Appendix A. Additional Experiments

Appendix A.1. Defending Black-box Attacks

Table A.1: Robust accuracy under transfer-based black-box attacks

Defense Models	Attacked Models (CIFAR-100)				Attacked Models (SVHN)			
	Vanilla Training		Adversarial Training		Vanilla Training		Adversarial Training	
	FGSM	PGD|CW20	FGSM	PGD|CW20	FGSM	PGD|CW20	FGSM	PGD|CW20
AT [1]	**59.57**	60.30|60.24	56.62	55.42|56.58	88.31	89.54|89.60	77.65	73.43|74.34
TRADES [2]	59.29	59.52|59.93	54.90	53.70|55.20	**90.47**	91.90|91.89	83.29	78.10|78.79
ATLD	56.51	57.14|57.32	57.46	55.14|56.97	89.40	91.74|92.15	91.05	87.25|88.39
ATLD+	56.31	**61.14|62.08**	**58.15**	**60.62|61.56**	90.40	**92.71|92.97**	**92.24**	**90.03|90.60**

We also performed transfer-based black-box attack experiments on CIFAR-100 and SVHN. Two different agent models (Resnet-18) are used for generating test time attacks, including the Vanilla Training model and the Adversarial Training with PGD model. As demonstrated by the results in Table A.1, our proposed approach can achieve competitive performance in almost all cases. Specifically, ATLD+ outperforms AT and TRADES in 10 out of 12 cases while demonstrating comparable or slightly worse accuracy in the other 2 cases. The performance of our two methods shows marginally inferior to AT or TRADES against FGSM data, while our method outperforms AT and TRADES significantly against PGD20 and CW20 adversarial attacks both on the vanilla training model and the adversarial training with PGD model.

Appendix B. Detailed Derivation

In the main content of this paper, we defined our proposed main objective as:

$$
\begin{aligned}
\min_{\theta}\Big\{ & \sum_{i=1}^{N} \underbrace{L(x_i^{adv}, y_i; \theta)}_{L_f} + \\
& \sup_{W} \sum_{i=1}^{N} \underbrace{[\log D_W(f_\theta(x_i^{adv})) + (1 - \log D_W(f_\theta(x_i)))]}_{L_d} \Big\} \\
\text{s.t.} \quad & x_i^{adv} = \arg \max_{x_i' \in B(x_i, \epsilon)} [\log D_W(f_\theta(x_i')) \\
& \qquad + (1 - \log D_W(f_\theta(x_i)))]
\end{aligned}
\tag{B.1}
$$

In this section, we provide the details about the derivation for the main objective function Equation (B.1) (exactly the same as Equation (6) in the main paper) and elaborate on how to compute adversarial examples and transformed examples.

Appendix B.1. Derivation for Main Objective Function (Equation (B.1))

We start with minimizing the largest f-divergence between latent distributions P_θ and Q_θ induced by perturbed example x' and natural example x. And we denote their corresponding probability density functions as $p(z)$ and $q(z)$. According to Equation (3) in the main paper, we have

$$
\begin{aligned}
& \min_{\theta} \max_{Q_\theta} D_f(P_\theta || Q_\theta) \\
& = \min_{\theta} \max_{q(z)} \int_{\mathcal{Z}} q(z) \sup_{t \in \mathrm{dom} f^*} \{ t \frac{p(z)}{q(z)} - f^*(t) \} dx \\
& \geq \min_{\theta} \max_{q(z)} \sup_{T \in \tau} \Big(\int_{\mathcal{Z}} p(z) T(z) dz \\
& \qquad\qquad - \int_{\mathcal{Z}} q(z) f^*(T(z)) dz \Big) \\
& = \min_{\theta} \max_{Q_\theta} \sup_{W} \Big\{ \mathbb{E}_{z \sim P_\theta} [g_f(V_W(z))] \\
& \qquad\qquad + \mathbb{E}_{z \sim Q_\theta} [-f^*(g_f(V_W(z)))] \Big\} \\
& = \min_{\theta} \sup_{W} \Big\{ \mathbb{E}_{x \sim \mathcal{D}} \Big\{ \max_{x' \in B(x, \epsilon)} [g_f(V_W(f_\theta(x')))] \\
& \qquad\qquad + [-f^*(g_f(V_W(f_\theta(x))))] \Big\} \Big\}
\end{aligned}
\tag{B.2}
$$

To compute the Jensen-Shannon divergence between P_θ and Q_θ, we set $g_f(t) = -\log(1+e^{-t})$ and $f^*(g) = -\log(2-e^g)$. Then, we have

$$\begin{aligned}
&\min_\theta \max_{Q_\theta} D_{JS}(P_\theta || Q_\theta) \\
&\geq \min_\theta \sup_W \Big\{ \mathbb{E}_{x\sim\mathcal{D}} \big\{ \max_{x' \in B(x,\epsilon)} [\log D_W(f_\theta(x'))] \\
&\qquad\qquad + [1 - \log D_W(f_\theta(x)))] \big\} \Big\}
\end{aligned} \tag{B.3}$$

where $D_W(x) = 1/(1+e^{-V_W(x)})$ is equivalent to optimize the lower bound of Jensen-Shannon divergence between P_θ and Q_θ. With disentangling the computation of adversarial examples from Eq. (B.3) and further considering the classification loss for the classifier L_f and the discriminator $L_d^{1:C}$, we can obtain the final objective:

$$\begin{aligned}
\min_\theta \Big\{ &\sup_W \sum_{i=1}^{N} [\underbrace{\log D_W^0(f_\theta(x_i^{adv})) + (1 - \log D_W^0(f_\theta(x_i))}_{L_d^0}] \\
&+ \underbrace{L(x_i^{adv}, y_i; \theta)}_{L_f} \\
&+ \min_W [\underbrace{l(D_W^{1:C}(f_\theta(x_i)), y_i) + l(D_W^{1:C}(f_\theta(x_i^{adv})), y_i)}_{L_d^{1:C}}] \Big\}, \\
\text{s.t.} \quad x_i^{adv} = \arg &\max_{x_i' \in B(x_i,\epsilon)} [\log D_W^0(f_\theta(x_i')) \\
&\qquad + (1 - \log D_W^0(f_\theta(x_i))]
\end{aligned} \tag{B.4}$$

Appendix B.2. Computation for Adversarial Example and Transformed Example

To compute the adversarial example, we need to solve the following problem:

$$x_i^{adv} = \arg \max_{x_i' \in B(x_i,\epsilon)} \underbrace{[\log D_W^0(f_\theta(x_i')) + (1 - \log D_W^0(f_\theta(x_i))]}_{L_d^0} \tag{B.5}$$

It can be reformulated as computing the adversarial perturbation as follows:

$$r_i^{adv} = \arg \max_{\|r\|_\infty \leq \epsilon} [L_d^0(x_i + r_i, \theta)] \tag{B.6}$$

We first consider the more general case $\|r\|_p \leq \epsilon$ and expand (B.6) with the first order Taylor expansion as follows:

$$r_i^{adv} = \arg\max_{\|r\|_p \leq \epsilon} [L_d^0(x_i, \theta)] + \nabla_x \mathcal{F}^T r_i \tag{B.7}$$

where $\mathcal{F} = L(x_i, \theta)$. The problem (B.7) can be reduced to:

$$\max_{\|r_i\|_p = \epsilon} \nabla_x \mathcal{F}^T r_i \tag{B.8}$$

We solve it with the Lagrangian multiplier method and we have

$$\nabla_x \mathcal{F} r_i = \lambda(\|r_i\|_p - \epsilon) \tag{B.9}$$

Then we make the first derivative with respect to r_i:

$$\nabla_x \mathcal{F} = \lambda \frac{r_i^{p-1}}{p(\sum_j (r_i^j)^p)^{1-\frac{1}{p}}} \tag{B.10}$$

$$\nabla_x \mathcal{F} = \frac{\lambda}{p} (\frac{r_i}{\epsilon})^{p-1}$$

$$(\nabla_x \mathcal{F})^{\frac{p}{p-1}} = (\frac{\lambda}{p})^{\frac{p}{p-1}} (\frac{r_i}{\epsilon})^p \tag{B.11}$$

If we sum over two sides, we have

$$\sum (\nabla_x \mathcal{F})^{\frac{p}{p-1}} = \sum (\frac{\lambda}{p})^{\frac{p}{p-1}} (\frac{r_i}{\epsilon})^p \tag{B.12}$$

$$\|\nabla_x \mathcal{F}\|_{p^*}^{p^*} = (\frac{\lambda}{p})^{p^*} * 1 \tag{B.13}$$

where p^* is the dual of p, i.e. $\frac{1}{p} + \frac{1}{p^*} = 1$. We have

$$(\frac{\lambda}{p}) = \|\nabla_x \mathcal{F}\|_{p^*} \tag{B.14}$$

By combining (B.11) and (B.14), we have

$$\begin{aligned} r_i^* =& \epsilon \text{sgn}(\nabla_x \mathcal{F})(\frac{|\nabla_x \mathcal{F}|}{\|\nabla_x \mathcal{F}\|_{p^*}})^{\frac{1}{p-1}} \\ =& \epsilon \text{sgn}(\nabla_x L_d^0)(\frac{|\nabla_x L_d^0|}{\|\nabla_x L_d^0\|_{p^*}})^{\frac{1}{p-1}} \end{aligned} \tag{B.15}$$

In this paper, we set p to ∞. Then we have

$$\begin{aligned} r_i^* &= \epsilon \lim_{p\to\infty} sgn(\nabla_x L_d^0)(\frac{|\nabla_x L_d^0|}{\|\nabla_x L_d^0\|_{p^*}})^{\frac{1}{p-1}} \\ &= \epsilon \text{sgn}(\nabla_x L_d^0)(\frac{|\nabla_x L_d^0|}{\|\nabla_x L_d^0\|_1})^0 \\ &= \epsilon \text{sgn}(\nabla_x L_d^0) \end{aligned} \tag{B.16}$$

Then we can obtain the adversarial example:

$$x_i^* = x_i + \epsilon \text{sgn}(\nabla_x L_d^0) \tag{B.17}$$

To compute the transformed example, we need to solve the following problem:

$$r^* = \arg \min_{\|r\|_\infty \leq \epsilon} \log D_W^0(f_\theta(x + r)). \tag{B.18}$$

With the similar method, we can easily get the transformed example x^t

$$x^t = x - \epsilon \text{sgn}(\nabla_x \log D_W^0). \tag{B.19}$$

References

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations (ICLR), OpenReview.net, 2018.

[2] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, M. I. Jordan, Theoretically principled trade-off between robustness and accuracy, in: International Conference on Machine Learning (ICML), Vol. 97, PMLR, 2019, pp. 7472–7482.