7 Appendix

7.1 Details of CutMix

To sample the binary mask $\mathbf{M}$, we first sample the bounding box coordinates $\mathbf{B} = (r_x, r_y, r_w, r_h)$ indicating the cropping regions on x_i and x_j. The region $\mathbf{B}$ in x_i is removed and filled in with the patch cropped from $\mathbf{B}$ of x_j. In our experiments, we sample rectangular masks $\mathbf{M}$ whose aspect ratio is proportional to the original image. The box coordinates are uniformly sampled according to the:

$$\begin{aligned} r_x &\sim \mathrm{Unif}(0, W), \quad r_w = W\sqrt{1-\lambda}, \\ r_y &\sim \mathrm{Unif}(0, H), \quad r_h = H\sqrt{1-\lambda} \end{aligned} \tag{11}$$

making the cropped area ratio $\frac{r_w r_h}{WH} = 1 - \lambda$. With the cropping region, the binary mask $\mathbf{M} \in \{0, 1\}^{W \times H}$ is decided by filling with 0 within the bounding box $\mathbf{B}$, otherwise 1.

7.2 More Comparison Results

In this section, we conduct comparison studies with mixup-based long-tailed methods, including Remix [21], Unimix [22], and CMO [20] under various loss functions. We summarize the results on CIFAR-LT-10 and CIFAR-LT-100 in Table 6 and list the results on ImageNet-LT and iNaturalist 2018 in Table 5. We can find that our OTmix surpasses related mixup-based methods with varying loss functions and enhances the imbalanced classification. Besides, OTmix with different losses produces different classification performances due to their characteristics, where OTmix with BALMS can achieve better or competing performance than OTmix with other loss functions. These results reveal the effectiveness of our proposed method when combined with other loss functions.

Table 5: Top-1 errors (%) of mixup-based long-tailed methods under various loss functions on ImageNet-LT and iNaturalist 2018. "⋆": results reported in CMO. "†": results reported in origin paper.

Method		**ImageNet-LT**				**iNaturalist 2018**			
Mixup	Loss	ALL	Many	Medium	Few	ALL	Many	Medium	Few
None⋆	ERM	58.4	36.0	66.2	94.2	39.0	26.1	36.5	44.5
Remix†	ERM	58.3	–	–	–	38.7	–	–	–
CMO†	ERM	50.9	33.0	57.7	79.5	31.1	**23.1**	30.7	33.4
OTmix	ERM	**48.0**	**30.0**	**54.1**	**77.7**	**30.5**	30.7	**29.5**	**31.6**
None	ERM-DRW	49.9⋆	38.3⋆	52.7⋆	71.2⋆	36.3†	–	–	–
Remix†	ERM-DRW	–	–	–	–	29.5	–	–	–
CMO†	ERM-DRW	48.6	39.2	51.4	**64.5**	29.1	31.8	29.8	**27.8**
OTmix	ERM-DRW	**46.6**	**33.0**	**51.0**	69.6	**28.9**	**29.4**	**28.1**	29.6
None⋆	LDAM-DRW	50.2	39.6	53.1	69.3	30.0	30.0	29.8	30.1
CMO†	LDAM-DRW	48.9	38.0	52.6	**69.2**	30.9	**24.7**	30.5	32.7
OTmix	LDAM-DRW	**47.5**	**37.2**	**48.9**	71.8	**30.4**	31.5	**29.8**	**30.6**
None⋆	BALMS	49.0	39.1	51.2	67.9	30.0	30.0	29.8	30.1
CMO†	BALMS	47.7	38.0	50.9	63.3	29.1	31.2	30.0	**27.7**
OTmix	BALMS	**44.4**	**36.0**	**47.6**	**57.3**	**28.5**	**28.9**	**28.0**	29.2

7.3 More Analytical Results

Confusion Matrix To verify whether our method improves the performance of minority classes, we show the confusion matrices of ERM-DRW, ERM-DRW+CMO, and ERM-DRW+OTmix on CIFAR-LT-10 with $\varphi = 100$ in Fig. 6. We can find that ERM-DRW suffers a severe performance drop in the minority classes even though it can almost accurately predict the samples in the majority classes. ERM-DRW+CMO can improve the accuracy of the minority classes, which coincides with the statement of CMO. OTmix further enhances the generalization of minority classes and

Table 6: Top-1 (%) errors of mixup-based long-tailed methods with various loss functions on CIFAR-LT-10 and CIFAR-LT-100. "‡ ": our reproduced results. "†": results reported in the original paper.

Method		**CIFAR-LT-10**			**CIFAR-LT-100**		
Mixup	Loss	100	50	10	100	50	10
None‡	ERM	26.9	22.9	12.9	63.2	56.3	43.4
CMO	ERM	25.0‡	18.6‡	11.5‡	56.1†	51.7†	40.5†
Remix†	ERM	24.6	–	11.8	58.1	–	40.6
UniMix†	ERM	23.5	–	–	58.5	–	–
OTmix	**ERM**	**21.7**	**16.6**	**9.8**	**53.6**	**49.3**	**38.4**
None‡	LDAM	26.3	21.6	13.4	61.1	56.0	44.4
CMO‡	LDAM	25.9	22.7	13.1	58.0	54.3	44.1
UniMix†	LDAM	24.6	–	–	58.3	–	–
OTmix	**LDAM**	**22.3**	**18.0**	**12.0**	**56.3**	**50.9**	**41.5**
None‡	ERM-DRW	24.3	18.9	11.9	58.0	54.3	41.8
Remix†	ERM-DRW	20.2	–	11.0	53.2	–	38.8
CMO	ERM-DRW	19.5‡	16.6‡	11.3‡	53.0†	49.1†	38.3†
OTmix	**ERM-DRW**	**16.9**	**13.8**	**9.4**	**52.0**	**47.4**	**37.3**
None‡	LDAM-DRW	23.0	19.1	11.8	57.4	52.2	45.0
Remix†	LDAM-DRW	20.7	–	13.2	55.0	–	40.5
CMO	LDAM-DRW	19.0‡	16.2‡	12.4‡	52.8†	48.3†	41.6†
OTmix	**LDAM-DRW**	**18.2**	**16.0**	**11.8**	**52.0**	**47.6**	**41.0**
None‡	BALMS	22.7	19.1	11.8	58.0	53.1	41.6
CMO	BALMS	19.7‡	15.9‡	11.0‡	53.4†	48.6†	**37.7**†
OTmix	**BALMS**	**16.0**	**13.5**	**9.8**	**53.2**	**47.7**	**37.7**
None‡	Focal	30.4	23.4	13.6	61.6	56.3	45.0
CMO‡	Focal	29.0	22.6	12.5	58.1	53.6	41.8
OTmix	**Focal**	**27.0**	**20.9**	**12.0**	**57.8**	**53.3**	**40.6**
None‡	CB Softmax	26.0	21.1	12.3	–	–	–
CMO‡	CB Softmax	25.7	20.2	12.2	–	–	–
OTmix	**CB Softmax**	**24.0**	**18.8**	**12.0**	–	–	–
None‡	CB Sigmoid	26.5	22.4	12.5	–	–	–
CMO‡	CB Sigmoid	27.4	20.8	12.1	–	–	–
OTmix	**CB Sigmoid**	**24.9**	**20.6**	**11.8**	–	–	–

maintains performance in majority classes, which thus outperforms the strong baselines on the overall performance. In Fig.7, we plot the classification results for each class on CIFAR-LT-10 with $\varphi = 100$, where we adopt the BALMS loss and LDAM loss, respectively. Compared with these baselines, OTmix provides a significant improvement in minority classes. We specifically note that the proposed method improves the accuracy over BALMS by 23% and over LDAM by 15% for the least frequent class 9 while degrading the accuracy for class 0 and class 1 by less than 2%. These results indicate that ours can achieve a more balanced classifier and ameliorate the generalization of minority classes.

Discussion of Hyper-parameters and Training Settings To analyze the effect of different hyper-parameters and settings of OTmix, we conduct analytical experiments on CIFAR-LT-10 with $\varphi = 100$. The hyper-parameters include ω, α, and r, respectively. ω in (7) is employed to manage the degree of combination of the confusion matrix and feature information in Fig. 8(a). The best performance is achieved when ω = 0.05, indicating the class-level cost based on the confusion matrix dominating the cost function. α affects the combined ratio in Cutmix in Appendix 7.1. As shown in Fig. 8(b), the larger the value of α is, the more complementary the distribution tends to be closer to a uniform distribution. The OTmix achieves an accuracy of 46.4% when α = 4.

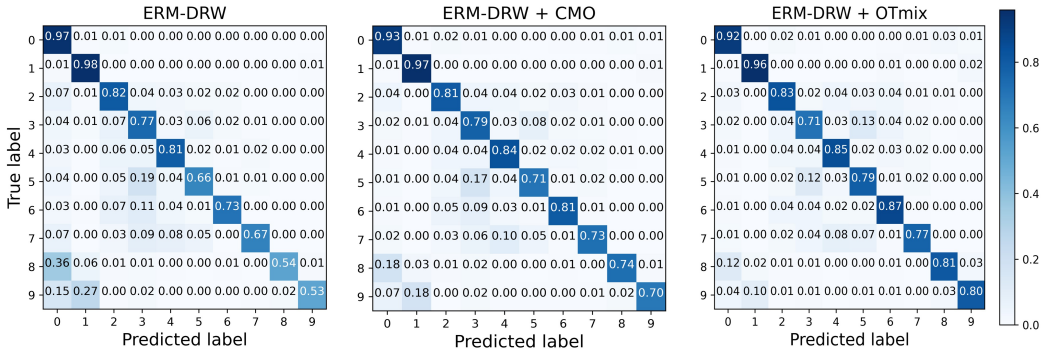
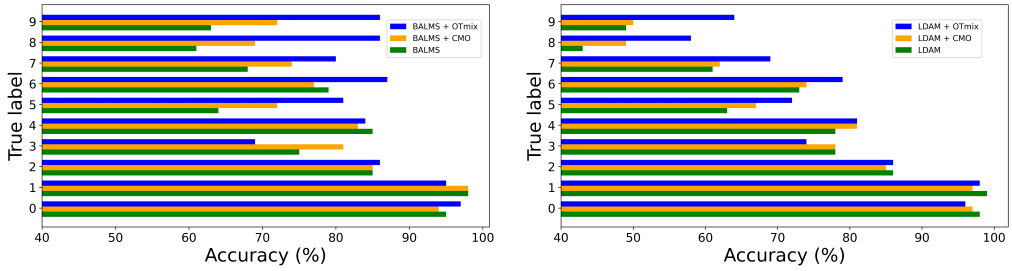
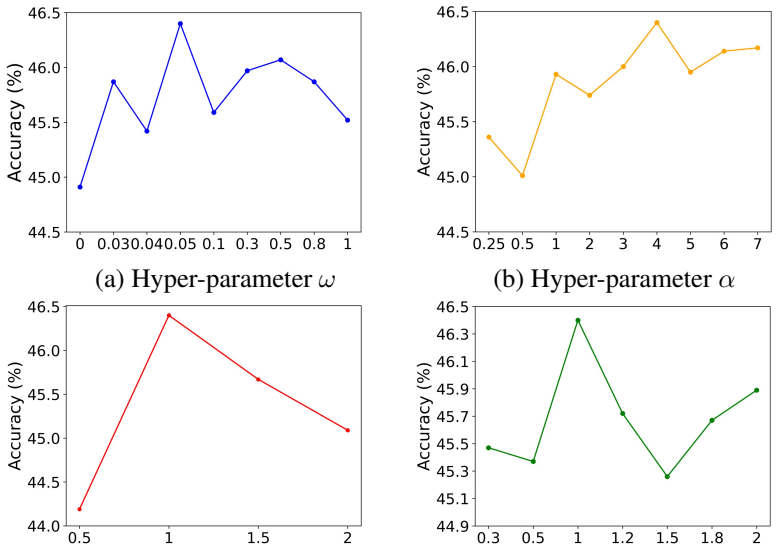


Figure 6: Confusion matrices of the ERM-DRW, ERM-DRW+CMO, and ERM-DRW+OTmix on CIFAR-LT-10 with $\varphi = 100$.



(a) BALMS baseline (b) LDAM baseline

Figure 7: The classification results of different methods for each class on CIFAR-LT-10 with $\varphi = 100$, where (a) uses the BALMS loss and (b) adopts the LDAM loss. Class 0 stands for the majority class, and class 9 stands for the minority class.



(a) Hyper-parameter ω (b) Hyper-parameter α

(c) Hyper-parameter r (d) Training factor μ

Figure 8: Analytical experiments of different hyper-parameters and settings of the proposed method on CIFAR-LT-100 with $\varphi = 100$: (a-c) with various hyper-parameters, (d) with the setting of training factor.

The hyper-parameter r controls the smoothness of q_k, which further decides the sample weight of foreground distribution in (4). As shown in Fig. 8(c), the smoothness factor $r = 1$ achieves the best performance, which indicates the inverse class frequency. Recalling that we adopt $y_{\text{random}} \sim \text{Bernoulli}(\frac{t}{T})$ to decide whether mixing the background and foreground pair with our OTmix, where we use μ to set $y_{\text{random}} \sim \text{Bernoulli}(\frac{t}{T})^{\mu}$ and explore the effect of μ. As plotted in Fig. 8(d), ours with $\mu = 1$ produces the best performance. That is to say, we will randomly mix the background and foreground images with a high probability during the first half of training epochs, and we are more likely to use OTmix to mix the background and foreground pairs in the latter half of epochs.

Mixed Fractions To examine the indispensability of the images unchanged in every batch, we conduct an additional experiment with the fraction of the mixed images in every batch on CIFAR-LT-10 and CIFAR-LT-100 under different methods and fractions. We denote $\boldsymbol{Fraction} = \frac{\boldsymbol{Mixed}}{\boldsymbol{Overall}}$ in every batch. From the results in Table 7, when the number of mixed samples decreases ($\boldsymbol{Fraction} \downarrow$), the performance of OTmix deteriorates significantly, which suggests that it is better to use mixed samples alone in every batch in our method.

Table 7: Classification errors of ResNet-32 on CIFAR-LT-10 and CIFAR-LT-100 under different methods and mixed fractions.

Method	$\boldsymbol{Fraction}$	**CIFAR-LT-10**			**CIFAR-LT-100**		
		100	50	10	100	50	10
ERM	/	26.9	22.9	12.9	63.2	56.3	43.4
ERM + CMO	100%	25.0	18.6	11.5	56.1	51.7	40.5
ERM + OTmix	**100%**	**21.7**	**16.6**	**9.8**	**53.6**	**49.3**	**38.4**
ERM + OTmix	50%	24.7	20.0	12.3	55.1	52.0	40.4
ERM + OTmix	25%	23.6	20.1	11.1	58.1	53.1	40.5
ERM + OTmix	12.5%	27.2	21.8	12.1	57.1	52.7	41.1

Discussion of Confusion matrix Considering the confusion matrix played an important role in learning the cost function based class-level, we discuss the performance of our method and the computational cost of the normalized confusion matrix calculated on a balanced validation set, an imbalanced training set, and a small balanced subset sampled from imbalanced training set in Table 8. Meanwhile, the confusion matrix is represented as two states, fixed and adaptive. The former denotes where the confusion matrix remains unchanged in our approach and the latter changes dynamically in each epoch. From Table 8, we can draw a few observations: (1) With the same settings, the adaptive methods perform significantly better overall and few than the fixed methods. (2) Compared with the fixed balanced training setting D^{bal}_{fixed}, the fixed imbalanced training setting D^{im}_{fixed} is preferable by providing more sample information ($D^{im}_{fixed} \gg D^{bal}_{fixed}$). However, large-scale samples can drastically increase the computational cost, making it difficult to implement adaptively. (3) Despite the time spent in the balanced training setting less, the balanced validation setting exhibits superior in terms of overall performance. To summarize, OTmix with the adaptively balanced validation setting enhances the suitability of OTmix for long-tailed classification.

Table 8: Classification errors of the ERM+OTmix with different confusion matrices under various methods and calculated settings on iNaturalist 2018.

Method	**Setting**	**ALL**	**Many**	**Medium**	**Few**	**Time**
Fix	Balanced validation	**31.0**	**29.3**	**29.5**	33.3	86s
Fix	Imbalanced training	31.5	31.0	30.2	**33.2**	1277s
Fix	Balanced training	31.8	32.7	30.3	33.5	**58s**
Adaptive	Balanced validation (OTmix)	**30.5**	**30.7**	**29.5**	31.6	86s
Adaptive	Balanced training	31.1	32.9	30.6	**31.1**	**58s**

7.4 Computational Cost

The optimal transport (OT) problem in our method between probability distributions is computed by the Sinkhorn algorithm [68], which introduces the entropic regularization term for fast computation. To compute the OT distance between n dimensional discrete distributions, the Sinkhorn algorithm requires the computational cost of $\mathcal{O}(n^2 log(n)/\varepsilon^2)$ reach ε-accuracy. In our case, n corresponds to the batchsize, which is set to 128 in our experiments. We compare the computational cost of different methods on a Pentium PC with a single GTX 3060 GPU.

Table 9: Computational cost (s) per training epoch on long-tailed datasets.

Method	**CIFAR-LT-10**	**CIFAR-LT-100**	**ImageNet-LT**	**iNaturalist 2018**
ERM	2.85	2.17	310.47	1251.29
ERM+CMO	3.35	2.85	319.73	1360.66
ERM+OTmix	4.59	3.79	333.64	1382.54

In addition, we also report the computational cost (s) per training epoch on long-tailed and balanced datasets, respectively. As shown in Table 9 and Table 10, mixup-based methods usually take more time than ERM. It is reasonable since mixup-based methods need to mix images. Besides, OTmix spends more time than CMO since we solve an OT problem to pair a background image and a foreground image. Still, introducing OTmix to existing mixup-based methods in balanced classification consumes more time. However, it is worth noticing that we only need to employ the OTmix during the late half phase of the training process. In summary, combining ours with others produces a better performance on long-tailed and balanced datasets with an acceptable cost.

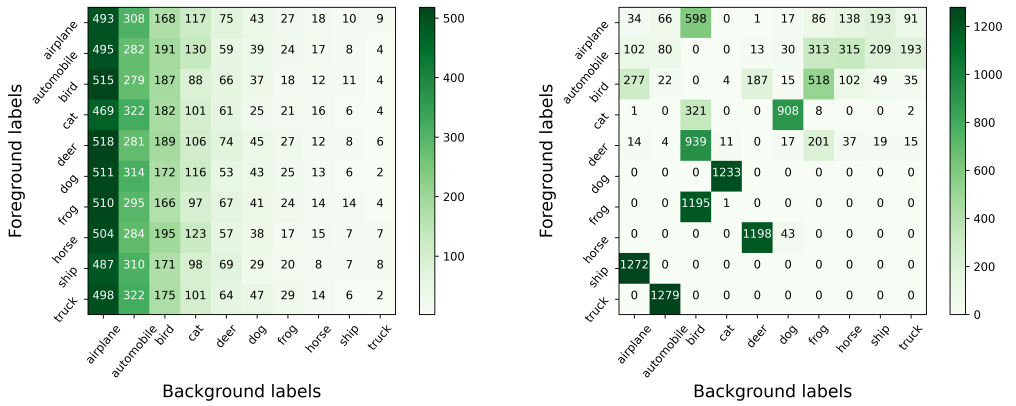
Table 10: Computational cost (s) per training epoch on balanced datasets.

Method	**CIFAR-10**	**CIFAR-100**
ERM	11.9	12.8
ERM+Mixup	12.1	13.8
ERM+OTmix (Mixup)	17.1	18.1
ERM+Cutmix	13.2	15.7
ERM+OTmix (Cutmix)	19.1	21.5
ERM+SaliencyMix	14.2	17.9
ERM+OTmix (SaliencyMix)	21.5	24.9

7.5 More Visualization Results and Analysis

Statistical Results of Mixed Images To intuitively reveal that OTmix is more effective than CMO, we show the statistical results of the mixed images generated by CMO and OTmix on CIFAR-LT-10, respectively. Specifically, we summarize the 10×10 matrix m for the ten-class classification task in one training epoch, where element m_{ij} denotes the number of pairs between the foreground images from the i-th class and the background images from the j-th class. As shown in Fig. 9(a), we can see that regardless of the foreground image from which class, CMO mainly mixes it with the background image from the majority classes. For example, the foreground images from the "truck" will be mixed with the background images from the "airplane" class, even if the truck is more similar to the automobile. However, Fig. 9(b) indicates that our proposed method builds more reasonable pairs by selecting the most relevant background image for each foreground image. For example, the "horse" is more easily confused with "deer" than "airplane". These results validate that ours can provide more reasonable generated samples than CMO.

Visualization Results of Mixed Image Pairs To gain a more intuitive insight into the dynamic changes within the mixing process of our method, we provide more visualization results of mixed image pairs on iNaturalist 2018 in Fig. 10. The foreground image and the selected background image commonly have significant semantic similarity. It suggests that OTmix has the capacity to generate reasonably mixed samples for long-tailed classification.



(a) CMO (b) OTmix

Figure 9: Image-mixing statistical results of OTmix and CMO on CIFAR-LT-10 with $\varphi = 100$.

Figure 10: The visualization results of image pairs used to semantically meaningful mixed images on iNaturalist 2018.

7.6 Negative Societal Impacts and Limitations

This work develops a simple and effective image-mixing method for long-tailed learning, which has the potential to encourage researchers to derive new and better methods for the line of mixing images or long-tailed learning. However, if there is a sufficiently malicious or ill-informed choice of a long-tailed classification task or an image-mixing task, it may indirectly lead to a negative impact. Employing an imprecise or incorrect confusion matrix can mislead our method to build unsatisfactory pairs for image-mixing.