Long-Tailed Backdoor Attack Using Dynamic Data Augmentation Operations

Supplementary Material

1. Experimental Details of Backdoor Attacks

We compare our method with other four state-of-the-art stealthy backdoor attacks, including Label-consistent Backdoor Attack (LC), Sinusoidal Signal Backdoor Attack (SIG), Input-aware Backdoor Attack (IAB) and Warp-based Backdoor Attack (WN). LC and SIG are classic clean-label attacks. IAB and WaNet are representative sample-specific attacks. All experiments are conducted under the all-to-one setting. The implementation of our code is based on BackdoorBench [7] V1 [1]. We follow the original paper (or code) of these attacks to conduct experiments, and the details of attack setting are introduced below:

Label-consistent Backdoor Attack (LC). LC is a clean-label backdoor attack, and utilizes a trigger comprising four 3×3 checkerboards positioned at the four corners of an image. We follow the paper [6] to generate adversarial perturbations using projected gradient descent (PGD) [2]. The adversarial model is trained with bounded in l_{inf} norm. For CIFAR10-LT, we follow the paper [6] to set $\epsilon = 16$. For CIFAR100-LT, we set $\epsilon = 8$ to attack model successfully. We poison 50% samples from target label to attack model for CIFAR10-LT. For CIFAR100-LT, we poison 80% samples from target label to successfully attack models.

Sinusoidal Signal Backdoor Attack (SIG). SIG is a clean-label attack with sinusoidal signal as triggers. In the case of CIFAR10-LT, we follow original paper [1] to set $\Delta = 20$ and $f = 6$. For CIFAR100-LT, we set $\Delta = 40$ and $f = 6$ to successfully attack the model. Similar to LC, we poison 50% samples from target label to attack model for CIFAR10-LT, while for CIFAR100-LT, we poison 80% samples from target label to achieve successfully attacks.

Input-aware Backdoor Attack (IAB). IAB is a sample-specific backdoor attack using two auto-encoders, one for generating triggers and the other for producing masks. Following the original code [3], we first train a mask generator and then train the classification model and trigger generator simultaneously. The injection rate ρ is set as 0.1. We follow the paper to set ρ_a as 0.1 and ρ_c as 0.1 for both datasets.

Warp-based Backdoor Attack (WN). WaNet employs elastic warping triggers to poison images. We follow the original code [4] to train backdoored model with $\rho_a = 0.1$ and $\rho_n = 0.2$ for both CIFAR10-LT and CIFAR100-LT. The injection rate ρ is also set as 0.1.

Lotus. We do not include Lotus [2] in the main comparison since it is a partition-conditioned backdoor attack whose trigger activation is restricted to specific subsets of the victim class. In contrast, our work focuses on all-to-one attacks, where a universal trigger is expected to generalize across inputs. Due to this fundamental difference in activation scope, their attack success rates are not directly comparable under a unified evaluation protocol.

[1] https://github.com/SCLBD/BackdoorBench/tree/v1

[2] https://github.com/MadryLab/label-consistent-backdoor-code

[3] https://github.com/VinAIResearch/input-aware-backdoor-attack-release

[4] https://github.com/VinAIResearch/Warping-based_Backdoor_Attack-release

2. Comparison Results Using Logits Adjustments (LA)

We present comparison results on CIFAR10-LT and CIFAR100-LT using Logits Adjustments (LA) in Table 1 and Table 2, respectively. Comparing to the results without using LA in the main paper, we observe that LA can increase ACC largely. This is consistent with the traditional long-tailed visual recognition methods, which usually integrate re-sampling and re-weighting methods to improve the classification accuracy. However, LA sometimes hinders the attack performance (ASR). For example, the average attack performance of IAB using LA decreases. A possible reason is that LA changes the decision boundary between clean images and backdoored images.

3. Comparison Results with ViT

To evaluate the performance of our method with advanced architectures, we conduct experiments on CIFAR10-LT using the Vision Transformer (ViT) [3]. The results of the comparison are presented in Table 3. Our method demonstrates superior attack performance (ASR) compared to other backdoor attacks, while maintaining comparable clean accuracy (ACC). Notably, since we utilize the pre-trained ViT model on ImageNet, the ACC is higher than those obtained with other architectures, such as ResNet18.

4. Results on ImageNet

We also conducts experiments on a large-scale dataset called ImageNet10-LT. The results in Table 4 show that our method outperforms other baselines on complex dataset. Although WN is stealthier than IAB, it struggles with complex datasets.

5. Results with Different Splitting Classes

In addition to the main paper, we report an alternative split on CIFAR10-LT: Many (first 2/10 classes), Medium (next 3/10), and Few (last 5/10). The results are shown in Table 5.

Metric	Attack	Target Label: Many				Target Label: Medium				Target Label: Few				Avg
		Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few	All	
ACC	IAB+LA	88.88	76.74	83.53	83.10	89.17	75.63	82.87	82.59	88.91	76.24	82.67	82.61	82.75
	LC+LA	89.67	78.78	83.83	84.13	91.44	76.33	84.00	83.96	92.08	78.75	82.56	84.21	84.11
	SIG+LA	89.33	79.11	83.67	84.01	91.78	75.00	82.67	83.13	91.58	78.42	80.75	83.30	83.46
	WN+LA	89.66	74.97	79.38	81.14	89.51	76.28	78.54	81.15	89.56	73.93	78.06	80.27	80.79
	Ours+LA	90.72	77.19	80.96	82.76	89.46	77.00	81.70	82.62	88.44	75.44	82.82	82.29	82.53
ASR	IAB+LA	89.15	79.64	64.06	74.83	97.64	93.78	84.71	91.04	**99.57**	98.44	94.50	97.50	88.76
	LC+LA	88.67	88.33	82.25	85.74	90.11	**99.33**	93.25	93.59	29.08	42.08	22.33	31.17	66.27
	SIG+LA	94.67	87.78	90.25	90.47	93.56	93.17	92.17	92.82	81.00	68.75	68.83	72.84	84.12
	WN+LA	93.88	94.38	89.62	92.16	96.59	96.70	**95.67**	96.21	97.75	**98.83**	**97.67**	**98.09**	95.74
	Ours+LA	**97.63**	**95.99**	**95.67**	**96.21**	**98.36**	95.58	94.91	**96.21**	98.72	96.62	93.48	96.28	**96.24**

Table 1. Comparison results using Logits Adjustment (LA) on CIFAR10-LT.

Metric	Attack	Target Label: Many				Target Label: Medium				Target Label: Few				Avg
		Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few	All	
ACC	IAB+LA	65.62	58.33	51.26	58.33	65.71	58.38	51.03	58.30	65.51	58.47	50.54	58.10	58.24
	LC+LA	67.26	60.16	53.15	60.12	68.02	60.16	52.55	60.17	68.48	60.49	53.15	60.63	60.31
	SIG+LA	67.49	59.84	52.74	59.95	66.07	56.92	51.40	58.06	67.71	59.65	51.24	59.45	59.16
	WN+LA	68.02	60.22	52.65	60.22	67.35	59.76	53.40	60.10	67.67	59.43	52.56	59.81	60.04
	Ours+LA	66.22	57.89	50.57	58.15	66.14	58.91	49.65	58.15	65.97	59.25	50.00	58.32	58.21
ASR	IAB+LA	97.95	96.96	96.46	97.11	**98.92**	97.92	97.80	98.21	**99.05**	**98.23**	98.04	98.44	97.93
	LC+LA	42.30	41.31	46.70	43.48	61.91	59.38	64.09	61.84	1.12	1.74	1.71	1.52	35.27
	SIG+LA	94.41	94.33	96.00	94.93	93.58	93.28	95.42	94.11	91.02	92.03	93.13	92.06	93.69
	WN+LA	92.63	93.86	90.50	92.31	93.09	94.34	90.93	92.75	94.91	96.25	92.69	94.62	93.24
	Ours+LA	**98.80**	**98.48**	**98.45**	**98.57**	98.39	**98.27**	**98.38**	**98.35**	98.97	98.15	**98.48**	**98.54**	**98.49**

Table 2. Comparison results using Logits Adjustment (LA) on CIFAR100-LT.

Metric	Attack	Target Label: Many				Target Label: Medium				Target Label: Few				Avg
		Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few	All	
ACC	IAB	98.00	89.00	81.08	88.51	98.11	88.56	80.33	88.07	97.83	89.00	80.50	88.29	88.29
	LC	97.89	89.22	83.67	89.64	98.11	89.44	82.33	89.29	97.83	88.75	83.38	89.39	89.44
	SIG	97.67	89.33	83.33	89.45	98.11	89.00	82.58	89.21	97.75	89.17	82.19	89.02	89.20
	WN	97.22	88.67	82.50	88.72	97.44	89.56	83.00	89.28	97.17	89.17	84.94	89.94	89.38
	Ours	97.57	89.40	79.56	87.91	98.19	88.86	81.88	88.86	98.06	89.02	81.03	88.53	88.45
ASR	IAB	94.50	96.67	95.08	95.46	97.22	98.00	96.92	97.25	95.50	96.50	95.00	95.68	96.08
	LC	9.00	19.00	22.08	18.13	18.22	41.00	31.33	29.08	0.25	0.75	0.33	0.52	14.37
	SIG	56.50	47.11	62.75	56.07	24.33	31.67	36.17	31.22	3.83	2.33	4.58	3.62	27.64
	WN	93.50	92.78	96.75	94.72	94.11	93.67	97.42	95.50	95.42	95.00	97.92	96.07	95.49
	Ours	**97.65**	**97.41**	**98.38**	**97.90**	**98.93**	**98.38**	**98.26**	**98.51**	**99.22**	**98.94**	**99.17**	**99.11**	**98.57**

Table 3. Comparison results with ViT on CIFAR10-LT.

6. Analysis on Other Target labels

In our main paper, we conduct analysis experiments on CIFAR10-LT with target label as 0 ("Many" classes). We extend our analysis with the target label set as 4 ("Medium" classes). These experiments follow a similar structure to those presented in the main paper. Below are the details of our ablation studies:

Data Augmentation Strength q in Backdoored Operation Selector. Similar to the experiments conducted in the main paper, we keep the other hyperparameters fixed. The results are presented in Tab. 6. We can observe that the model gets higher ASRs when q is set as 2 or 4. Compared to models without using data augmentation ($q = 0$), ASRs increase about 2%. This observation underscores the effectiveness of employing weak data augmentation policies, consistent with our findings in the main paper. Since ASRs can be higher for both $q = 2$ and $q = 4$, data augmentation strengths can be different for different target labels. Tuning hyperparameter q for different target labels can get better overall performance.

Softmax Temperature T in Optimization Objective of Network h. We fix other hyperparameters and only change

Metric	ACC				ASR			
Attack	IAB	SIG	WN	Ours	IAB	SIG	WN	Ours
Target Label: Many								
Many	91.33	91.33	86.67	87.33	86.00	82.00	15.00	**100.00**
Med.	70.00	70.67	65.33	75.33	**97.33**	84.67	22.67	96.00
Few	47.00	36.50	36.00	45.50	90.50	87.33	44.00	**99.00**
All	67.20	63.20	60.00	67.00	91.78	84.66	30.44	**98.22**
Target Label: Medium								
Many	91.33	90.67	85.67	90.33	98.67	42.67	28.33	**99.00**
Med.	71.33	68.33	74.67	75.33	**100.00**	54.00	25.00	99.50
Few	40.00	38.75	34.00	47.50	**98.50**	48.00	45.50	97.50
All	64.80	63.20	61.70	68.70	**98.89**	48.22	35.22	98.44
Target Label: Few								
Many	90.00	92.00	88.00	88.67	97.33	0.67	12.67	**99.33**
Med.	66.67	70.00	74.00	71.33	99.33	6.67	15.33	**100.00**
Few	43.00	39.00	36.50	46.00	**99.33**	11.33	25.33	98.00
All	64.20	64.20	63.20	66.40	98.67	6.22	17.78	**99.11**
Avg.	65.28	63.60	61.79	67.27	96.67	46.36	26.81	**98.64**

Table 4. Comparison results on ImageNet10-LT.

Metric	ACC					ASR				
Attack	IAB	LC	SIG	WN	Ours	IAB	LC	SIG	WN	Ours
Target Label: Many										
Many	93.71	95.22	94.33	94.00	94.22	95.37	82.22	90.03	94.33	**98.31**
Med.	77.33	78.70	76.44	73.20	73.22	93.33	89.21	90.66	96.47	**97.99**
Few	66.21	69.30	63.24	64.00	63.23	93.27	89.26	93.44	96.78	**97.43**
All	77.67	80.21	77.12	76.40	76.90	93.20	87.32	91.83	94.89	**97.60**
Target Label: Medium										
Many	93.42	96.11	95.47	94.42	94.88	94.42	91.32	81.00	96.32	**99.20**
Med.	76.21	75.33	71.53	74.32	75.78	95.45	**99.30**	93.54	97.89	98.36
Few	64.78	69.81	62.43	64.25	66.89	92.00	97.44	85.56	97.19	**97.94**
All	76.12	78.32	75.60	76.32	77.58	93.21	93.70	85.89	95.73	**98.60**
Target Label: Few										
Many	94.48	96.89	95.27	93.49	95.09	95.92	29.27	45.43	97.22	**98.28**
Med.	76.29	77.84	75.09	75.03	76.20	95.19	45.82	41.29	**98.37**	97.27
Few	65.73	67.32	58.67	66.29	65.00	94.27	31.20	39.12	**98.11**	97.24
All	75.32	75.40	74.31	75.11	74.67	93.33	34.27	44.55	95.22	**97.66**
Avg.	76.39	78.05	73.95	75.41	76.00	93.31	67.93	69.19	95.90	**97.01**

Table 5. Comparison results on CIFAR10-LT with another splitting on classes.

Strength q	All	Many	Medium	Few
0	95.49 / 76.23	95.80 / 92.03	97.00 / 75.97	94.50 / 64.58
1	93.51 / 77.03	93.27 / 93.23	95.35 / 76.13	92.77 / 65.55
2	97.26 / 75.90	97.43 / 93.07	97.40 / 74.93	97.05 / 63.75
3	94.82 / 75.97	95.03 / 92.80	95.20 / 75.87	94.47 / 63.42
4	97.64 / 74.69	98.10 / 92.33	97.80 / 75.37	97.23 / 60.95

Table 6. Analysis of data augmentation strength q in backdoored operation selector.

temperature T. The results are shown in Tab. 7. From the results, we can see that changing temperature T does not significantly affect average ASRs. This suggests that the choice of T does not play a crucial role in achieving better performance on ASRs when the target label is 4.

T	All	Many	Medium	Few
2	97.82 / 74.86	98.07 / 92.97	97.80 / 75.07	97.65 / 61.13
1	97.26 / 75.90	97.43 / 93.07	97.40 / 74.93	97.05 / 63.75
3	97.89 / 75.77	98.47 / 92.33	98.20 / 72.87	97.30 / 65.52

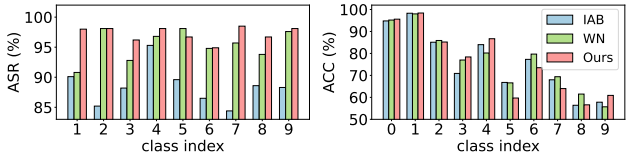
Table 7. Results on CIFAR10-LT when changing temperature T in optimization objective $\mathcal{L}_h$ of backdoored operation selector.

λ_{div}	All	Many	Medium	Few
0.01	97.26 / 75.90	97.43 / 93.07	97.40 / 74.93	97.05 / 63.75
0.05	97.39 / 75.86	97.70 / 92.80	98.70 / 75.73	96.50 / 63.25
0.1	97.59 / 74.37	97.27 / 91.73	98.15 / 75.97	97.55 / 60.15
0.5	89.02 / 76.06	90.30 / 93.73	91.00 / 74.30	87.08 / 64.12

Table 8. Analysis of trigger diversity loss weight λ_{div}.

α	All	Many	Medium	Few
0.01	19.18 / 74.13	12.10 / 92.80	25.80 / 74.33	21.18 / 59.98
0.05	82.28 / 76.38	82.73 / 94.07	85.00 / 75.37	80.58 / 63.87
0.1	97.26 / 75.90	97.43 / 93.07	97.40 / 74.93	97.05 / 63.75
0.15	98.89 / 75.66	98.90 / 93.77	99.20 / 73.47	98.72 / 63.73
0.2	98.14 / 75.98	98.77 / 93.17	98.25 / 74.43	97.62 / 64.25

Table 9. Analysis of trigger strength α.



(a) Class-wise ASR. (b) Class-wise ACC.

Figure 1. Class-wise performance compared to other two attacks when target label is 4.

Weight (λ_{div}) of Trigger Diversity Loss. The results of varying λ_{div} are presented in Tab. 8. It is noticeable that ASR experiences a significant decrease with a larger value of λ_{div}. The reason is that increasing λ_{div} can increase the difficulty of training trigger generator. Essentially, λ_{div} represents the trade-off between the effectiveness and diversity of the trigger generator, which aligns with the observations made in the main paper.

Strength α of Trigger. We conducted experiments with fixed hyperparameters while adjusting α, and the results are summarized in Tab. 9. It is observed that ASRs will increase when increasing trigger strength α. This observation can be attributed to the heightened visibility of the trigger as its strength increases. A stronger trigger is more conspicuous, making it easier for the classifier to discern and learn its features, thus resulting in higher ASRs.

Class-wise Performance. We further analyzed the class-wise performance, comparing our method with two other sample-specific backdoor attacks, namely IAB [5] and WaNet [4]. Class-wise ASR and ACC are depicted in Fig-

ure 1a and Fig. 1b, respectively. From Fig. 1a, it is evident that our method achieves state-of-the-art ASRs for most classes. Additionally, our method demonstrates superior performance on head classes and some tail classes (*e.g.*, class 9) in terms of ACC, as illustrated in Fig. 1b.

7. Negative Impact and Limitations.

We brings attention to the threat of this practical long-tailed backdoor attack task. While our method achieves a good attack performance, we mainly focus on data augmentation-based methods, and does not consider re-weighting-based techniques. We leave them for an interesting future work.

References

[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019. 1

[2] Siyuan Cheng, Guanhong Tao, Yingqi Liu, Guangyu Shen, Shengwei An, Shiwei Feng, Xiangzhe Xu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Lotus: Evasive and resilient backdoor attacks through sub-partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24798–24809, 2024. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021. 3

[5] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020. 3

[6] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 1

[7] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1