Table 1: Comparison of model repair performance on CIFAR-10N and corrupted Tiny-ImageNet datasets. The best performance among all baselines is in bold. The networks for CIFAR-10N and Tiny-ImageNet are ResNet20 and ResNet18, respectively.

	CIFAR-10N			Tiny-ImageNet	
	Worst (40.21%)	Random (17.23%)	Aggregate (9.03%)	Symmetric	Asymmetric
Base Model	77.76	84.68	85.78	48.41	42.84
Linear Infuence Repair	79.30	85.22	87.70	49.15	44.50
SGD Influence Repair	71.50	77.21	76.08	57.91	61.22
EWC Repair	71.80	76.22	77.18	57.99	61.36
RDIA	62.90	62.96	64.62	59.01	62.47
BIR	**85.50**	**88.44**	**89.60**	**59.14**	**63.15**

Table 2: Number of identified cause data on corrupted MNIST and CIFAR-10 datasets using different strategies.

	MNIST-Label	MNIST-Input	MNIST-Adv	CIFAR10-Label	CIFAR10-Input	CIFAR10-Adv
ground-truth amount	432	432	432	7200	5400	7200
threshold = 0	1115	1789	1919	23955	27449	30457
threshold = log(1.1)	469	122	198	8745	5076	6003
threshold = log(1.05)	487	157	251	9952	6362	7686
threshold = log(1.01)	531	268	393	12836	9716	11849

Table 3: The repair performance on corrupted MNIST datasets under different cause set sizes using different identified strategies. The number of cause set size is availabel in Tab. 2. The network is LeNet.

	MNIST-Label			MNIST-Input			MNIST-Adv		
	All	Clean	Noisy	All	Clean	Noisy	All	Clean	Noisy
ground-truth amount	95.60	95.93	95.10	96.16	96.20	96.10	96.44	96.47	96.40
threshold = 0	95.84	96.30	95.15	96.32	96.13	96.60	96.48	96.63	96.25
threshold = log(1.1)	95.64	95.77	95.45	96.22	96.13	96.35	96.40	96.57	96.15
threshold = log(1.05)	95.66	96.20	94.85	96.28	96.07	96.60	96.30	96.43	96.10
threshold = log(1.01)	95.56	96.07	94.8	96.14	95.97	96.40	96.34	96.53	96.05

Table 4: The repair performance on corrupted CIFAR-10 datasets under different cause set sizes using different identified strategies. The number of cause set size is availabel in Tab. 2. The network is ResNet50.

	CIFAR10-Label			CIFAR10-Input			CIFAR10-Adv		
	All	Clean	Noisy	All	Clean	Noisy	All	Clean	Noisy
ground-truth amount	83.02	91.73	69.95	88.64	91.80	83.90	90.26	93.20	85.85
threshold = 0	83.26	91.33	71.15	88.90	92.20	83.55	90.14	93.43	85.20
threshold = log(1.1)	83.00	91.50	70.25	88.50	91.63	83.80	90.20	93.30	85.55
threshold = log(1.05)	83.14	91.83	70.10	88.62	91.70	84.00	90.12	93.03	85.75
threshold = log(1.01)	82.98	91.57	70.10	88.66	92.07	83.55	89.86	93.07	85.05

Table 5: The repair performance on corrupted MNIST datasets with different sizes of preserved clean data. The network is LeNet.

	MNIST-Label				MNIST-Input				MNIST-Adv			
Size of preserved clean data	Identification AUC	ALL	Clean	Noisy	Identification AUC	ALL	Clean	Noisy	Identification AUC	ALL	Clean	Noisy
30	0.76	83.67	95.99	69.70	0.68	96.48	97.07	95.60	0.57	96.36	97.07	95.30
100	0.98	92.82	93.66	91.55	0.70	96.24	96.60	95.70	0.63	96.46	96.90	95.80
200	0.99	95.24	95.83	94.35	0.70	96.34	96.47	96.15	0.68	96.26	96.50	95.90
300	0.99	95.60	95.93	95.10	0.71	96.16	96.20	96.10	0.70	96.44	96.47	96.40

Table 6: Comparison of model repair performance on corrupted MINST dataset with mixture noise (30% of label noise and 20% of adversarial attacks. The best performance among all baselines is in bold.

	All	Clean	Noisy
Base Model	90.10	97.13	79.55
Linear Infuence Repair	91.28	97.00	82.70
SGD Influence Repair	95.78	**97.33**	93.45
EWC Repair	95.96	97.27	94.00
RDIA	85.66	95.87	70.35
BIR	**96.60**	97.10	**95.85**

Table 7: Comparison of model repairing and adversrial training (AT) on MINST-Adv and CIFAR10-Adv dataset with mixture noise (30% of label noise and 20% of adversarial attacks. The best performance among all baselines is in bold.

Dataset	Network	Method	ALL	Clean	Noisy
MNIST-Adv	LeNet	AT	93.68	92.73	95.10
		BIR(Ours)	96.44	96.47	96.40
MNIST-Adv	Conv	AT	92.16	96.20	86.10
		BIR(Ours)	98.12	98.43	97.65
CIFAR10-Adv	ResNet20	AT	83.16	91.10	71.25
		BIR(Ours)	87.94	92.17	81.60
CIFAR10-Adv	Resnet50	AT	84.52	92.13	73.10
		BIR(Ours)	90.26	93.20	85.85