CIARD: Cyclic Iterative Adversarial Robustness Distillation

Supplementary Material

In the supplementary materials, we provide an extensive elaboration on additional experiments and results for the proposed CIARD. On one hand (Section A), we present a more systematic description of CIARD's complete algorithmic workflow 1. On the other hand (Section B), we have expanded our experimental scope by incorporating a wider range of models and attack methods for testing, thereby offering more comprehensive experimental results.

A. Cyclic Iterative ARD

This section provides a comprehensive description of the Cyclic Iterative Adversarial Robustness Distillation (CIARD) algorithm, including its adversarial example generation pipeline, multi-teacher knowledge transfer strategy, and dynamic optimization mechanisms. The detailed algorithm description of CIARD is outlined in Algorithm 1.

B. Supplementary Experiments

Training Dynamics Analysis. Figure 1 illustrates the evolution of both clean accuracy and robust accuracy throughout the training process. Unlike traditional ARD methods that typically show sharp trade-offs between these metrics, CIARD demonstrates a more harmonious progression where robust accuracy steadily improves without substantially sacrificing clean performance. Notably, after epoch 50 when the robust teacher begins updating, we observe an accelerated improvement in the student's adversarial robustness.

In addition, Figure 2 shows the ablations study that only includes the learnable teacher and does not include the pushing loss component. The results clearly show that without the pushing loss, the robustness of our clean teacher decreases, and the robustness accuracy of the student model cannot be significantly improved. This highlights the crucial role of the pushing loss mechanism in our framework, which helps maintain the performance of the clean teacher while transferring effective robustness to the student model.

Computational Efficiency Analysis. While our Iterative Teacher Training (ITT) introduces additional computation, the overhead is moderate: when training the results in Table 3 (main paper) using WRN-34-10 as the robust teacher on a single RTX 4090 GPU, the per-epoch time increases from 111.06s (w/o ITT) to 140.56s (w/ ITT). This overhead largely depends on the teacher model size and remains relatively minor for lightweight architectures. We believe the trade-off is acceptable given the consistent robustness gains.

Complete Results on CIFAR-10 & CIFAR-100. In Tables 1 and 2 of the supplementary material, we present comprehensive evaluation results of our proposed CIARD method compared with state-of-the-art adversarial robustness approaches on the CIFAR-10 and CIFAR-100 datasets. We conduct experiments using ResNet-18 and MobileNet-V2 architectures against four different attack methods: FGSM, PGD$_{\text{SAT}}$, PGD$_{\text{TRADES}}$, and CW$_\infty$.

For the CIFAR-10 dataset, CIARD consistently outperforms all baseline methods across both architectures. With ResNet-18, CIARD achieves the highest clean accuracy

Algorithm 1: Cyclic Iterative ARD (CIARD)

```
Input: Clean teacher model $T_{\text{nat}}$, robust teacher model $T_{\text{adv}}$, student model $S(x|\theta_s)$, clean images $x$ and labels $y$, perturbation bound $\Omega$, training epochs $T$, temperature $\tau$
Input: weights $w_{\text{nat}} = 0.5$, $w_{\text{adv}} = 0.5$, learning rate $\alpha = 0.1$, teacher learning rate $\alpha_t = 0.01$, weight learning rate $\eta = 0.025$
for $epoch = 1$ to $T$ do
    for each mini-batch $(x, y)$ do
        // * Adversarial Example Generation * //
        Generate adversarial examples via PGD:
            $x^* = \arg\max_{\delta\in\Omega} \text{CE}(S(x+\delta), y)$;
        // * Compute $\mathcal{L}_{\text{student}}$ * //
        i) Clean knowledge transfer:
            $\mathcal{L}_{\text{nat}} = \text{KL}(S(x), T_{\text{nat}}(x))$;
        ii) Robust knowledge transfer:
            $\mathcal{L}_{\text{adv}} = \text{KL}(S(x^*), T_{\text{adv}}(x^*))$;
        iii) Push loss for robust specialization:
            $\mathcal{L}_{\text{push}} = \text{Push}(S(x^*), T_{\text{nat}}(x^*))$;
        iv) Adaptive weight update:
            $\hat{\mathcal{L}}_{\text{nat}} = \mathcal{L}_{\text{nat}} / \mathcal{L}_{\text{nat}}^{\text{init}}$;
            $\hat{\mathcal{L}}_{\text{adv}} = \mathcal{L}_{\text{adv}} / \mathcal{L}_{\text{adv}}^{\text{init}}$;
            $w_{\text{nat}} = w_{\text{nat}} - \eta(w_{\text{nat}} - \frac{\hat{\mathcal{L}}_{\text{nat}}}{\hat{\mathcal{L}}_{\text{nat}} + \hat{\mathcal{L}}_{\text{adv}}})$;
            $w_{\text{adv}} = 1 - w_{\text{nat}}$;
        v) Total student loss:
            $\mathcal{L}_{\text{student}} = w_{\text{adv}}\mathcal{L}_{\text{adv}} + w_{\text{nat}}\mathcal{L}_{\text{nat}} - \lambda\mathcal{L}_{\text{push}}$;
        // * Compute $\mathcal{L}_{\text{adv\_teacher}}$ * //
        $\mathcal{L}_{\text{adv\_teacher}} = \text{CE}(T_{\text{adv}}(x^*), y)$;
        Update student model $S$ using $\nabla_{\theta_S}\mathcal{L}_{\text{student}}$;
        if $epoch > 50$ then
            Update $T_{\text{adv}}$ using $\nabla_{\theta_T}\mathcal{L}_{\text{adv\_teacher}}$;
        end
    end
end
```

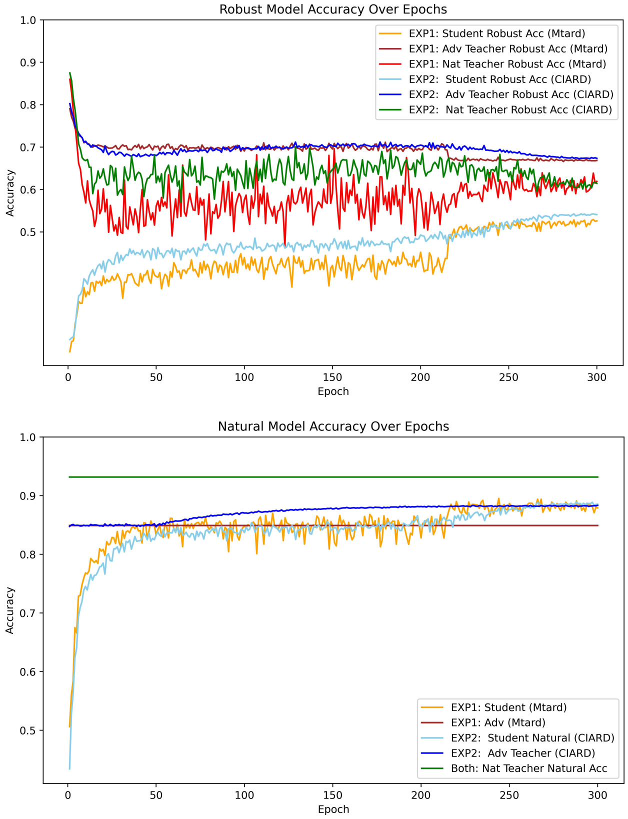

Figure 1. MTARD vs. CIARD.

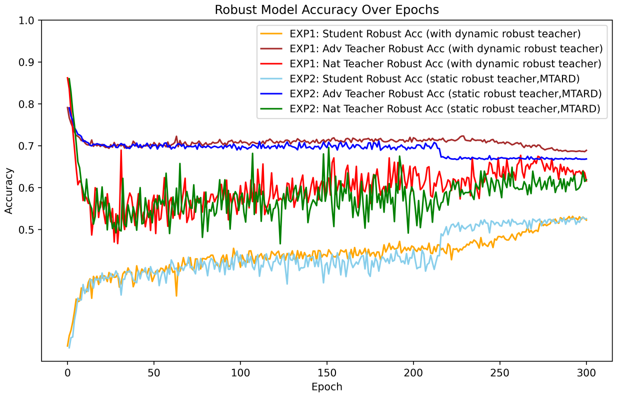

Figure 2. Robust Model Accuracy Over Epochs.

(88.87%) while maintaining superior robust accuracy under various attacks. For example, under FGSM attack, CIARD achieves 61.88% robust accuracy and 75.38% weighted robust accuracy, surpassing both single-teacher methods like RSLAD (60.41%, 72.20%) and dual-teacher approaches like B-MTARD (61.42%, 74.81%).

Similar performance advantages are observed with MobileNet-V2, where CIARD achieves 89.51% clean accuracy and demonstrates consistent improvements in both robust and weighted robust metrics across all attack types. For instance, under FGSM attack, CIARD achieves 59.10% robust accuracy and 74.31% weighted robust accuracy, outperforming B-MTARD (58.79%, 73.94%).

For the CIFAR-100 dataset, CIARD also demonstrates superior performance. With ResNet-18 under FGSM attack, CIARD achieves 65.73% clean accuracy, 34.47% robust accuracy, and 50.10% weighted robust accuracy, exceeding B-MTARD's performance (65.08%, 34.21%, 49.65%). These results consistently hold across different attack types, with CIARD maintaining its advantage in the more challenging CIFAR-100 dataset.

The comprehensive experimental results confirm that our cyclic iterative approach effectively enhances adversarial robustness while preserving high clean accuracy across different model architectures and datasets, successfully addressing the accuracy-robustness trade-off challenge.

Results on Tiny-ImageNet Dataset. To further evaluate the scalability of our proposed method to more complex datasets, we conduct extensive experiments on the Tiny-ImageNet dataset. Table 3 presents the white-box robustness results using both PreActResNet-18 (RN-18) and MobileNet-V2 (MN-V2) architectures, while Table 4 displays the black-box robustness results against various attacks.

On the Tiny-ImageNet dataset, CIARD achieves state-of-the-art performance with 57.42% clean accuracy and 28.26% robust accuracy under FGSM attacks for PreActResNet-18, surpassing B-MTARD by 0.61% and 0.14% respectively. Similar improvements are observed with MobileNet-V2, where CIARD achieves 53.05% clean accuracy and 25.45% robust accuracy, outperforming B-MTARD (52.98%, 25.60%).

For black-box attacks, CIARD demonstrates even more pronounced advantages. Under PGDtrades attack, CIARD achieves 36.80% robust accuracy and 47.11% weighted robust accuracy with PreActResNet-18, exceeding B-MTARD's performance (36.65%, 46.73%). Against the more challenging Square Attack (SA), CIARD maintains superior performance with 44.48% robust accuracy and 50.95% weighted robust accuracy, compared to B-MTARD's 44.46% and 50.64% respectively.

These comprehensive results on the more complex Tiny-ImageNet dataset further confirm the effectiveness and scalability of our approach across different network architectures and attack types, demonstrating CIARD's ability to maintain the balance between clean accuracy and adversarial robustness even on more challenging visual recognition tasks.

Robustness Evaluation against AutoAttack. To provide a more comprehensive and rigorous evaluation of adversarial robustness, we tested our proposed method against **AutoAttack** [3], which is widely considered a state-of-the-art and parameter-free ensemble of attacks. This evaluation provides a reliable estimate of a model's worst-case robustness under strong, adaptive threat models, thereby addressing the need for evaluation beyond standard PGD-based attacks. The eval-

Table 1. White-box Adversarial Robustness of ResNet-18 on CIFAR-10 and CIFAR-100 Datasets. The best results are **bolded**, and the second best results are underlined.

Student Model	Attack	Type	Defense	CIFAR-10			CIFAR-100		
				Clean	Robust	W-Robust	Clean	Robust	W-Robust
ResNet-18	FGSM [5]	Single-Teacher	SAT [6]	84.20	55.59	69.90	56.16	25.88	41.02
			TRADES [9]	83.00	58.35	70.68	57.75	31.36	44.56
			ARD [4]	84.11	58.40	71.26	60.11	33.61	46.86
			RSLAD [13]	83.99	60.41	72.20	58.25	34.73	46.49
			SCORE [7]	84.43	59.84	72.14	56.40	32.94	44.67
			Fair-ARD [8]	83.41	58.91	71.16	57.81	34.39	46.10
			ABSLD [11]	83.21	60.22	71.72	56.77	**34.94**	45.86
		Dual-Teacher	MTARD [10]	87.36	61.20	74.28	64.30	31.49	47.90
			B-MTARD [12]	88.20	61.42	74.81	65.08	34.21	49.65
			CIARD	**88.87**	**61.88**	**75.38**	**65.73**	34.47	**50.10**
ResNet-18	PGD_{sat} [6]	Single-Teacher	SAT [6]	84.20	45.85	65.08	56.16	21.18	38.67
			TRADES [9]	83.00	52.35	67.68	57.75	28.05	42.90
			ARD [4]	84.11	50.93	67.52	60.11	29.40	44.76
			RSLAD [13]	83.99	53.94	68.97	58.25	31.19	44.72
			SCORE [7]	84.43	53.72	69.08	56.40	30.27	43.34
			Fair-ARD [8]	83.41	52.00	67.71	57.81	30.64	44.23
			ABSLD [11]	83.21	**54.63**	68.92	56.77	**32.41**	44.59
		Dual-Teacher	MTARD [10]	87.36	50.83	69.05	64.30	24.95	44.63
			B-MTARD [12]	88.20	51.68	69.94	65.08	28.50	46.79
			CIARD	**88.87**	51.70	**70.29**	**65.73**	28.05	**46.89**
ResNet-18	$\text{PGD}_{\text{trades}}$ [9]	Single-Teacher	SAT [6]	84.20	48.12	66.16	56.16	22.02	39.09
			TRADES [9]	83.00	53.83	68.42	57.75	28.88	43.32
			ARD [4]	84.11	52.96	68.54	60.11	30.51	45.31
			RSLAD [13]	83.99	55.73	69.86	58.25	32.05	45.15
			SCORE [7]	84.43	55.21	69.82	56.40	30.56	43.48
			Fair-ARD [8]	83.41	53.77	68.59	57.81	31.50	44.66
			ABSLD [11]	83.21	**56.10**	69.66	56.77	**32.99**	44.88
		Dual-Teacher	MTARD [10]	87.36	53.60	70.48	64.30	26.75	45.53
			B-MTARD [12]	88.20	54.40	71.30	65.08	29.94	47.51
			CIARD	**88.87**	54.46	**71.67**	**65.73**	29.45	**47.59**
ResNet-18	CW_{∞} [2]	Single-Teacher	SAT [6]	84.20	45.97	65.09	56.16	20.90	38.53
			TRADES [9]	83.00	50.23	66.62	57.75	24.19	40.97
			ARD [4]	84.11	50.15	67.13	60.11	27.56	43.84
			RSLAD [13]	83.99	**52.67**	68.33	58.25	**28.21**	43.23
			SCORE [7]	84.43	50.46	67.45	56.40	26.30	41.35
			Fair-ARD [8]	83.41	51.07	67.24	57.81	27.84	42.83
			ABSLD [11]	83.21	52.04	67.63	56.77	26.99	41.88
		Dual-Teacher	MTARD [10]	87.36	48.57	67.97	64.30	23.42	43.86
			B-MTARD [12]	88.20	49.88	69.04	65.08	25.45	**45.27**
			CIARD	**88.87**	50.61	**69.74**	**65.73**	24.43	45.08

uation was conducted on the CIFAR-10 dataset using both ResNet-18 and MobileNetV2 as student architectures. The results, including clean accuracy, robust accuracy, and the holistic Weighted Robustness (W-R) metric, are presented in Table 5.

The results clearly demonstrate the superiority of our proposed CIARD framework. It achieves the highest Weighted Robustness (W-R) on both architectures, reaching **68.88% on ResNet-18** and **67.61% on MobileNetV2**. This state-of-the-art overall performance highlights its exceptional ability to balance high accuracy on clean samples with strong defense against adversarial attacks. Compared to the strong B-MTARD baseline, CIARD yields a significant W-R improvement of **+1.06%** on ResNet-18 and **+0.77%** on Mo-

Table 2. White-box Adversarial Robustness of MobileNet-V2 on CIFAR-10 and CIFAR-100 Datasets. The best results are **bolded**, and the second best results are underlined.

Student Model	Attack	Type	Defense	CIFAR-10			CIFAR-100		
				Clean	Robust	W-Robust	Clean	Robust	W-Robust
MobileNet-V2	FGSM [5]	Single-Teacher	SAT [6]	83.87	55.89	69.88	59.19	30.88	45.04
			TRADES [9]	77.95	53.75	65.85	55.41	30.28	42.85
			ARD [4]	83.43	57.03	70.23	60.45	32.77	46.61
			RSLAD [13]	83.20	**59.47**	71.34	59.01	33.88	46.45
			SCORE [7]	82.32	58.43	70.38	49.38	29.28	39.33
			Fair-ARD [8]	82.65	56.37	69.51	59.18	34.07	46.63
			ABSLD [11]	82.50	58.47	70.49	56.67	33.85	45.26
		Dual-Teacher	MTARD [10]	89.26	57.84	73.55	**67.01**	32.42	49.72
			B-MTARD [12]	89.09	58.79	73.94	66.13	**34.36**	**50.25**
			CIARD	**89.51**	59.10	**74.31**	66.72	33.56	50.14
MobileNet-V2	PGD$_{\text{sat}}$ [6]	Single-Teacher	SAT [6]	83.87	46.84	65.36	59.19	25.64	42.42
			TRADES [9]	77.95	49.06	63.51	55.41	23.33	39.37
			ARD [4]	83.43	49.50	66.47	60.45	28.69	44.57
			RSLAD [13]	83.20	53.25	68.23	59.01	30.19	44.60
			SCORE [7]	82.32	**53.42**	67.87	49.38	27.03	38.21
			Fair-ARD [8]	82.65	50.50	66.58	59.18	30.15	44.67
			ABSLD [11]	82.50	52.98	67.74	56.67	**31.28**	43.98
		Dual-Teacher	MTARD [10]	89.26	44.16	66.71	**67.01**	25.14	46.08
			B-MTARD [12]	89.09	47.56	68.33	66.13	28.47	**47.30**
			CIARD	**89.51**	47.67	**68.59**	66.72	27.02	46.87
MobileNet-V2	PGD$_{\text{trades}}$ [9]	Single-Teacher	SAT [6]	83.87	49.14	66.51	59.19	26.96	43.08
			TRADES [9]	77.95	50.27	64.11	55.41	28.42	41.92
			ARD [4]	83.43	51.70	67.57	60.45	29.63	45.04
			RSLAD [13]	83.20	**54.76**	68.98	59.01	31.19	45.10
			SCORE [7]	82.32	54.46	68.39	49.38	27.53	38.46
			Fair-ARD [8]	82.65	52.12	67.39	59.18	31.26	45.22
			ABSLD [11]	82.50	54.49	68.50	56.67	**31.90**	44.29
		Dual-Teacher	MTARD [10]	89.26	47.99	68.63	**67.01**	27.10	47.06
			B-MTARD [12]	89.09	50.44	69.77	66.13	29.82	**47.98**
			CIARD	**89.51**	50.71	**70.11**	66.72	28.95	47.84
MobileNet-V2	CW$_{\infty}$ [2]	Single-Teacher	SAT [6]	83.87	46.62	65.25	59.19	25.01	42.10
			TRADES [9]	77.95	46.06	62.01	55.41	27.72	41.57
			ARD [4]	83.43	48.96	66.20	60.45	26.55	43.50
			RSLAD [13]	83.20	**51.78**	67.49	59.01	**27.98**	43.50
			SCORE [7]	82.32	49.18	65.75	49.38	23.29	36.34
			Fair-ARD [8]	82.65	51.07	66.86	59.18	27.55	43.37
			ABSLD [11]	82.50	50.20	66.35	56.67	26.40	41.54
		Dual-Teacher	MTARD [10]	89.26	43.42	66.34	**67.01**	24.14	45.58
			B-MTARD [12]	89.09	46.81	67.95	66.13	26.50	**46.32**
			CIARD	**89.51**	46.88	**68.20**	66.72	25.54	46.13

bileNetV2. It is worth noting that while a specialized method like RSLAD achieves the highest raw robust accuracy, it does so at the cost of a considerable drop in clean accuracy. In contrast, CIARD maintains a high clean accuracy (e.g., 88.87% on ResNet-18) while delivering competitive robustness. This confirms that our proposed contrastive push loss and iterative teacher training are highly effective at mitigating the common accuracy-robustness trade-off. In summary, the strong performance under the demanding AutoAttack benchmark further validates the effectiveness of CIARD in producing lightweight models that are both robust and accurate, making them more reliable for real-world deployment.

Ablation Study on the Push Loss Weight λ. Our proposed CIARD framework introduces a key hyperparameter,

Table 3. The white-box robustness of the Tiny-ImageNet dataset is tested using PreActResNet-18 (RN-18) and MobileNet-V2 (MN-V2), respectively.

Attack	Defense	Tiny-ImageNet(RN-18)			Tiny-ImageNet(MN-V2)		
		Clean	Robust	W-Robust	Clean	Robust	W-Robust
FGSM	SAT	50.08	25.35	37.72	49.03	23.38	36.21
	TRADES	48.45	23.59	36.02	43.81	20.10	31.96
	ARD	53.22	27.97	40.60	45.53	22.88	33.21
	RSLAD	48.78	27.26	38.02	45.69	24.09	34.89
	SCORE	10.05	7.80	8.93	28.27	17.47	22.87
	Fair-ARD	46.64	25.81	36.23	47.24	25.31	36.28
	MTARD	52.98	26.41	39.70	50.50	23.94	37.22
	B-MTARD	56.81	28.12	42.47	52.98	**25.60**	**39.29**
	CIARD	**57.42**	**28.26**	**42.84**	**53.05**	25.45	39.25
PGD_{sat}	SAT	50.08	22.24	36.16	49.03	20.31	34.67
	TRADES	48.45	21.59	35.02	43.81	18.16	30.99
	ARD	53.22	**24.92**	39.07	45.53	20.43	32.98
	RSLAD	48.78	25.00	36.89	45.69	22.30	34.00
	SCORE	10.05	7.65	8.85	28.27	16.48	22.38
	Fair-ARD	46.64	23.91	35.28	47.24	**23.37**	35.31
	MTARD	52.98	22.55	37.77	50.50	20.45	35.48
	B-MTARD	56.81	23.93	40.37	52.98	21.58	**37.28**
	CIARD	**57.42**	23.95	**40.69**	**53.05**	21.38	37.22
$\text{PGD}_{\text{trades}}$	SAT	50.08	23.05	36.57	49.03	21.15	35.09
	TRADES	48.45	22.09	35.27	43.81	18.36	31.09
	ARD	53.22	**25.71**	39.47	45.53	21.00	33.27
	RSLAD	48.78	25.45	37.12	45.69	22.74	34.22
	SCORE	10.05	7.67	8.86	28.27	16.69	22.48
	Fair-ARD	46.64	24.29	35.47	47.24	**23.77**	35.51
	MTARD	52.98	23.41	38.20	50.50	21.20	35.85
	B-MTARD	56.81	24.94	40.88	52.98	22.58	**37.78**
	CIARD	**57.42**	24.89	**41.16**	**53.05**	22.42	37.74
CW_{∞}	SAT	50.08	20.48	35.28	49.03	18.69	33.86
	TRADES	48.45	17.33	32.89	43.81	13.47	28.66
	ARD	53.22	**21.41**	37.32	45.53	16.81	31.17
	RSLAD	48.78	20.87	34.83	45.69	18.63	32.16
	SCORE	10.05	6.19	8.13	28.27	13.25	20.76
	Fair-ARD	46.64	19.59	33.12	47.24	**20.04**	33.64
	MTARD	52.98	19.36	36.17	50.50	17.45	33.98
	B-MTARD	56.81	19.69	38.25	52.98	18.08	35.53
	CIARD	**57.42**	19.56	**38.49**	**53.05**	18.20	**35.63**

Table 4. Black-box Adversarial Robustness of MobileNet-V2 on CIFAR-10 and CIFAR-100 Datasets.

Attack	Defense	Tiny-ImageNet(RN-18)			Tiny-ImageNet(MN-V2)		
		Clean	Robust	W-Robust	Clean	Robust	W-Robust
$\text{PGD}_{\text{trades}}$	SAT	50.08	33.40	41.74	49.03	33.47	41.25
	TRADES	48.45	31.01	39.73	43.81	28.35	36.08
	ARD	53.22	34.74	43.98	45.53	30.73	38.13
	RSLAD	48.78	32.85	40.82	45.69	31.20	38.45
	SCORE	10.05	8.74	9.40	28.27	21.82	25.05
	Fair-ARD	46.64	31.58	39.11	47.24	31.80	39.52
	ABSLD	47.21	31.84	39.53	48.08	32.89	40.49
	MTARD	52.98	34.48	43.73	50.50	32.75	41.63
	B-MTARD	56.81	36.65	46.73	52.98	34.25	43.62
	CIARD	**57.42**	**36.80**	**47.11**	**53.05**	**34.50**	**43.78**
CW_{∞}	SAT	50.08	33.20	41.63	49.03	33.13	41.08
	TRADES	48.45	30.72	39.59	43.81	28.64	36.23
	ARD	53.22	33.32	43.27	45.53	30.23	37.88
	RSLAD	48.78	32.09	40.44	45.69	31.10	38.40
	SCORE	10.05	8.82	9.44	28.27	22.19	25.23
	Fair-ARD	46.64	31.38	39.01	47.24	31.40	39.32
	ABSLD	47.21	31.66	39.44	48.08	32.43	40.26
	MTARD	52.98	**33.80**	43.39	50.50	32.05	41.28
	B-MTARD	56.81	33.40	45.11	52.98	**33.50**	43.24
	CIARD	**57.42**	33.69	**45.56**	**53.05**	33.45	**43.25**
SA [1]	SAT	50.08	38.72	44.40	49.03	37.95	43.49
	TRADES	48.45	36.58	42.52	43.81	32.39	38.10
	ARD	53.22	42.58	47.90	45.53	34.60	40.07
	RSLAD	48.78	37.64	43.21	45.69	35.18	40.44
	SCORE	10.05	8.67	9.36	28.27	22.16	25.22
	Fair-ARD	46.64	35.81	41.23	47.24	36.53	41.89
	ABSLD	47.21	36.77	41.99	48.08	38.18	43.13
	MTARD	52.98	41.70	47.34	50.50	38.88	44.69
	B-MTARD	56.81	44.46	50.64	52.98	**40.62**	**46.80**
	CIARD	**57.42**	**44.48**	**50.95**	**53.05**	39.96	46.51

λ, which controls the weight of the contrastive push loss term responsible for robust specialization. To analyze the sensitivity of our method to this parameter, we conducted a dedicated ablation study. In our main experiments, we set $\lambda = 1.0$ by default. For this analysis, to isolate the effect of λ, we kept the other primary loss weights fixed at $\alpha = 1$ and $\beta = 1$. The study was performed using the ResNet-18 student model on the CIFAR-10 dataset.

The results, presented in Table 6, show the model's performance under various white-box attacks and the comprehensive AutoAttack suite as λ is varied.

The findings reveal a clear and expected trade-off. As λ increases, a greater emphasis is placed on pushing the student's predictions away from the clean teacher's vulnerabilities, which generally leads to improved adversarial robustness. This is evidenced by the rising robust accuracy against most attacks, particularly the strong AutoAttack benchmark (from 48.48% to 49.25%). This gain in robustness is accompanied by a slight and graceful degradation in clean accuracy (from 89.06% to 88.60%). Our chosen default value of $\lambda = 1.0$ strikes an effective balance, achieving strong performance across both clean and adversarial examples. The relative stability of the results across the tested range also indicates that our method is not overly sensitive to this hyperparameter.

C. Future Research Directions

Scaling of multi-modal data. We can extend CIARD to multi-modal data, such as images, text and audio, to enhance the robustness of the model across different data modalities. The fusion and processing of multi-modal data is crucial for

Table 5. Performance evaluation against AutoAttack on CIFAR-10. We report clean accuracy (Clean), robust accuracy (Robust), and Weighted Robustness (W-R). The **best** results are in bold, and the second-best are underlined. Our CIARD method achieves the highest W-R on both architectures.

Defense Method	**ResNet-18**			**MobileNetV2**		
	Clean (%)	Robust (%)	W-R (%)	Clean (%)	Robust (%)	W-R (%)
RSLAD	83.99	**50.98**	67.49	83.20	**50.23**	66.72
Fair-ARD	83.41	49.21	66.31	82.65	47.68	65.17
ABSLD	83.21	50.60	66.91	82.50	48.65	65.58
MTARD	87.36	46.18	66.77	**89.26**	41.02	65.14
B-MTARD	88.20	47.44	67.82	89.09	44.58	66.84
CIARD (Ours)	**88.87**	48.88	**68.88**	88.90	46.31	**67.61**

Table 6. Ablation study on the push loss weight λ. Experiments were run with ResNet-18 on CIFAR-10. We report clean accuracy and robust accuracy under multiple attack scenarios. The setting used in our main experiments ($\lambda = 1.0$) is highlighted.

Weight λ	**Clean (%)**	**PGD-T (%)**	**PGD-S (%)**	**FGSM (%)**	**CW$_\infty$ (%)**	**AutoAttack (%)**
0.8	89.06	53.85	51.21	61.12	50.31	48.48
1.0	**88.87**	**54.46**	**51.70**	**61.88**	**50.61**	**48.88**
1.2	88.60	54.55	51.98	61.81	50.92	49.25

improving model robustness. Future research could focus on designing a unified framework that enables CIARD to process multiple data types and utilise the complementary information between these modalities to improve robustness and accuracy. This approach will significantly expand the applicability and effectiveness of CIARD in a variety of realistic scenarios.

Cross-domain applications. We can try to apply CIARD to various domains such as healthcare, finance, and autonomous driving to verify its versatility and usefulness. Different domains have unique data characteristics and application requirements, and future research could explore how CIARD can be adapted and optimised for these specific scenarios. By doing so, the effectiveness and value of CIARD can be validated in real-world applications, thus ensuring its wider applicability and impact in different domains.

Integration with mainstream architectures. In this paper, we focus on DNN models because they have established benchmarks and are widely used in current applications. However, we also recognize the importance of evaluating our approach on newer architectures. In our current work, we not only focus on the latest models like ViT and Swin-Transformer, but also combine the methods of this paper with LLMs to carry out NLP tasks as a way to improve the robustness of the target model.

Through these research directions, the performance and application scope of CIARD can be further improved, laying the foundation for building more robust and efficient edge intelligence systems.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501. Springer, 2020. 5

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 3, 4

[3] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 2

[4] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3996–4003, 2020. 3, 4

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3, 4

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3, 4

[7] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, pages 17258–17277. PMLR, 2022. 3, 4

[8] Xinli Yue, Mou Ningping, Qian Wang, and Lingchen Zhao. Revisiting adversarial robustness distillation from the perspective of robust fairness. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4

[9] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled

trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 3, 4

[10] Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*, pages 585–602. Springer, 2022. 3, 4

[11] Shiji Zhao, Xizhe Wang, and Xingxing Wei. Improving adversarial robust fairness via anti-bias soft label distillation. *arXiv preprint arXiv:2312.05508*, 2023. 3, 4

[12] Shiji Zhao, Xizhe Wang, and Xingxing Wei. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–14, 2024. 3, 4

[13] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16443–16452, 2021. 3, 4