## A1 Basic Sanity Checks to Evaluation

To further verify the reliability of our evaluation, we report our results on the basic sanity checks introduced in Athalye *et al.* [1].

- Table 4 shows that iterative attacks (PGD [26] and MI-FGSM [6]) are stronger than one-step attacks (FGSM [10]).
- Table 4 shows that white-box attacks are stronger than black-box attacks [27] (by MI-FGSM).
- Unbounded attacks reach 100% attack success rate (accuracy drops to 0.0%) on all the three datasets.
- Fig. 4 shows that increasing distortion bound increases attack success (decreases accuracy).

## A2 Loss Weight of Self-Supervised AT

We can impose a hyperparameter on our ARTUDA training scheme. Specifically, we can add a loss weight $\lambda$ to Eq. (11), and it is shown as follows:

$$\begin{aligned} &\mathcal{L}_{CE}\big(C(x_s), y_s\big) + \lambda \mathcal{L}_{KL}\big(C(\tilde{x}_t), C([x_t]_{sg})\big) \\ &+ \mathcal{L}_{DA}\big(x_s, x_t\big) + \mathcal{L}_{DA}\big(x_s, \tilde{x}_t\big). \end{aligned} \tag{15}$$

The loss weight $\lambda$ controls the ratio of the Self-Supervised AT objective to the overall objective. In all of our previous experiments, we set $\lambda$ to 1. In this section, we train multiple ARTUDA models with varied $\lambda$, where we use the experimental setup described in Sec. 4. The results are reported in Table 6.

We can find that the robust accuracy significantly increases along with the increase of $\lambda$, while the clean accuracy does not vary obviously. This implies that the robustness of the proposed ARTUDA can be further improved with a larger $\lambda$ though it already outperforms the state-of-the-art methods.

**Table 6.** Results (%) of ARTUDA models with varied hyperparameter $\lambda$.

| $\lambda$ | Clean | FGSM |
|---|---|---|
| 0.2 | 68.9 | 33.3 |
| 0.5 | 66.1 | 39.3 |
| 1.0 | 69.0 | 41.1 |
| 2.0 | 66.5 | 48.5 |
| 5.0 | 68.0 | 54.4 |

## A3 Class-wise Accuracy on VisDA-2017

In Table 7, we report class-wise accuracy under PGD attacks [26] on the VisDA-2017 dataset [29]. The results correspond to the PGD column in Table 4. We can see that ARTUDA achieves the best accuracy across the majority of the classes.

**Table 7.** Class-wise accuracy (%) under PGD attacks on the VisDA-2017 dataset.

| Training method | aero | bicycle | bus | car | horse | knife | motor | person | plant | skate | train | truck | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Natural Training | 4.8 | 0.9 | 1.5 | 0.0 | 0.2 | 0.9 | 0.3 | 3.2 | 0.2 | 0.1 | 0.7 | 0.0 | 0.9 |
| PGD-AT [26] | 49.6 | 20.4 | 15.2 | 8.7 | 34.3 | 7.3 | 27.3 | 32.8 | 35.2 | **17.4** | 19.8 | 3.2 | 21.3 |
| TRADES [42] | 61.8 | 24.5 | 32.0 | 11.4 | 42.9 | 30.6 | 34.1 | **49.1** | 50.1 | 5.6 | 33.1 | 4.8 | 29.7 |
| ARTUDA (ours) | **75.0** | **32.1** | **61.5** | **25.9** | **53.3** | **65.1** | **66.4** | 48.2 | **52.3** | 9.2 | **58.8** | **7.8** | **44.3** |