examples by the model f_θ among $l + 1$ randomly augmented samples with strength l. V_{correct} is defined as:

$$V_{\text{Correct}}(\mathcal{D}_c, l, f_\theta, T) = \textstyle\sum_{x \in \mathcal{D}'_c} \mathbb{1}_{\{f_\theta(\mathcal{O}(x;l))=c\}} \quad \text{where } \mathcal{D}'_c \subset \mathcal{D}_c.$$

Note that $\mathcal{D}'_c$ is a randomly sampled subset of $\mathcal{D}_c$ with replacement and its size is $T(l + 1)$.

The key philosophy of this criterion is two fold. (1) If samples in the class c are trained sufficiently with an augmentation strength of L_c^e, the model is ready to learn a more difficult version with augmentation strength of $L_c^{e+1} \leftarrow L_c^e + 1$. In contrast, if the model predicts incorrectly, it should re-learn the easier sample with an augmentation strength of $L_c^{e+1} \leftarrow L_c^e - 1$. (2) As the strength parameter increases, the number of candidates for the sequential augmentation operation $\mathcal{O}(x; L)$ increases exponentially. For example, the amount of increment is $N^L(N - 1)$ when L is increases to $L + 1$. To control the LoL in a large sequential augmentation operation space, we take more random samples to check as the strength parameter gets bigger. In our experiments, linearly increasing the number of samples to evaluate corresponding to the strength with a small additional computation time was sufficient. V_{LoL} is described in Algorithm 2.



(a) All (b) Many (c) Median (d) Few

Figure 2: Validation accuracy on CIFAR-100-LT dataset. This experiments conducted on ResNet-32 architecture on CIFAR-100-LT with imbalance ratio 100.

Curriculum of DA. By combining two components, including DA with a strength parameter and LoL, our CUDA provides class-wise adaptive augmentation to enhance the performance of the others without losing its own information. As shown in algorithm 1, we measure the LoL score L_c for all classes in the training dataset to determine the augmentation strength for every epoch. Based on L_c, we generate the augmented version $\mathcal{O}(x; L_c)$ for $x \in \mathcal{D}_c$ and train the model with the augmented samples. Additionally, we randomly use the original sample instead of the augmented sample with probability p_{aug} so that the trained models do not forget the original information. In our experiments, this operation improved performance robustly on a wide range of p_{aug} values.

Advantage of CUDA design. Our proposed approach mainly has three advantages. (1) CUDA adaptively finds proper augmentation strengths for each class without need for a validation set. (2) Following the spirits of existing curriculum learning methods [19, 62, 55], CUDA enables modeling by first presenting easier examples earlier during training to improve generalization. This encourages the model to learn difficult samples (*i.e.,* within high augmentation strength) better. (3) Moreover, owing to the universality of data augmentation, CUDA is easily compatible with other LTR algorithms, such as [6, 46, 54].

3 Experiments

In this section, we present empirical evaluation, the results of which demonstrate the superior performance of our proposed algorithm for class imbalance. We describe the experimental results on CIFAR-100-LT, a synthetic long-tailed benchmark. CIFAR-100-LT is examined with imbalance ratio of 100, where an imbalance ratio is defined as $N_{\text{max}}/N_{\text{min}}$. The related works, implementation setups, and additional results are detailed in Appendix.

In Figure 2, we report the performance when CUDA is applied to the various algorithms: CE, CE-DRW [6], LDAM-DRW [6], BS [46], RIDE [54] with 3 experts, RIDE+CMO [42], and BCL [63]. We include four different categories of accuracy: all, many, med(ium), and few. Each represents the average accuracy of all samples, classes containing more than 100 samples, 20 to 100 samples, and under 20 samples, respectively. Compared to the case without CUDA cases, balanced validation performance is increased when we apply the proposed approach.

4 Conclusion

In this study, we proposed CUDA to address the class imbalance problem. The proposed approach is also compatible with existing methods. To design a proper augmentation for LTR, we first studied the impact of augmentation strength for LTR. We found that the strength of augmentation for a specific type of class (*e.g.,* major class) could affect the performance of the other type (*e.g.,* minor class). From this finding, we designed CUDA to adaptively find an appropriate augmentation strength without any further searching phase by measuring the LoL score for each epoch and determining the augmentation accordingly. To verify the superior performance of the proposed approach, we examined each performance with various methods and obtained the best performance among the methods compared to CIFAR-100-LT, synthetically generated benchmarks.

References

[1] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6907, 2022.

[2] Shin Ando and Chun Yuan Huang. Deep over-sampling framework for classifying imbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 770–785. Springer, 2017.

[3] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyper-parameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

[4] 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.

[5] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–121, 2021.

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

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

[8] Tsz-Him Cheung and Dit-Yan Yeung. Adaaug: Learning class-and instance-adaptive data augmentation policies. In *International Conference on Learning Representations*, 2021.

[9] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019.

[10] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *European Conference on Computer Vision*, pages 95–110. Springer, 2020.

[11] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *European Conference on Computer Vision*, pages 694–710. Springer, 2020.

[12] 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.

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[14] 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.

[15] 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.

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

[17] Denis Gudovskiy, Luca Rigazio, Shun Ishizaka, Kazuki Kozuka, and Sotaro Tsukizawa. Autodo: Robust autoaugment for biased data with label noise via scalable probabilistic implicit differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16601–16610, 2021.

[18] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019.

[19] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pages 2535–2544. PMLR, 2019.

[20] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster autoaugment: Learning augmentation strategies using backpropagation. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[22] 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.

[23] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021.

[24] 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.

[25] 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.

[26] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2020.

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

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

[29] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3524–3533, 2017.

[30] 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.

[31] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2022.

[32] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246, 2020.

[33] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5212–5221, 2021.

[34] 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.

[35] Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M Robertson, and Yongxin Yang. Dada: Differentiable automatic data augmentation. *arXiv preprint arXiv:2003.03780*, 2020.

[36] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

[37] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. *Advances in Neural Information Processing Systems*, 32, 2019.

[38] Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *International Conference on Learning Representations*, 2022.

[39] 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.

[40] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

[41] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. {BOIL}: Towards representation change for few-shot learning. In *International Conference on Learning Representations*, 2021.

[42] 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.

[43] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 735–744, 2021.

[44] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[45] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[46] 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.

[47] Cédric Rommel, Thomas Moreau, Joseph Paillard, and Alexandre Gramfort. Cadda: Class-wise automatic differentiable data augmentation for eeg signals. *arXiv preprint arXiv:2106.13695*, 2021.

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[49] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Ricap: Random image cropping and patching data augmentation for deep cnns. In *Asian conference on machine learning*, pages 786–798. PMLR, 2018.

[50] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5486–5494, 2018.

[51] 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.

[52] Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pages 935–942, 2007.

[53] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

[54] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.

[55] Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? In *International Conference on Learning Representations*, 2021.

[56] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263. Springer, 2020.

[57] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[58] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[59] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[60] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[61] 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.

[62] Tianyi Zhou, Shengjie Wang, and Jeffrey Bilmes. Curriculum learning by dynamic instance hardness. *Advances in Neural Information Processing Systems*, 33:8602–8613, 2020.

[63] 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.

Appendix
CUDA: Curriculum of Data Augmentation for Long-tailed Learning

Owing to the page limitation of the main manuscript, we provide detailed information in this supplementary as follows. (1) In Appendix A, we provide the related works for long-tailed recognition and data augmentation. (2) In Appendix B, we summarize the experimental setup of Figure 1, and further explain why augmentation on one side causes performance degradation on the opposite side. (3) In Appendix C, we describe in detail our experimental setting, including dataset configuration, data preprocessing, and training implementation. (4) In Appendix D, we show ImageNet-LT performance on different size and architecture networks, training time analysis, and accuracy on the balanced dataset case. (5) In Appendix E, we present in detail the augmentation operations that `CUDA` utilizes. (6) In Appendix F, we describe the experimental setting of Figure 9d.

A Related works

Long-tailed Recognition (LTR). The datasets with class imbalances can lead DNNs to learn biases toward training data, and their performance may decrease significantly on the balanced test data. To improve the robustness of such models to imbalance, LTR methods have been evolving in two main directions: (1) reweighting [15, 6, 43] methods that reweight the loss for each class by a factor inversely proportional to the number of data points, and (2) resampling methods [28, 7, 2] that balance the number of training samples for each class in the training set. However, studies along these lines commonly sacrifice performance on majority classes to enhance that on minority classes, because the overfitting problem occurs with limited information on minority classes as a result of increasing the weight of a small number of minority samples.

Several methods have recently been developed to alleviate the overfitting issues in various categories: (1) two-stage training [6, 24, 39], (2) ensemble methods [61, 56, 54, 5], and (3) contrastive learning approach [23, 14, 63, 31, 34]. To re-balance the classifier layers after achieving a good representation on the imbalanced training dataset in an early phase, Cao et al. [6] proposed deferred resampling (DRS) and reweighting (DRW) approaches. Kang et al. [24] decoupled the learning procedure into representation learning and training linear classifier, achieved higher performance than previous balancing methods. Wang et al. [54] and Cai et al. [5] suggested efficient ensemble methods using multiple experts with a routing module and a shared architecture for experts to capture various representations. Liu et al. [38] found that self-supervised representations are more robust to class imbalance than supervised representations, and some works have developed supervised contrastive learning methods [25] for imbalanced datasets [14, 63, 34].

Another line of research has considered augmentation methods in terms of both input and feature spaces [26, 11, 33]. Recently, Park et al. [42] mixed minority and majority images by using CutMix with different sampling strategies to enhance balancing and robustness simultaneously. These methods commonly focus on utilizing the rich context of majority samples to improve the diversity of minority samples. Moreover, these augmentation-based methods are relatitvely in easy to apply orthogonally with other LTR methods.

Data Augmentation (DA). DA has been studied to mitigate overfitting which may occur due to a lack of data samples. Some works have been proposed to erase random parts of images to enhance the generalization performance of neural networks [16, 60, 29, 9]. Recently, variants of MixUp [59] have been proposed; this method combines two images with specific weights [50, 18, 49, 16, 53]. By aggregating two approaches, CutMix [58] was proposed to erase and replace a small rectangular part of an image into another image. In another line of research, methods have been proposed to automatically configure augmentation operations [12, 37, 35, 20, 17]. In addition, Cubuk et al. [13] randomly selected augmentation operations using the given hyperparameters of the number of sampling augmentation and their magnitudes. Recently, class-wise or per-sample auto-augmentation methods have also been proposed [8, 47].

B Detail for Figure 1

B.1 Experimental Settings

Major and minor group decomposition. To check the impact of augmentation on majority and minority classes, we split the training dataset into two clusters. The majority cluster is the top 50

classes by sorting through the number of samples for each class. The bottom 50 classes are in the minority cluster. For simplicity, we utilize class indices of 0 to 49 as the majority and 50 to 99 as the minority, respectively. For the balanced case, we utilize 0 to 49 classes as cluster 1, and the others as cluster 2.

Controlling augmentation strength. We set the augmentation strength as the number of augmentation and its augmentation magnitude by following the augmentation rule of `CUDA`. For example, the samples in the majority classes with magnitude parameter 4 represents that they are augmented with randomly sampled 4 augmentations with their own pre-defined augmentation magnitude.

Training setting. For heatmaps in Figure 1, we follow the training recipe of CIFAR-100-LT for CE case, *e.g.,* ResNet-32, learning rate of 0.1, and so on. Further details, hyperparameters, and datasets are described in section 3 and Appendix C.

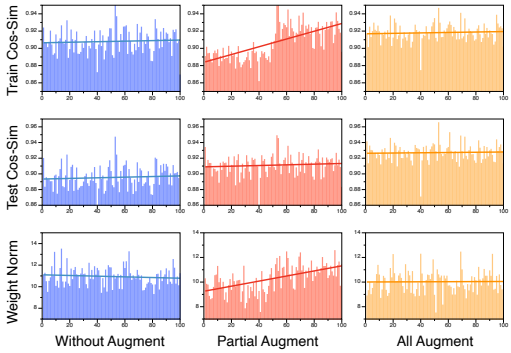


Figure 3: Analysis on Balanced CIFAR-100.

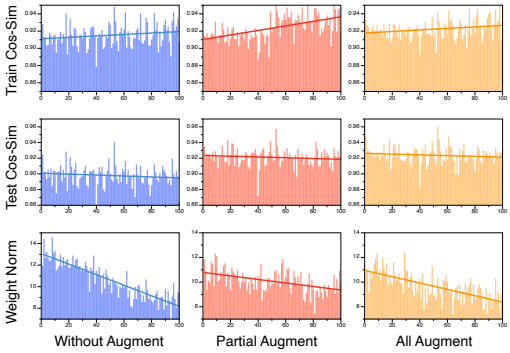


Figure 4: Analysis on CIFAR-100-LT (IR 100).

B.2 Analysis

Analysis for Figure 1. To figure out the reason for the phenomena in Figure 1, we conduct further analysis as shown in Figure 3 and Figure 4. Our experimental setups are as follows:

- Train the networks with three augmentation strategies, respectively (without, partial, and all), then measure the class-wise feature alignment and linear classifier weight norm for all networks. (Experiment 1)
- From a trained network without augmentation in Experiment 1, we freeze the feature extractor and train the linear classifier layer with augmenting partial classes. Then, we measure the class-wise L1-norm for all linear classifiers. (Experiment 2)

From the Figure 3 and Figure 4, we have three observations from Experiment 1:

1. When we conduct augmentation only for partial classes (0-49 classes), the feature alignment for augmented classes of the training dataset is degraded compared to the non-augmented classes. This is because the augmentation classes have more diversified training data than non-augmentation classes, which leads to more diversification in feature space. We observe the balance between alignment between classes in the cases of without augmentation and with all augmentation since all classes have similar diversity. (See the first rows in Figure 3, 4)
2. However, all three augmentation strategies have balanced class-wise feature alignment for the same test dataset. This tendency can be observed in both balanced and imbalanced datasets. This result is consistent with [24]. Furthermore, the values for feature alignment are increased when we conduct augmentation partially or all, compared to without augmentation. This result shows that augmentation enhances the feature extraction ability, which is consistent with conventional studies. (See the second rows in Figure 3, 4)
3. When we conduct augmentation only for partial classes on a balanced dataset, the class-wise weight norm of the linear classifier is larger for non-augmentation classes. This result incurs

performance improvement for non-augmentation classes and reduction for augmentation classes since this linear classifier has a tendency to classify non-augmented classes with larger weight values. However, we observe that class-wise weight norms are balanced in "without augmentation" and "all augmentation" cases. (See the third row in Figure 3)

4. We observe that the class-wise weight norm of the linear classifier is larger for majorities for all classes that have the same augmentation strength. These results are consistent with previous works [24, 1]. However, when we conduct augmentation only for majorities, the class-wise weight norm is more balanced. This phenomenon is similar to the balanced case in that partial augmentation incurs a reduction in the norm of the linear classifier for augmented classes. (See the third row in Figure 4)

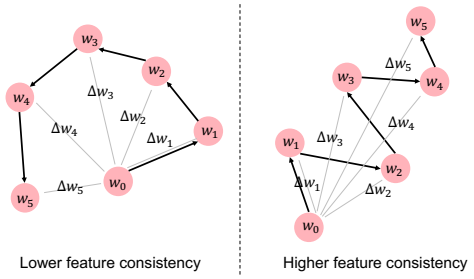


Figure 5: Concept of the impact of lower feature alignment on linear classifier.

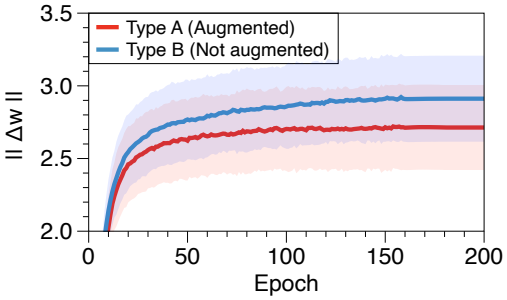


Figure 6: The difference of linear classifier norm $\|\Delta \mathbf{w}\|$ along training epoch.

Our observations from Experiment 1 are highly consistent in both balanced and imbalanced datasets. The results in Figure 1, Figure 3 and Figure 4 highly motivate the design of CUDA. Moreover, our results for Experiment 2 can explain these observations as shown in Figure 5 and Figure 6.

We observe that in the presence of feature alignment degradation from augmentation, the corresponding norm is relatively small, as shown in Figure 5. This is because in the class that has lower feature alignment, the variation of the gradient for the linear classifier is larger than in the class with high feature alignment. As shown in Figure 6, from Experiment 2, we observe that $\|\Delta \mathbf{w}\|$, the norm of class-wise difference of between current and initialize linear classifier parameters $\Delta \mathbf{w} := \mathbf{w} - \mathbf{w}_0$, have smaller value in augmented classes than non-augmented classes. From our experimental analysis in Figure 3, 4, and 6, we can conclude that augmentation breaks the consistency of feature alignment and it makes the weight norm of the linear classifier decreases.

C Implementation detail in section 3

C.1 Implementation Detail

Baselines. We compare CUDA with previous long-tailed learning algorithms , including cross-entropy loss (CE), two-stage approaches: CE-DRW [6] and cRT [24], balanced loss approaches: LDAM-DRW [6] and Balanced Softmax (BS; [46]), the ensemble method: RIDE with three experts [54], re-sampling algorithms: Remix [10] and CMO [42], and contrastive learning-based approach: BCL [63]. We integrate CUDA with CE, CE-DRW, LDAM-DRW, BS, RIDE, and BCL algorithms. For longer epochs, we compare CUDA with PaCo [14], BCL, and NCL [31], by combining CUDA with BCL and NCL. For a fair comparison of the computational cost, we train the network with the official one-stage implementation of RIDE (*i.e.,* without distillation and routing).

Implementation. For CIFAR-100-LT dataset, almost all implementations follow the general setting from [6], whereas cRT [24], BCL, NCL and RIDE follow the settings used in their original implementation. Following [6], we use ResNet-32 [22] as a backbone network for CIFAR-100-LT. The network is trained on SGD with a momentum of 0.9 and a weight decay of 2×10^{-4}. The initial learning rate is 0.1 and a linear learning rate warm-up is used in the first 5 epochs to reach the initial learning rate. During training over 200 epochs, the learning rate is decayed at the 160th and 180th epochs by 0.01. For the ImageNet-LT and iNaturalist, the ResNet-50 is used as a backbone network and is trained

for 100 epochs. The learning rate is decayed at the 60th and 80th epochs by 0.1. As with CIFAR, for cRT, RIDE, and BCL, we follow the original experimental settings of the official released code. For the hyperparameter values of CUDA, we apply a p_{aug} of 0.5 and T of 10 for all experiments. For γ, we set the values as 0.6 for CIFAR-100-LT and 0.4 for ImageNet-LT and iNaturalist 2018. The detailed implementation for baselines are in Appendix C.

CIFAR-100-LT. CIFAR-100-LT is a subset of CIFAR-100. Following [54, 42, 63], we use the same long-tailed version for a fair comparison. The number of samples of kth class is determined as follows: (1) Compute the imbalanced factor $N_{\text{max}}/N_{\text{min}}$, which reflects the degree of imbalance in the data. (2) $|\mathcal{D}_k|$ between $|\mathcal{D}_1| = N_{\text{max}}$ and $|\mathcal{D}_{100}| = N_{\text{min}}$ follows an exponential decay (*i.e.,* $|\mathcal{D}_k| = |\mathcal{D}_1| \times (N_{\text{max}}/N_{\text{min}})^{k/100}$). The imbalance factors used in the experiment are set to 100, 50, and 10.

ImageNet-LT. ImageNet-LT [39] is a modified version of the large-scale real-world dataset [48]. Subsampling is conducted by following the Pareto distribution with power value $\alpha = 0.6$. It consists of 115.8K images of $1,000$ classes in total. The most common or rare class has $1,280$ or 5 images, respectively.

iNaturalist 2018. iNaturalist [51] is a large-scale real-world dataset which consists of 437.5K images from $8,142$ classes. It has long-tailed property by nature, with an extremely class imbalanced. In addition to long-tailed recognition, this dataset is also used for evaluating the fine-grained classification task.

C.2 Data Preprocessing

For data preprocessing, we follow the default settings of [6]. For CIFAR-100-LT, each side of the image is padded with 4 pixels, and a 32×32 crop is randomly selected from the padded image or its horizontal flip. For ImageNet-LT and iNaturalist 2018, after resizing each image by setting the shorter side to 256 pixels, a 224×224 crop is randomly sampled from an image or its horizontal flip.

For BCL and NCL, which use AutoAugment [12] or RandAugment [13] as default data augmentation, we apply them after random cropping by following their original papers [63, 31]. Then, we finally conduct CUDA after all default augmentation operations, and then normalize the image with following mean and standard deviation values sequentially: CIFAR-100-LT ((0.4914, 0.4822, 0.4465), (0.2023, 0.1994, 0.2010)), ImageNet-LT ((0.485, 0.456, 0.406), (0.229, 0.224, 0.225)), and iNaturalist 2019 ((0.466, 0.471, 0.380), (0.195, 0.194, 0.192)).

C.3 Detailed Implementation

Because some official codes do not open their entire implementations, we re-implement by following the rules. For re-implementation, we reproduce the code based on their partial code and the authors' responses.

RIDE. We follow the officially offered code[1]. Among various experimental configurations of official code (*e.g.,* one-stage RIDE, RIDE-EA, Distill-RIDE), for fair comparison (to leverage similar computation resources), we utilize one-stage training (*i.e.,* one-stage RIDE) for all cases. We confirm that CMO [42] also utilizes this setup for RIDE + CMO from the response of the authors.

CMO. We re-implement all CMO results from their official code[2] in our work. However, the official code of CMO does not contain code for RIDE + CMO. Therefore, we re-implement by injecting the CMO part for BS in the official code (weighted sampler and mixup part) into the RIDE code. Furthermore, for iNaturalist 2018, we train the model for 100 epochs for a fair comparison with other methods (whereas the original RIDE + CMO is trained for 200 epochs on iNaturalist 2018).

BCL. The officially released code[3] of BCL only contains ImageNet-LT and iNaturalist 2018. Whereas the official code applies a cosine classifier for ImageNet-LT and iNaturalist 2018, we apply an ordinary linear classifier for CIFAR-100-LT from the author's response. All hyperparameters are the same as the experiment settings of the original work [63].

[1] https://github.com/frank-xwang/RIDE-LongTailRecognition

[2] https://github.com/naver-ai/cmo

[3] https://github.com/FlamieZhu/Balanced-Contrastive-Learning

Table 1: Validation accuracy on CIFAR-100-LT dataset. This experiments are conducted on ResNet-32 on CIFAR-100-LT with different imbalance ratios. † are from [42] and ‡, ⋆ are from the original papers [26, 63]. Other results are from our implementation. We format the first and second best results as **bold** and underline .

Algorithm	Imbalance Ratio (IR)			Statistics (IR 100)		
	100	50	10	Many	Med	Few
CE	38.7	43.4	56.5	66.2	37.3	8.2
CE + CMO [42]	42.0	47.0	60.0	69.1	41.2	11.3
CE + CUDA	42.7	47.2	59.6	**71.6**	42.3	9.4
CE + CMO + CUDA	43.5	48.7	60.0	70.0	43.4	12.7
CE-DRW [6]	41.4	45.5	57.8	62.8	41.7	16.1
CE-DRW + Remix [10]†	45.8	49.5	59.2	-	-	-
CE-DRW + CUDA	47.7	52.4	61.6	64.3	49.2	26.7
LDAM-DRW [6]	42.5	47.4	57.6	62.8	42.3	19.0
LDAM + M2m [26]‡	43.5	-	57.6	-	-	-
LDAM-DRW + CUDA	47.6	51.1	58.4	67.3	50.4	21.4
BS [46]	43.3	46.9	58.3	61.6	42.3	23.0
BS + CUDA	47.7	52.1	61.7	63.3	48.4	28.7
RIDE (3 experts) [54]†	48.6	51.4	59.8	-	-	-
RIDE (3 experts)	49.7	52.7	60.2	67.7	51.5	26.7
RIDE + CMO [42]†	50.0	53.0	60.2	-	-	-
RIDE + CMO	49.9	53.0	58.9	67.3	51.3	28.1
RIDE (3 experts) + CUDA	50.7	53.7	60.2	69.2	52.8	27.3
BCL [63]⋆	51.0	54.9	64.4	67.2	53.1	32.9
BCL + CUDA	**52.3**	**56.2**	**64.6**	66.4	**54.2**	**33.9**

Table 2: Validation accuracy on ImageNet-LT and iNaturalist 2018 datasets. † indicates reported results from the [42] and ‡ indicates those from the original paper [24]. ⋆ means we train the network with the official code in an one-stage RIDE.

Algorithm	ImageNet-LT				iNaturalist 2018			
	Many	Med	Few	All	Many	Med	Few	All
CE†	64.0	33.8	5.8	41.6	73.9	63.5	55.5	61.0
CE + CUDA	**67.1**	47.1	13.4	47.2	**74.6**	65.0	57.2	62.5
CE-DRW [6]	61.7	47.3	28.8	50.1	68.2	67.3	66.4	67.0
CE-DRW + CUDA	61.7	48.4	30.5	51.1	68.8	67.9	66.5	67.4
LWS [24]‡	57.1	45.2	29.3	47.7	65.0	66.3	65.5	65.9
cRT [24]‡	58.8	44.0	26.1	47.3	69.0	66.0	63.2	65.2
cRT + CUDA	62.3	47.2	28.1	50.2	68.2	67.9	66.4	67.3
LDAM-DRW [6]†	60.4	46.9	30.7	49.8	-	-	-	66.1
LDAM-DRW + CUDA	63.2	48.2	31.2	51.5	68.0	67.5	66.8	67.3
BS [46]	60.9	48.8	32.1	51.0	65.7	67.4	67.5	67.3
BS + CUDA	61.8	49.1	31.8	51.5	67.6	68.2	68.3	68.2
RIDE (3 experts) [54]⋆	64.9	50.4	34.4	53.6	70.4	71.8	71.8	71.6
RIDE + CMO [42]⋆	65.6	50.6	34.8	54.0	68.0	70.6	72.0	70.9
RIDE (3 experts) + CUDA⋆	66.0	51.7	34.7	54.7	70.6	72.6	**72.7**	**72.4**
BCL [63]	65.3	53.5	36.3	55.6	69.4	72.4	71.8	71.8
BCL + CUDA	66.8	**53.9**	**36.6**	**56.3**	70.8	**72.7**	72.0	72.2

Table 3: Comparison for CIFAR-LT-100 performance on ResNet-32 with 400 epochs.

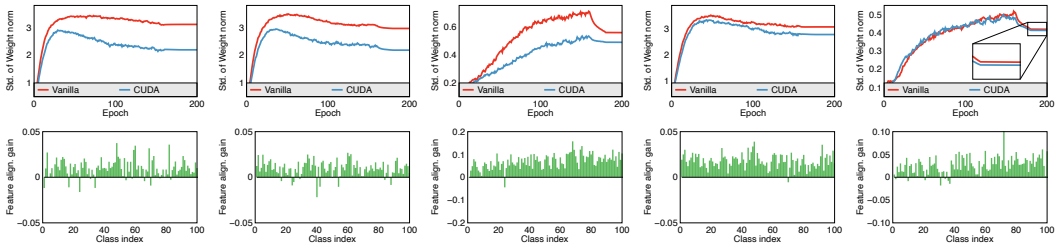
Algorithm	Imbalance Ratio	
	100	50
PaCo	52.0	56.0
BCL	52.6	57.2
NCL	54.2	58.2
BCL + CUDA	53.5	57.4
NCL + CUDA	**54.8**	**59.6**

Table 4: Augmentation analysis on CIFAR-100-LT with IR 100. AA [12], FAA [37], DADA [35], and RA [13] with $n = 1, m = 2$ policies are used. C, S, I represent CIFAR, SVHN, and ImageNet policy.

	Vanilla	AA			FAA		DADA		RA	CUDA
		C	S	I	C	I	C	I		
CE	38.7	41.7	40.7	40.1	42.3	40.8	41.0	41.2	40.5	**42.7**
CE-DRW	41.4	46.5	44.7	45.5	46.3	44.8	45.6	45.7	45.8	**47.4**
LDAM-DRW	42.5	47.0	44.7	44.9	46.6	45.6	45.9	46.5	44.0	**47.2**
BS	43.3	47.0	46.1	45.5	46.5	45.0	45.0	46.9	45.2	**47.7**
RIDE (3 experts)	49.7	49.5	47.3	45.5	49.8	50.6	50.4	50.5	47.9	**50.7**

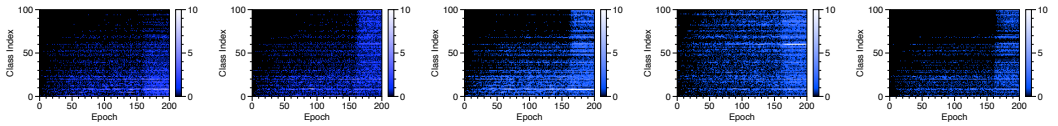
D Analyses

CIFAR-100-LT. In Figure 2, we report the performance when CUDA is applied to the various algorithms: CE, CE-DRW [6], LDAM-DRW [6], BS [46], RIDE [54] with 3 experts, RIDE+CMO [42],



(a) CE (b) CE-DRW (c) LDAM-DRW (d) BS (e) RIDE

Figure 7: Analysis of how CUDA improves long-tailed recognition performance, classifier weight norm (top row) and feature alignment (bottom row) of the CIFAR-100-LT validation set. Notably that weight norm and feature alignment represent class-wise weight magnitude of classifier and ability of feature extractor, respectively. The detailed analysis is described in Section D.1.



(a) CE (b) CE-DRW (c) LDAM-DRW (d) BS (e) RIDE

Figure 8: Evolution of LoL score on various algorithms, CE, CE-DRW, LDAM-DRW, BS, and RIDE.

and BCL [63]. Compared to the case without CUDA cases, balanced validation performance is increased when we apply the proposed approach.

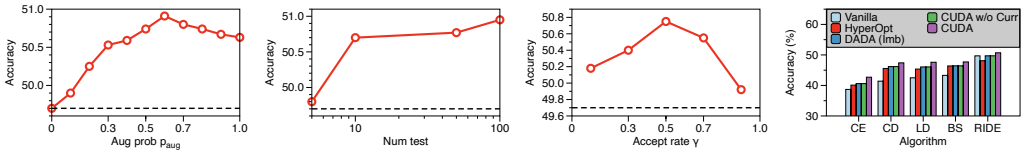
Recently, some works [14, 1, 63, 31] have shown impressive performances with diverse augmentation strategies and longer training epochs. For a fair comparison with these methods, we examine CUDA using the same experimental setups from PaCo ([14]; 400 epochs with batch size of 64). Table 3 shows that augmented image using CUDA can enhance LTR performance compare to the other baselines. In particular, CUDA with NCL obtains the best performance over 400 epochs. As noted by [31], the NCL algorithm utilizes six times as much memory compared to the vanilla architecture with three experts. Hereinafter in large-scale benchmarks, we focus on the cases with similar network size.

ImageNet-LT and iNaturalist 2018. To evaluate the performance of CUDA on larger datasets, we conduct experiments on ImageNet-LT [39] and iNaturalist 2018 [51]. Table 2 summarizes the performance of various LTR methods and the performance gain when integrated with CUDA. Our proposed method consistently improves performance regardless of the LTR method and target dataset by simply adding class-wise data augmentation without complicated methodological modification. Additionally, to evaluate the performance gain of CUDA on other architectures, we experiment with CUDA on ImageNet-LT with ResNet-10 [39] and ResNeXt-50 [57], as reported in Appendix D.

D.1 Analysis

We design our analyses to answer the following questions. (1) How does CUDA perform? (2) Does CUDA perform better than other augmentation methods? (3) How does LoL score change over training epochs when combined with various LTR methods? (4) Which part of CUDA is important to improvet performance? These analyses provide additional explanations to understand CUDA. All experiments are conducted on CIFAR-100-LT with imbalance ratio of 100.

How does CUDA mitigate the class imbalance problem? To deeply understand CUDA, we observe two types of metrics: (1) variance of classifier weight L1-norm, (2) feature alignment gain (*i.e.,* cosine similarity with and without CUDA) on validation dataset. The classifier weight norm is usually used to measure how balanced the model consider the input from a class-wise perspective [24, 1]. Feature alignment, especially feature cosine similarity amongst samples belonging to the same class, is a measure of the extent to which the extracted features are aligned [41]. As shown in Figure 7, CUDA has two forces for alleviating imbalance. For all cases, CUDA reduces the variance of the weight norm (*i.e.,* balance the weight norm), and thus the trained model consider the minority classes in a balanced manner. Note that because LDAM-DRW and RIDE utilize a cosine classifier (*i.e.,* utilizing L2 normalized linear weight), their standard deviation scale is quite different from those other methods. Because LDAM-DRW, BS, and RIDE include balancing logic in their loss function, they exibit lower



(a) Aug prob. p_{aug}. (b) Num test T. (c) Accept rate γ. (d) Curriculum.

Figure 9: Additional analysis of CUDA. (a) sensitivity of augmentation probability p_{aug}, (b) sensitivity analysis of number of sample coefficient T, (c) sensitivity of acceptance threshold γ, and (d) impact of curriculum. The dotted lines in (a), (b) and (c) represents the performance of CE.

variance reduction compared to the CE and CE-DRW. Second, as shown in the bottom row in Figure 7, CUDA obtains feature alignment gains for almost all classes. This shows that CUDA facilitates a network to learn to extract meaningful features.

Compared with other augmentations. To verify the impact of CUDA, we examine the other augmentation methods as follows. We compare five augmentation methods, including AutoAugment (AA, [12]), Fast AutoAugment (FAA, [37]), DADA [35], RandAugment (RA, [13]), and the proposed method CUDA. Because AA, FAA, and DADA provide their policies searched by using CIFAR, SVHN (for AA), and ImageNet, we leverage their results. Furthermore, RA suggests using their parameter $(n, m) = (1, 2)$ for CIFAR, and we follow their guidelines. As shown in Table 4, even though the augmentation policies use additional computation resources to search, CUDA outperforms the other pre-searched augmentations. This shows that CUDA is computationally efficient.

Dynamics of LoL score. We evaluate how LoL scores vary with algorithms: CE, CE-DRW, LDAM-DRW, BS, and RIDE. Note that we set a lower class index (*i.e.,* 0) as the most common class (*i.e.,* the number of samples is 500), while an index of 100 represents the rarest class (*i.e.,* with five samples). As described in Figure 8, as training progressed, the LoL score of all algorithms increase. After learning rate decay (*i.e.,* 160 epoch) all algorithms are able to learn to classify minority classes more easily than before. In particular, except for BS, the majority classes of most algorithms show a steep increment. The reason that BS exhibit a similar increasing speed for majority and minority classes is that it includes a module to balance the impact of majority and minority samples. Moreover, because RIDE can promote when all experts satisfying promotion criterion, its increasing speed is moderate.

Parameter sensitivity. For further analysis, we conduct a sensitivity analysis of hyperparameters in CUDA. More precisely, we study three kinds of parameters, including augmentation probability p_{aug} (Figure 9a), number of tests T (Figure 9b), and LoL update threshold γ (Figure 9c). We examine each hyperparameter sensitivity on a CUDA case with RIDE and the remainder of the hyperparameters are fixed to the default values in Section **??**. All results show that the performance gains of CUDA decreases if the parameters are adjusted to make the augmentation too strong or weak. For example, the augmentation strength of all classes steeply increases when γ becomes small. The strength cannot increase when γ becomes large, and thus it cannot improve the performance of the model. Moreover, as shown in Figure 9b, the performance of CUDA increases as T increases. However, larger T spends computational overhead, we set T as 10 and obtained cost-effective performance gain.

Impact of curriculum. In addition to studying the impact of CUDA, we examine its performance component-wise. In particular, we test the case where class-wise augmentation strength is searched based on the hyperparameter optimization algorithm. We check five cases overall: vanilla algorithm, hyperparameter optimization (HO), re-searched DADA for CIFAR-100-LT, CUDA without curriculum, (*i.e.,* re-training utilizing the final augmentation strength of CUDA), and CUDA. As described in Figure 9d, CUDA finds better augmentation strengths compare to the hyperparameter search case. This means that CUDA exhibits not only a lower searching time but also obtains better augmentation strength. Moreover, by comparing the performance of with or without curriculum, the curriculum also can provide additional advance to the model to achieve better generalization. Additionally, as Figure 8, lower augmentation strength at the beginning of training is more effective than static higher augmentation strength. These results are consistent with the results of previous studies on curriculum learning methods [19, 62].

Training Time Analysis. CUDA requires additional computation for computing LoL score. We measure the additional training time for adding CUDA on various algorithms. As shown in Figure 11, when utilizing CUDA additional training time is spent. However, the additional operation for searching

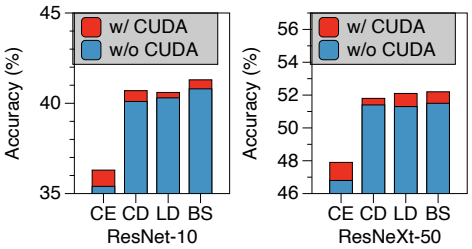


Figure 10: Network architecture.

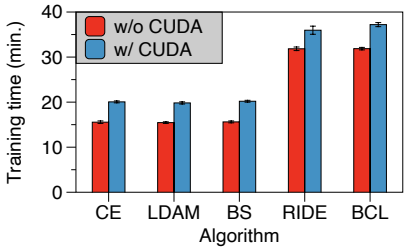


Figure 11: Training time.

the LoL score does not require a large value. For example, BS with CUDA spends $\times 1.29$ time to obtain adequate augmentation strength.

Network Architecture Analysis. We also present our ResNet-10 [39] and ResNeXt-50 [57] experiments on the ImageNet-LT dataset in Figure 10, respectively. These results show that CUDA consistently improves performance regardless of network sizes and corresponding LTR methods.

Table 5: Balanced case. † mark represents the reported value in [35]

Augmentation	Acc.	Searching time (Overhead)
CE	68.5	-
AutoAug	70.7	$5,000$ GPU hours†
RandAug	69.4	-
FAA	70.7	3.5 GPU hours†
DADA	70.9	0.2 GPU hours†
CUDA	70.4	0.07 GPU hours

What if CUDA is ran on the balanced dataset. We examine that if CUDA is applied to the balanced case, *i.e.,* imbalance ratio is 1. As described in the Table 5 CUDA obtains 1.9% accuracy gain, which is lower than the other auto augmentation methods. However, other autoaugmentation methods spend more computation time searching a good augmentation than CUDA. Furthermore, as described in Figure 8, CUDA has higher performance than the others when the class imbalance dataset is given.

E Augmentation Preset

E.1 Data augmentation operations used in CUDA.

There have been numerous data augmentation operations in vision tasks. We used totally 22 augmentations for CUDA with their own parameter set. Details of the operation set and parameters are described in Table 6. For augmentation magnitude parameter $m_k(s)$, we divide parameters into thirty values linearly. For example of, ShearX case, its max and min values are 3 and 0, respectively. Therefore, $m_{\text{ShearX}}(s) = (3-0)/30 * s$, thus $m_{\text{ShearX}}(1) = 0.01 = (3-0)/30 * 1$.

F Experimental setting of Figure 9d

To further analyze the impact of curriculum, we compare CUDA with the performance of previous hyper-parameter search algorithms and auto-augmentation methods, especially DADA [35]. We describe each setting in detail as follows.

Hyper-parameter search. We utilize the strength score-based augmentation module in CUDA to verify the hyper-parameter search. In other words, samples in each class utilize K augmentation operations. Therefore, we search the class-wise augmentation on the search space K^N where N is the number of classes. We leverage the hyper-parameter searching open-source library, Ray [36], for search K^N space efficiently. Among various search modules, we utilize the HyperOptSearch module, which is the implementation of the Tree-structured Parzen Estimator [3]. Moreover, for fast search, we use the Asynchronous Successive Halving Algorithm (ASHA) [32]. We run $1,000$ trials for each algorithms which spends almost 20 GPU hours (*i.e.,* $\times 80$ overhead compare to CUDA).

(a) Raw (b) Flip (c) Mirror (d) Edge Enhance (e) Detail (f) Smooth

(g) AutoContrast (h) Equalize (i) Invert (j) Gaussian Blur (k) Resize Crop (l) Rotate

(m) Posterize (n) Solarize (o) SolarizeAdd (p) Color (q) Contrast (r) Brightness

(s) Sharpness (t) ShearX (u) ShearY (v) Translate X (w) Translate Y

Table 6: Description of augmentation operations utilized in `CUDA`. We show the examples of each augmentation with maximum augmentation parameters.

Operation	Parameter	Description
`Flip`	On/Off	Flip top and bottom
`Mirror`	On/Off	Flip left and right
`Edge Enhancement`	On/Off	Increasing the contrast of the pixels around the targeted edges
`Detail`	On/Off	Utilize convolutional kernel $[[0,-1,0],[-1,10,-1],[0,-1,0]]$
`Smooth`	On/Off	Utilize convolutional kernel $[[1,1,1],[1,5,1],[1,1,1]]$
`AutoContrast`	On/Off	Remove a specific percent of the lightest and darkest pixels
`Equalize`	On/Off	apply non-linear mapping to make uniform distribution
`Invert`	On/Off	Negate the image
`Gaussian Blur`	[0,2]	Blurring an image using Gaussian function
`Resize Crop`	[1,1.3]	Resizing and center random cropping
`Rotate`	[0,30]	Rotate the image
`Posterize`	[0,4]	Reduce the number of bits for each channel
`Solarize`	[0,256]	Invert all pixel values above a threshold
`SolarizeAdd`	[0,110]	Adding value and run solarize
`Color`	[0.1, 1.9]	Colorize gray scale values
`Contrast`	[0.1,1.9]	Distance between the colors
`Brightness`	[0.1,1.9]	Adjust image brightness
`Sharpness`	[0.1,1.9]	Adjust image sharp
`Shear X`	[0,0.3]	Shearing X-axis
`Shear Y`	[0,0.3]	Shearing Y-axis
`Translate X`	[0,100]	Shift X-axis
`Translate Y`	[0,100]	Shifting Y-axis

Researched DADA operation on imbalanced CIFAR. Because the officially offered policies on CIFAR by DADA [35] are searched for a balanced CIFAR dataset, we have to re-search the

augmentation policy for the imbalanced dataset. We utilize the official code of DADA and replace the dataloader to re-search the operations. It spends 48 minutes for searching the augmentation policy ($\times 8.6$ than the overhead of CUDA). Despite this additional overhead, DADA outputs worse performance than `CUDA` (even `CUDA` without curriculum case). This is because (1) DADA does not consider class-wise augmentation and (2) it does not consider the impact of class imbalance.