A1 DETAILS OF THEOREM 1

Here, we provide a detailed proof of the decomposition of the robust risk given in Eq. 9.

Theorem 1. For a DNN f_θ, parameterized by θ , the robust risk $\mathcal{R}(\theta)$ for a batch of samples (X_B^t, Y_B^t) can be written as:

$$\begin{aligned}\mathcal{R}(\theta) = \mathbb{E}_{(X_B^t, Y_B^t)} \mathbb{1}\Big\{ F_\theta(X_B^t) \neq Y_B^t \Big\} \; + \\ p\Big(\exists \tilde{X}_B^t \in \mathcal{B}_p(X_B^t, \epsilon) : F_\theta(X_B^t) \neq F_\theta(\tilde{X}_B^t) \Big) \cdot p\Big(Y_B^t = F_\theta(X_B^t) \mid X_B^t \Big)\end{aligned} \tag{A1}$$

Proof. From Eq. 6, $\mathcal{R}(\theta) = \mathcal{R}_{nat}(\theta) + \mathcal{R}_{bdy}(\theta)$. where, $\mathcal{R}_{nat}(\theta) = \mathbb{E}_{(X_B^t, Y_B^t)} \mathbb{1}\left\{ F_\theta(X_B^t) \neq Y_B^t \right\}$ and $\mathcal{R}_{bdy}(\theta) = \mathbb{E}_{(X_B^t, Y_B^t)} \mathbb{1}\Big\{ \exists \tilde{X}_B^t \in \mathcal{B}_p(X_B^t, \epsilon) \; : \; F_\theta(X_B^t) \neq F_\theta(\tilde{X}_B^t),\, F_\theta(X_B^t) = Y_B^t \Big\}$.

Since,

$$\begin{aligned}\mathcal{R}_{bdy}(\theta) &= \mathbb{E}_{(X_B^t, Y_B^t)} \mathbb{1}\Big\{ \exists \tilde{X}_B^t \in \mathcal{B}_p(X_B^t, \epsilon) : F_\theta(X_B^t) \neq F_\theta(\tilde{X}_B^t),\, F_\theta(X_B^t) = Y_B^t \Big\} \\ &= \mathbb{E}_{(X_B^t, Y_B^t)} \mathbb{1}\Big\{ \exists \tilde{X}_B^t \in \mathcal{B}_p(X_B^t, \epsilon) : F_\theta(X_B^t) \neq F_\theta(\tilde{X}_B^t) \Big\} \cdot \mathbb{1}\left\{ F_\theta(X_B^t) = Y_B^t \right\} \\ &= \mathbb{E}_{X_B^t} \Big[\mathbb{1}\Big\{ \exists \tilde{X}_B^t \in \mathcal{B}_p(X_B^t, \epsilon) : F_\theta(X_B^t) \neq F_\theta(\tilde{X}_B^t) \Big\} \cdot \mathbb{E}_{(Y_B^t \mid X_B^t)} \mathbb{1}\left\{ F_\theta(X_B^t) = Y_B^t \right\} \Big] \\ &\qquad\qquad [\because \text{Law of Iterated Expectation}] \\ &= \mathbb{E}_{X_B^t} \Big[\mathbb{1}\Big\{ \exists \tilde{X}_B^t \in \mathcal{B}_p(X_B^t, \epsilon) : F_\theta(X_B^t) \neq F_\theta(\tilde{X}_B^t) \Big\} \cdot p\left(Y_B^t = F_\theta(X_B^t) \mid X_B^t \right) \Big] \\ &= \mathbb{E}_{X_B^t} \Big[\mathbb{1}\Big\{ \exists \tilde{X}_B^t \in \mathcal{B}_p(X_B^t, \epsilon) : F_\theta(X_B^t) \neq F_\theta(\tilde{X}_B^t) \Big\} \Big] \cdot \mathbb{E}_{X_B^t} \left[p\left(Y_B^t = F_\theta(X_B^t) \mid X_B^t \right) \right] \\ &= p\Big(\exists \tilde{X}_B^t \in \mathcal{B}_p(X_B^t, \epsilon) : F_\theta(X_B^t) \neq F_\theta(\tilde{X}_B^t) \Big) \cdot p\left(Y_B^t = F_\theta(X_B^t) \mid X_B^t \right)\end{aligned}$$

Thus the equality holds.

A2 EFFECT OF DNN ARCHITECTURE ON FCA

To examine whether the source DNN architecture significantly impacts TTA vulnerabilities, we evaluated the performance of FCA on the MobileNet family of DNNs, specifically using MobileNet-v2 as the source DNN. We assessed the performance of five baseline TTA methods across three benchmark datasets and report the results in Table 7. Our findings show that TTA methods exhibit similar vulnerabilities on CIFAR-10C and CIFAR-100C datasets as those observed with ResNet variants. However, for the ImageNet-C benchmark, MobileNet-v2 proved to be even more vulnerable, with performance degradation under FCA being $\sim 10\%$ greater compared to the ResNet-50 results.

A3 PERFORMANCE EVALUATION FOR ADVERSARIALLY TRAINED MODELS

A potential defense against the vulnerabilities highlighted by FCA is to proactively use an adversarially trained source DNN. To evaluate this, we utilized the adversarially trained WideResNet-28 with an l_∞ budget ($\epsilon_\infty = 8/255$) from Robustbench Croce et al. (2020) by Wu et al. (2020), and assessed its performance on CIFAR10-C and CIFAR-100C benchmark datasets. The results are reported in Table 8. Adversarially trained DNNs are highly effective against FCA when the same perturbation is used for both crafting adversarial examples and training the source DNN. However, with a different perturbation budget, such as an l_2 norm constraint of ($\epsilon_\infty = 8/255$), FCA can still degrade performance by approximately 4%. Furthermore, for the CIFAR-100C dataset, adversarially trained source DNNs result in more than a 10% increase in error rate during adaptation with benign data. This is unexpected, as TTA is generally intended to handle online data batches without adversarial perturbation, raising concerns about the robustness-utility trade-off in deploying adversarially

Table 7: (% Error) comparison on MobileNet architectures.

Dataset	Attack Method	TTA Method				
		TeBN	**TENT**	**EATA**	**SAR**	**SoTTA**
CIFAR10-C	w/o Attack	21.04	21.55	23.94	20.93	19.91
	DIA	35.54	35.11	35.56	34.44	34.48
	DIA (PL)	22.07	22.64	24.87	21.55	21.03
	TePA	22.44	22.61	24.81	21.63	21.15
	FCA	**39.33**	**38.34**	**40.01**	**38.07**	**38.09**
CIFAR100-C	w/o Attack	45.55	44.81	45.83	44.63	43.84
	DIA	**57.13**	**55.45**	57.03	**56.14**	**55.93**
	DIA (PL)	46.67	46.55	47.03	46.41	45.59
	TePA	46.74	46.51	46.98	46.55	45.71
	FCA	56.88	55.19	**57.21**	56.04	55.45
ImageNet-C	w/o Attack	54.2	52.97	53.78	52.83	51.29
	DIA	**71.56**	**70.37**	**70.45**	**70.87**	68.55
	DIA (PL)	57.5	56.44	56.29	55.31	54.92
	FCA	71.44	70.01	70.31	70.15	**70.22**

Table 8: (% Error) comparison on adversarially trained models.

Dataset	Evaluation Setup	TTA Method				
		TeBN	**TENT**	**EATA**	**SAR**	**SoTTA**
CIFAR10-C	Unattacked(Standard)	17.14	16.98	19.21	16.88	16.42
	Unattacked(Adv trained)	19.21	16.22	18.44	17.91	15.40
	FCA ($\epsilon_\infty = 8/255$)	21.44	18.01	20.25	19.83	17.17
	FCA ($\epsilon_2 = 0.5$)	23.45	20.14	22.03	21.55	19.03
CIFAR100-C	Unattacked(Standard)	31.27	30.91	31.87	30.9	29.3
	Unattacked(Adv trained)	42.04	41.59	42.14	41.51	41.04
	FCA ($\epsilon_\infty = 8/255$)	43.01	42.57	20.25	42.79	41.85
	FCA ($\epsilon_2 = 0.5$)	46.44	45.22	44.55	45.01	44.76

trained DNNs. Additionally, adversarial training is known to reduce accuracy on clean data Zhang et al. (2019); Tsipras et al. (2018). Thus, further scrutiny is required to develop computationally lightweight test-time defenses that are effective against FCA without impairing TTA performance on clean or benign samples from different domains.

Table 9: (% Error) comparison on robust models

Dataset	Evaluation Setup	TTA Method				
		TeBN	**TENT**	**EATA**	**SAR**	**SoTTA**
CIFAR10-C	Unattacked(AugMix)	15.37	14.81	16.47	14.53	14.11
	FCA ($\epsilon_\infty = 8/255$	24.33	23.21	24.98	23.05	22.87
CIFAR100-C	Unattacked(AugMix)	29.34	28.77	30.21	28.55	27.87
	FCA	36.41	35.22	37.02	34.75	34.28

A4 Performance Evaluation for robust models

To further understand how the robustness of the source DNN influences FCA, we analyzed the performance of FCA against source DNNs known for their robustness to distribution shifts. Specifically, we utilized the WideResNet-28 model trained with AugMix Hendrycks et al. (2019) from Robustbench Croce et al. (2020), and the evaluation results are presented in Table 9. While AugMix-trained

models are effective in enhancing robustness against various distribution shifts, they remain highly vulnerable to FCA when deployed for TTA.