Appendices

A. Theoretical Results

In this section, we provide the proofs of Theorems 3.1 and 3.2.

Lemma A.1. *([9]) Let*

$$z(\boldsymbol{x}) \in \underset{\boldsymbol{x}' \in \mathcal{B}_p(\boldsymbol{x}, \varepsilon)}{\operatorname{argmax}} \mathbb{1}\{F_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq F_{\boldsymbol{\theta}}(\boldsymbol{x}')\}.$$

Then,

$$\begin{aligned} &\mathbb{1}\left\{\exists \mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon) : F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(\mathbf{X}'), F_{\boldsymbol{\theta}}(\mathbf{X}') \neq Y\right\} \\ &\quad \leq \mathbb{1}\left\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X})), Y \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\right\}. \end{aligned} \tag{A.1}$$

The equality holds when $\mathcal{Y} = \{-1, 1\}$.

Theorem 3.1. *For a given score function* $f_{\boldsymbol{\theta}}$*, let*

$$z(\boldsymbol{x}) \in \underset{\boldsymbol{x}' \in \mathcal{B}_p(\boldsymbol{x}, \varepsilon)}{\operatorname{argmax}} \mathbb{1}\left\{F_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq F_{\boldsymbol{\theta}}(\boldsymbol{x}')\right\}.$$

Then, we have

$$\begin{aligned} \mathcal{R}_{rob}(\boldsymbol{\theta}) \leq\ & \mathbb{E}_{(\mathbf{X},Y)} \mathbb{1}\{Y \neq F_{\boldsymbol{\theta}}(\mathbf{X})\} \\ & + \mathbb{E}_{\mathbf{X}}\left\{\mathbb{1}\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\} \cdot p(Y \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))|\mathbf{X})\right\}. \end{aligned} \tag{6}$$

Proof. Note that $\mathcal{R}_{\text{rob}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{nat}}(\boldsymbol{\theta}) + \mathcal{R}_{\text{bdy}}(\boldsymbol{\theta})$ where $\mathcal{R}_{\text{nat}}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{X},Y)} \mathbb{1}\left\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq Y\right\}$ and $\mathcal{R}_{\text{bdy}}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{X},Y)} \mathbb{1}\left\{\exists \mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon) : F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(\mathbf{X}'), F_{\boldsymbol{\theta}}(\mathbf{X}) = Y\right\}$.

Since

$$\begin{aligned} \mathcal{R}_{\text{bdy}}(\boldsymbol{\theta}) &= \mathbb{E}_{(\mathbf{X},Y)} \mathbb{1}\left\{\exists \mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon) : F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(\mathbf{X}'), F_{\boldsymbol{\theta}}(\mathbf{X}) = Y\right\} \\ &\leq \mathbb{E}_{(\mathbf{X},Y)} \mathbb{1}\left\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X})), Y \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\right\} (\because \text{ Lemma A.1}) \\ &= \mathbb{E}_{\mathbf{X}} \mathbb{1}\left\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\right\} \mathbb{E}_{Y|\mathbf{X}} \mathbb{1}\left\{Y \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\right\} \\ &= \mathbb{E}_{\mathbf{X}}\left\{\mathbb{1}\left\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\right\} \cdot p(Y \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))|\mathbf{X})\right\}, \end{aligned}$$

the inequality (6) holds. □

Theorem 3.2. *For a given score function* $f_{\boldsymbol{\theta}}$*, let*

$$z(\boldsymbol{x}) \in \underset{\boldsymbol{x}' \in \mathcal{B}_p(\boldsymbol{x}, \varepsilon)}{\operatorname{argmax}} \mathbb{1}\{F_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq F_{\boldsymbol{\theta}}(\boldsymbol{x}')\}.$$

Then, we have

$$\begin{aligned} \mathcal{R}_{rob}(\boldsymbol{\theta}) \leq\ & \mathbb{E}_{(\mathbf{X},Y)} \mathbb{1}\{Y \neq F_{\boldsymbol{\theta}}(\mathbf{X})\} \\ & + \mathbb{E}_{\mathbf{X}}\left\{\mathbb{1}\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\} \cdot p(Y = F_{\boldsymbol{\theta}}(\mathbf{X})|\mathbf{X})\right\}. \end{aligned} \tag{7}$$

Proof. It suffices to show that $\mathcal{R}_{\text{bdy}}(\boldsymbol{\theta}) \leq \mathbb{E}_{\mathbf{X}}\left\{\mathbb{1}\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}'))\} \cdot p(\mathbf{Y} = F_{\boldsymbol{\theta}}(\mathbf{X})|\mathbf{X})\right\}$.

Since

$$\begin{aligned} \mathcal{R}_{\text{bdy}}(\boldsymbol{\theta}) &= \mathbb{E}_{(\mathbf{X},\mathbf{Y})} \mathbb{1}\left\{\exists \mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon) : F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(\mathbf{X}'), F_{\boldsymbol{\theta}}(\mathbf{X}) = \mathbf{Y}\right\} \\ &= \mathbb{E}_{(\mathbf{X},\mathbf{Y})} \mathbb{1}\left\{\exists \mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon) : F_{\boldsymbol{\theta}}(\mathbf{X}') \neq F_{\boldsymbol{\theta}}(\mathbf{X})\right\} \mathbb{1}\left\{\mathbf{Y} = F_{\boldsymbol{\theta}}(\mathbf{X})\right\} \\ &= \mathbb{E}_{(\mathbf{X},\mathbf{Y})} \mathbb{1}\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\} \mathbb{1}\left\{\mathbf{Y} = F_{\boldsymbol{\theta}}(\mathbf{X})\right\} \\ &= \mathbb{E}_{\mathbf{X}} \mathbb{1}\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\} \cdot \mathbb{E}_{\mathbf{Y}|\mathbf{X}} \mathbb{1}\left\{\mathbf{Y} = F_{\boldsymbol{\theta}}(\mathbf{X})\right\} \\ &= \mathbb{E}_{\mathbf{X}}\left\{\mathbb{1}(F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(\mathbf{X}')) \cdot p(\mathbf{Y} = F_{\boldsymbol{\theta}}(\mathbf{X})|\mathbf{X})\right\}, \end{aligned}$$

the inequality (7) holds. □

B. Experimental Setup

B.1. Hyperparameter setting

Table 9: **Selected hyperparameters.** Hyperparameters used in the numerical studies in Table 2 and 3.

Dataset	Model	**Method**	λ	γ	β	τ	Teacher
CIFAR-10	WRN-28-5,2	RST	5	-	-	-	Supervised
		UAT++	5	-	-	-	Supervised
		SRST-AWR	20	4	0.5	1.2	FixMatch
CIFAR-100	WRN-28-8	RST	5	-	-	-	Supervised
		UAT++	5	-	-	-	Supervised
		SRST-AWR	20	4	0.5	1.0	FixMatch
SVHN	WRN-28-2	RST	5	-	-	-	Supervised
		UAT++	5	-	-	-	Supervised
		SRST-AWR	15	4	0.5	1.0	FixMatch
STL-10	WRN-28-5	RST	5	-	-	-	Supervised
		UAT++	5	-	-	-	Supervised
		SRST-AWR	8	4	0.5	1.0	FixMatch

Table 9 presents the hyperparameters used in Table 2 and 3. Most of the hyperparameters are set to be the ones used in the previous studies.

B.2. Loss for Generating Adversarial Examples

As described in Section 2.2, KL-divergence and cross-entropy loss can be used to generate the adversarial examples. When using KL divergence, adversarial examples are generated based on the current predictions, whereas when using cross-entropy, the target label is required. In our experimental setting, we use the cross-entropy loss with target labels which are predicted by the teacher models except for CIFAR-10 since KL-divergence are not stable.

B.3. The teacher models

B.3.1 Hyperparameter setting

Table 10: **Selected hyperparameters.** Hyperparameters used in the numerical studies in Section 4.

Dataset	Model	**Method**	λ	Weight Decay	τ	num_labels	batch size (labeled, unlabeled)
CIFAR-10	WRN-28-5	FixMatch	1	$5e^{-4}$	0.95	4,000	(64, 128)
CIFAR-100	WRN-28-8	FixMatch	1	$5e^{-4}$	0.95	4,000	(64, 128)
SVHN	WRN-28-2	FixMatch	1	$5e^{-4}$	0.95	1,000	(64, 128)
STL-10	WRN-28-5	FixMatch	1	$5e^{-4}$	0.95	1,000	(64, 128)

Table 11: **Performance of Teachers**.

	CIFAR-10 ($n_l = 4,000$)	CIFAR-100 ($n_l = 4,000$)	SVHN ($n_l = 1,000$)	STL-10 ($n_l = 1,000$)
Supervised	81.82	45.96	83.50	56.60
FixMatch	95.87	64.82	97.11	92.46

We train the model using FixMatch. Table 10 and 11 shows the selected hyperparmeters for training teacher models and performance of them.

B.4. Comparison SRST-AWR to SRST-TRADES

Table 12: **Selected hyperparameters.** Hyperparameters used in the numerical studies in Table 5.

Dataset	Model	**Method**	λ	γ	β	τ	Teacher
CIFAR-10	WRN-28-5	SRST-TRADES	12	4	-	1.2	FixMatch
		SRST-AWR	20	4	0.5	1.2	FixMatch
CIFAR-100	WRN-28-8	SRST-TRADES	20	4	-	1.0	FixMatch
		SRST-AWR	20	4	0.5	1.0	FixMatch
STL-10	WRN-28-5	SRST-TRADES	8	4	-	1.0	FixMatch
		SRST-AWR	8	4	0.5	1.0	FixMatch

Table 12 presents the hyperparameters used in Table 5.

B.5. Comparison SRST-AWR to Other Competitors in Supervised Setting

Table 13: **Selected hyperparameters.** Hyperparameters used in the numerical studies in Table 7.

Method	λ	γ	β	τ	# of labeled data
PGD-AT	-	-	-	-	50,000(100%)
TRADES	6	-	-	-	50,000(100%)
MART	4	-	-	-	50,000(100%)
SRST-AWR	20	4	0.5	1.2	4,000(8%)
SRST-AWR	20	4	0.5	1.2	6,000(12%)

Table 13 presents the hyperparameters used in Table 2. Most of the hyperparameters are set to be the ones used in the previous studies [19, 36, 30].

B.6. Performance on Fully Labeled Data

Table 14: **Selected hyperparameters.** Hyperparameters used in the numerical studies in Table 8.

Method	λ	# of labeled data
TRADES	6	50,000(100%)
AWR	9	50,000(100%)
AWP-TRADES	6	50,000(100%)
AWP-AWR	9	50,000(100%)

Table 14 presents the hyperparameters used in Table 8. We select models with maximized robust accuracies against PGD^{10}.

C. Additional Experiments

C.1. Effect of knowledge distillation

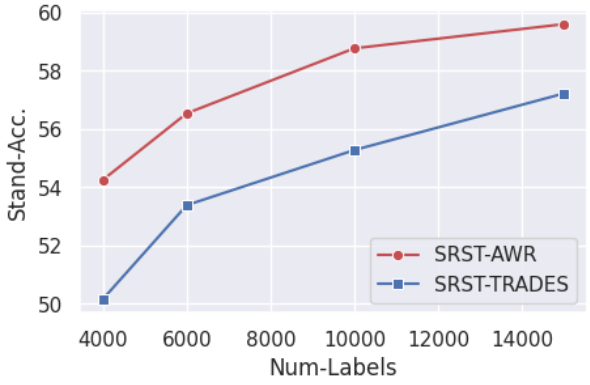
We conduct a comparison of knowledge distillation in both supervised and semi-supervised settings using the same architecture for the teacher and student models. For the supervised setting, we set α and τ to 0.9 and 20, respectively, as suggested in [13]. For the semi-supervised setting, we perform a grid search and found that setting α to 0.9 and τ to 1.1 yields the optimal results. The results are presented in Table 15, which show that while the student model can outperform the teacher model in the supervised setting, it cannot achieve comparable performance to the teacher model in the semi-supervised setting.

Table 15: **The Effect of Knowledge Distillation**.

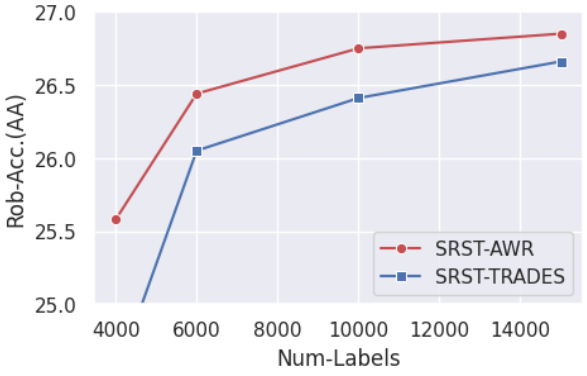
Setting	Teacher Acc.	Student Acc.	Diff.	# of labeled data
Supervised	96.04(0.11)	96.14(0.09)	+0.1	50,000(100%)
Semi-supervised	95.87(0.05)	95.37(0.04)	-0.5	4,000(8%)

C.2. The comparison SRST-AWR and SRST-TRADES with varying the number of labeled data

We compare the performance of SRST-AWR and SRST-TRADES with varying numbers of labeled data for CIFAR-100 and STL-10 to assess the effect of $w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T)$. Figure 3 and 4 present the results, showing that SRST-AWR consistently outperforms SRST-TRADES for both datasets across all labeled data sizes. For CIFAR-100, we use 4,000, 6,000, 10,000, and 15,000 labeled data , while for STL-10, we use 100, 250, 500, 1,000 and 2,000 labeled data. For both datasets, we observe that both standard and robust accuracies against AA improve as the number of labeled data increases. The margins between SRST-AWR and SRST-TRADES are relatively high, especially when the number of labeled data is 4,000 on CIFAR-100 and 1,000 on STL-10. Overall, our results demonstrate that SRST-AWR can outperform SRST-TRADES, even with limited amounts of labeled data, indicating the efficacy of $w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T)$ in improving adversarial robustness.

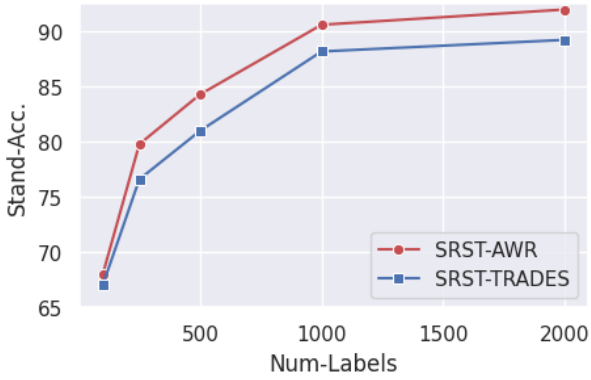


(a) y-axis : standard accuracies

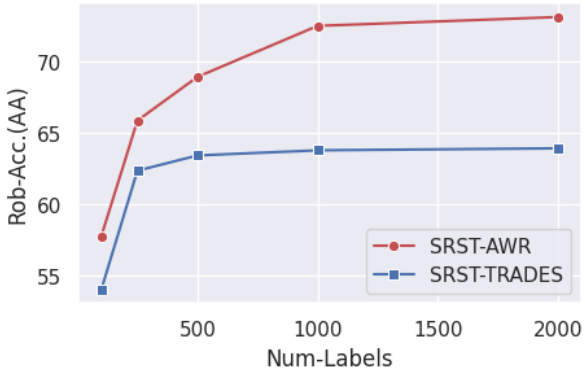


(b) y-axis : robust accuracies

Figure 3: **Comparison SRST-AWR to SRST-TRADES with varying the number of labeled data on CIFAR100**. The x-axis and y-axis are the number of labeled data and performance measure - standard accuracies and robust accuracies against AA , respectively.



(a) y-axis : standard accuracy



(b) y-axis : robust accuracy

Figure 4: **Comparison SRST-AWR to SRST-TRADES with varying the number of labeled data on STL-10**. The x-axis and y-axis are the number of labeled data and performance measure - standard accuracies and robust accuracies against AA , respectively.

C.3. Sensitivity Analysis on β

In this subsection, we perform a sensitivity analysis on β. Figure 5 illustrates that the highest level of robustness can be attained at $\beta = 0.5$ on CIFAR-10. Additionally, standard accuracy remains relatively high when β is in the range of [0, 0.5].

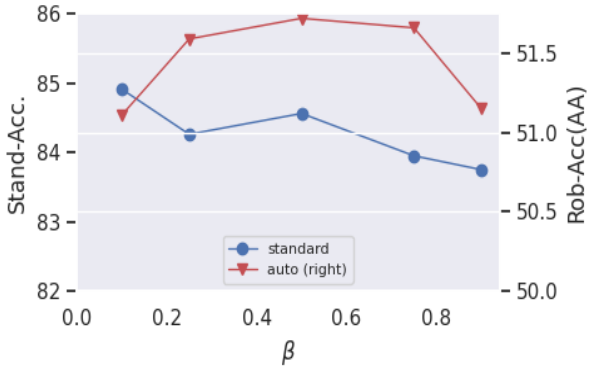


Figure 5: **Sensitivity Analysis for** β. We vary β from 0 to 1 in SRST-AWR. The x-axis is β and y-axes are standard accuracy and robust accuracy against AA, respectively.

C.4. Sensitivity Analysis of Temperature τ

We perform a sensitivity analysis on the temperature parameter τ for knowledge distillation on CIFAR-10. Figure 6 demonstrates that the selection of τ affects both standard and robust accuracies. Specifically, increasing τ to 1.4 enhances robustness but deteriorates generalization. On the other hand, if τ exceeds 1.2, both robustness and generalization decline. Therefore, the results obtained with $\tau = 1.2$ are more favorable compared to other choices for enhancing robustness on CIFAR-10.

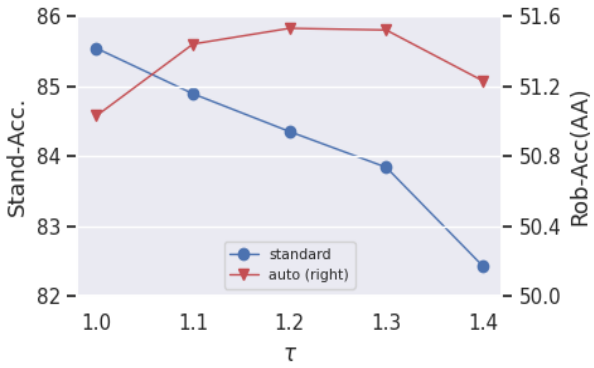


Figure 6: **Sensitivity Analysis for τ.** We vary τ from 1 to 1.4 in SRST-AWR. The x-axis is τ and y-axes are standard accuracy and robust accuracy against AA, respectively.