A Empirical Robustness and Domain Generalization

A.1 The effect of λ on robustness generalization

How does the parameter λ in equation 5 affect the target robustness? We study the effect of varying λ, which controls the robustness-accuracy trade-off in equation 5. Intuitively, the closer λ is to zero, the more robust the model is. However, this added robustness comes at the cost of clean data accuracy. Similar to 1, which shows the robust accuracies on various datasets with given $\lambda = 0.5$, in Figure 5 we visualize the evaluation results for PACS dataset for $\lambda = 0.1$ and $\lambda = 0.9$. The extreme case of robust only training ($\lambda = 0$) is visualized in Figure 6

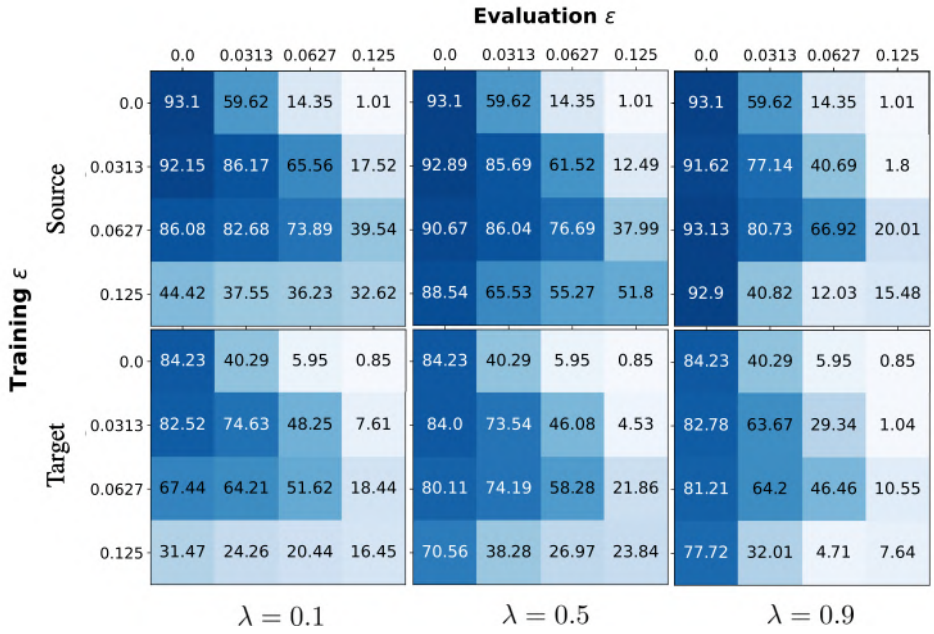


Figure 5: **The Effect of λ on Robustness Generalizability.** As we decrease the value of λ, the network sacrifices the clean accuracy to improve the robust accuracy.

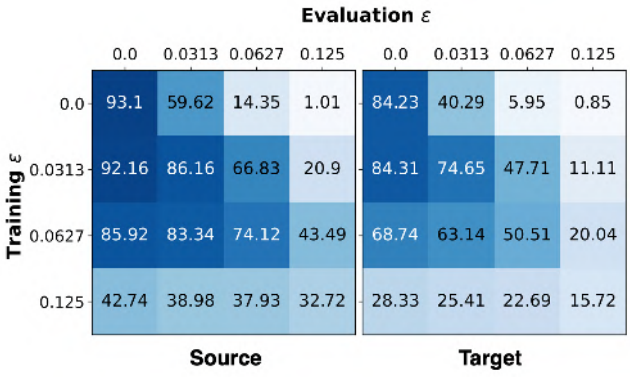


Figure 6: **Robust Training Only ($\lambda = 0$).** In the extreme case, $\lambda = 0$, the network is only trained on adversarial samples from the source domain without any clean source samples. This leads to a sharp drop in the network's accuracy on both the source and target domains.

A.2 The generalization of ℓ_2 robustness

Do the paper conclusions about the generalization of robustness hold if we consider ℓ_2 adversarial attacks? We repeat the experiments in Section 4.2 using ℓ_2 adversarial augmentations.

We set $\epsilon = \{0, 0.5, 1.0, 5.0\}$. Each model is trained on one ϵ but is evaluated on all ϵ values. We see from Figure 7 that the paper conclusions also hold when considering ℓ_2 robustness. Following Section 4.2, we answer the following questions:

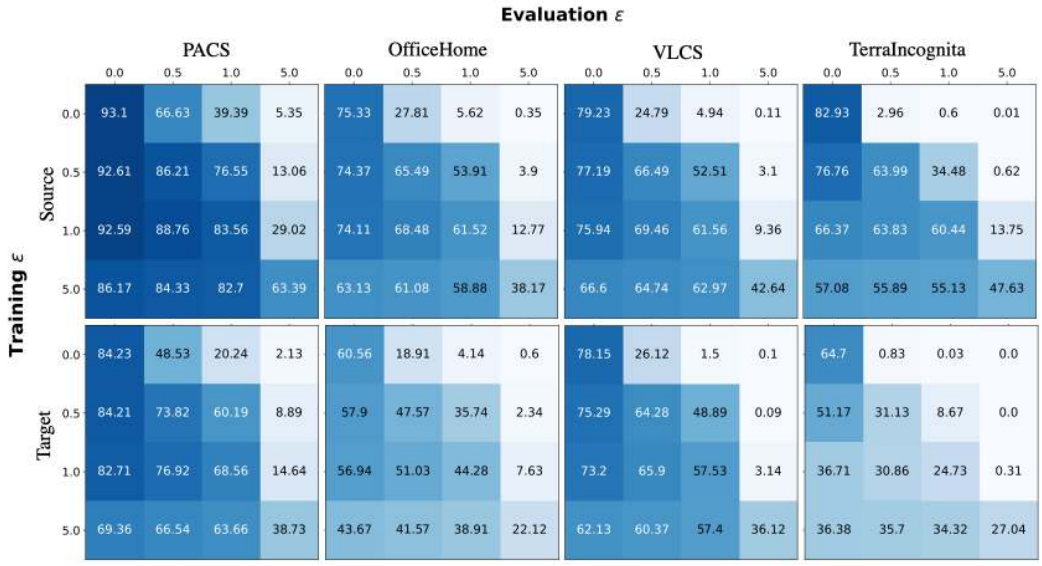


Figure 7: **Evaluation of ℓ_2 Robustness.** The robustness of the source domain (row 1) and the target domain (row 2) follow a similar trend for each dataset. Robustness transfers from the source domains to the target domains, a higher robustness in the source domain is associated with higher robustness in the target domain and vice versa.

Q1: Do adversarially robust models generalize better than their standard-trained counter-parts? Again the answer is no. Adversarially trained models tend to experience a drop in generalizability when compared to their standard-trained counter-parts.

Q2: Does a higher source-domain robustness correspond to a higher target-domain robustness? As expected, the answer is still yes. As observed across Table 7, when we have a higher robustness in the source domain we consistently observe a higher robustness in the target domain.

Q3: Does the robustness-accuracy trade-off generalize to unseen domains? Yes, similar to what was observed in ℓ_∞ experiments, the robustness-accuracy trade-off exists in unseen domains. Robustness in the target domain comes at the cost of clean accuracy.

A.3 THE EFFECT OF USING A STRONGER EMPIRICAL DEFENSE

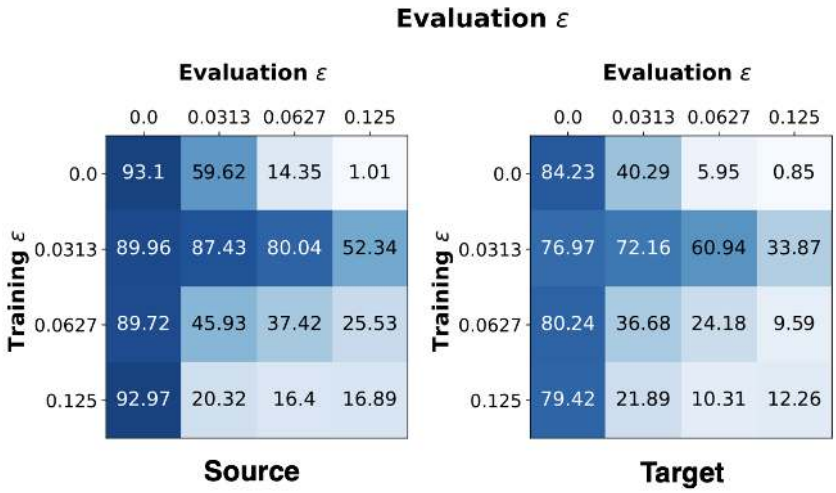


Figure 8: **Robustness Generalizability Using TRADES.** Utilizing a stronger adversarial training method , TRADES, does not provide guarantees towards higher target robustness .

Does the use of a stronger defense, e.g. TRADES (Zhang et al., 2019), improve the generalizability of DNN robustness? When a stronger adversarial robustness method is deployed, in general and unexpectedly we obtain lower robustness generalizability compared to standard adversarial training. We also happen to experience a sharper drop in target accuracy for $\epsilon = 8/255, 16/255$.

B CERTIFIED ROBUSTNESS AND DOMAIN GENERALIZATION

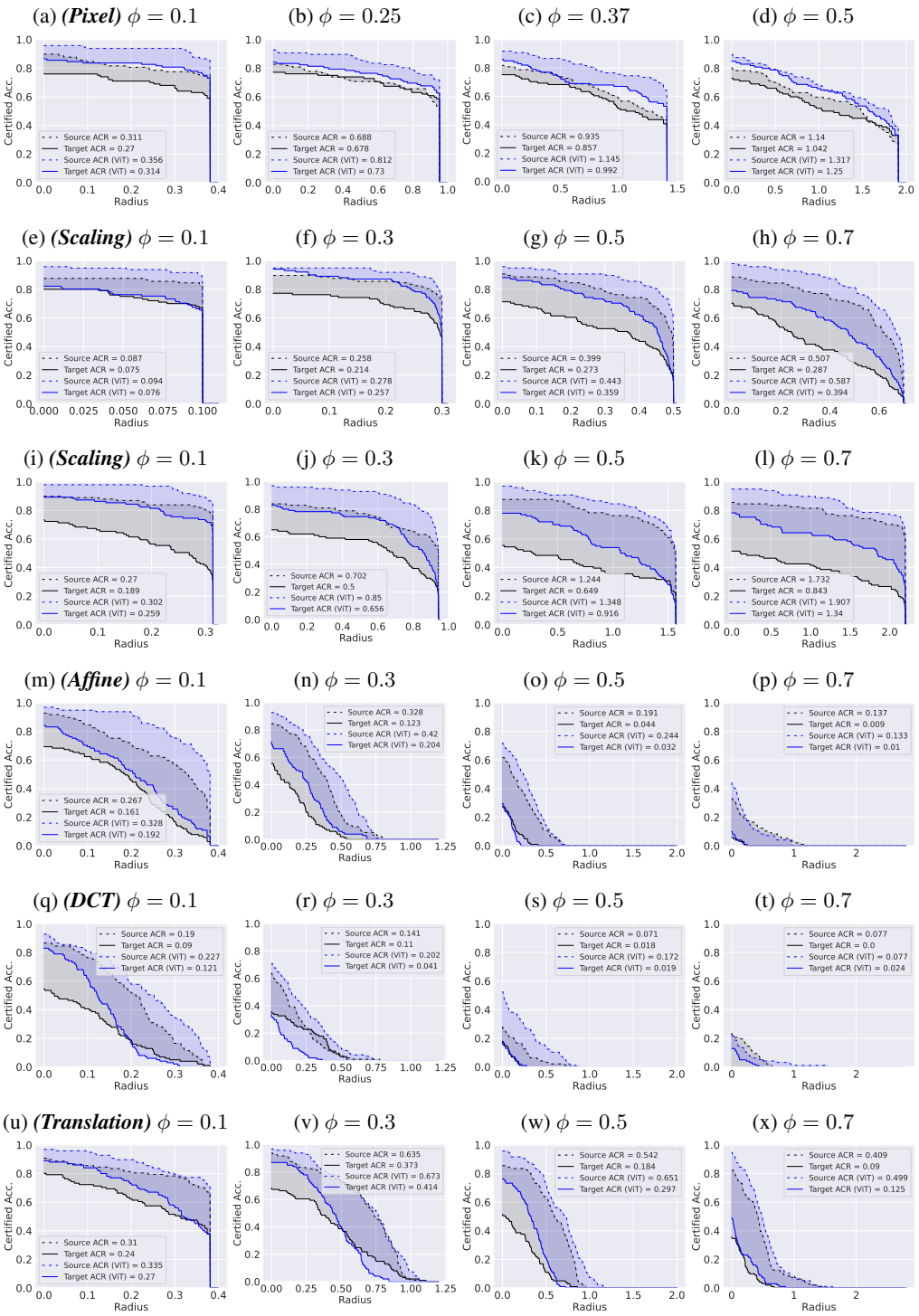


Figure 9: **Effect of Varying the Deformation Parameter ϕ.** We observe that 1) for Pixel Perturbations, Scaling, and Rotation, the higher ϕ gets, the larger the Average Certified Radius (ACR) becomes; and (2) for Affine, DCT, and Translation, high ϕ values can result in low ACRs.

B.1 THE EFFECT OF ϕ ON THE GENERALIZATION OF CERTIFIED ROBUSTNESS

To complement Figure 2, we investigate the behavior when the deformation parameter ϕ varies. Following Section 5.2, we certify ResNet-50 and ViT-Base against pixel perturbations and input deformations in the source and target domains of PACS. We break down each envelope curve in Figure 2 into multiple curves, each representing one choice of ϕ in Eq. 8. We label each curve with the corresponding ϕ value in Figure 9. We observe that the effect of ϕ largely depends on the type of perturbation. On the one hand, for Scaling and Pixel Perturbations, a higher ϕ values corresponds to a larger Average Certified Radius (ACR); on the other hand, for Affine, DCT, and Translation, a higher ϕ values might correspond to a smaller ACR. This is because for the latter group of deformations, a higher ϕ results in a completely deformed image, which hinders the certification ability of the model, even at a small radius. We visualize images from these deformations in Section D. Note that the trends in Figure 2 still stand. Specifically, (1) for ϕ values where there's a reasonable certified accuracy in the source, that certified accuracy generalizes to the target. Moreover, (2) A stronger architecture (ViT-Base) generally leads to a better source and target certified accuracy.

B.2 DOES VISUAL SIMILARITY CORRELATE WITH ROBUSTNESS GENERALIZABILITY?

We repeat the experiments in Section 5.2, which aim to evaluate the ability of FID/R-FID to predict the generalization of robustness, on pixel perturbations and the following deformations: rotation, affine, and DCT. We observe from Figure 10 that the FID/R-FID values do not predict the level of generalizability of certified robustness, which matches our paper findings.

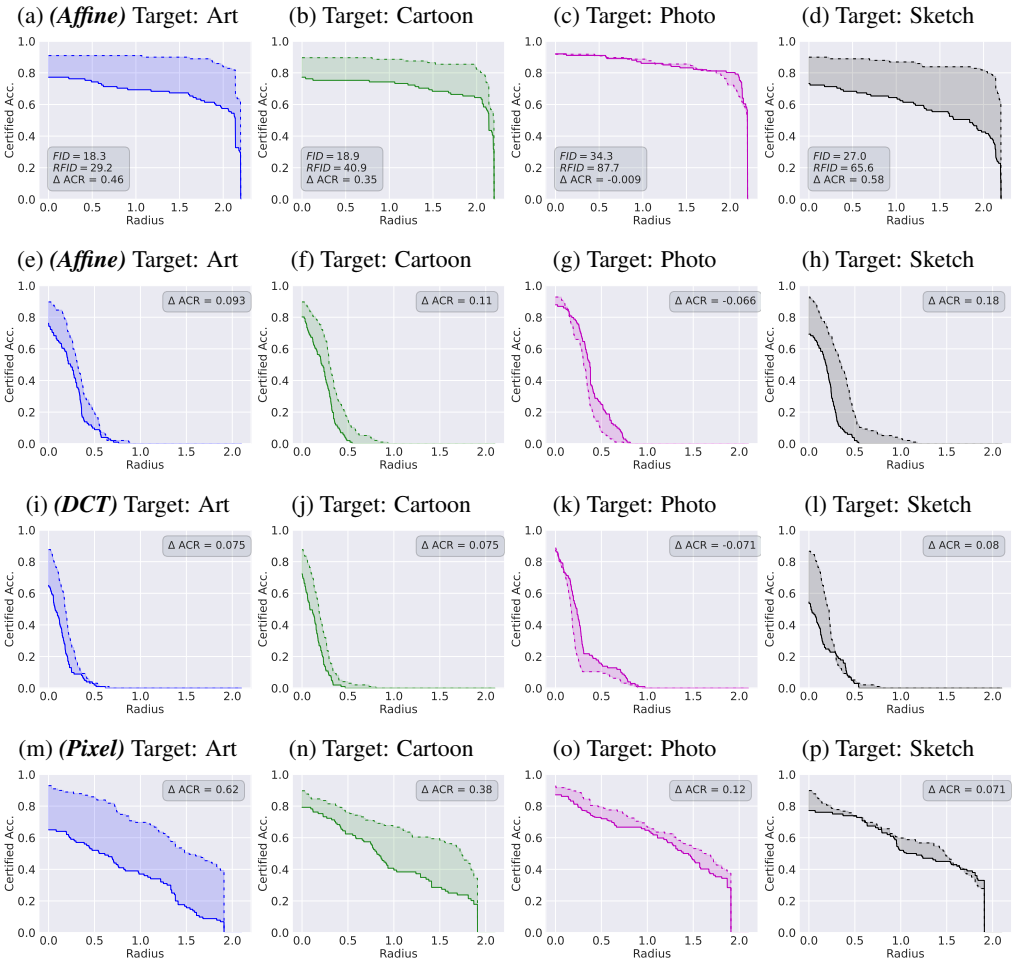


Figure 10: **Does visual similarity correlates with robustness generalizability?** We vary the target distribution and plot the certified accuracy curves for different deformations. The FID/R-FID distances between the source and target distributions are shown in the first row. Visual similarity (FID and R-FID) does not correlate with the level of robustness generalization to the target domain.

C Real-world Application: Medical Images

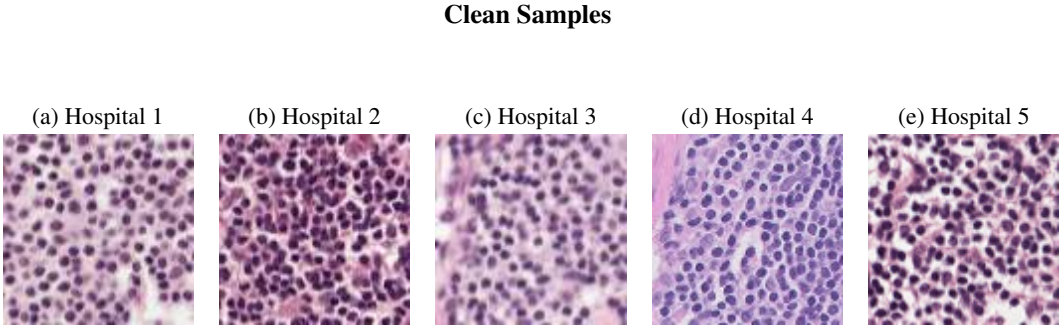


Figure 11: **A visualization of the images taken from the 5 hospitals in Camelyon17.**

We repeat the certified robustness experiments in Section 6 on the following deformations: affine, DCT, translation, and rotation. We observe from Figure 12 that the source-target certification gap

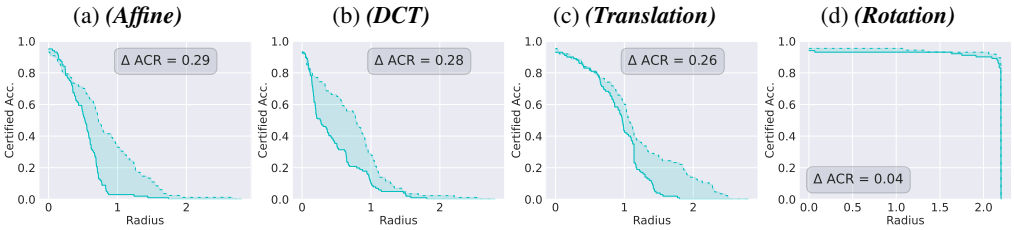


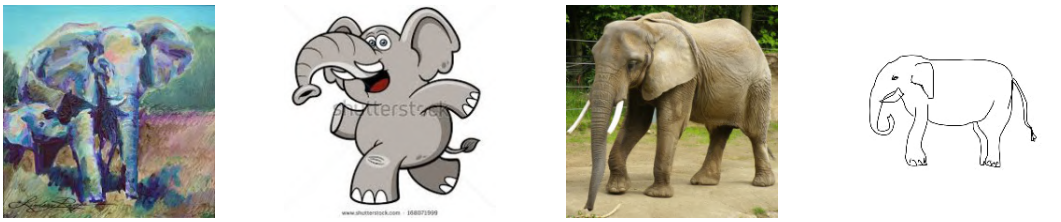
Figure 12: **Does visual similarity correlates with robustness generalizability?** We vary the target distribution and plot the certified accuracy curves for two deformations: scaling and translation. A sample from each distribution is shown in the second row. The FID/R-FID distances between the source distributions and each target are inset in the first row. Visual similarity, measured by FID and R-FID, does not correlate with the level of robustness generalization to the target domain.

is similar for affine, DCT, and translation. However, the certified accuracy curves for rotation are different. This makes sense when we consider the way the Camelyon17 dataset is constructed (Bándi et al., 2019; Koh et al., 2021). The dataset includes cropped histopathological images, each of which may contain a tumor tissue in the central 32x32 region. Due to the nature of this construction, rotated versions of the image look similar, which explains why the source and target certified radii remain almost constant. Samples from the Camelyon17 dataset are visualized in Figure 11.

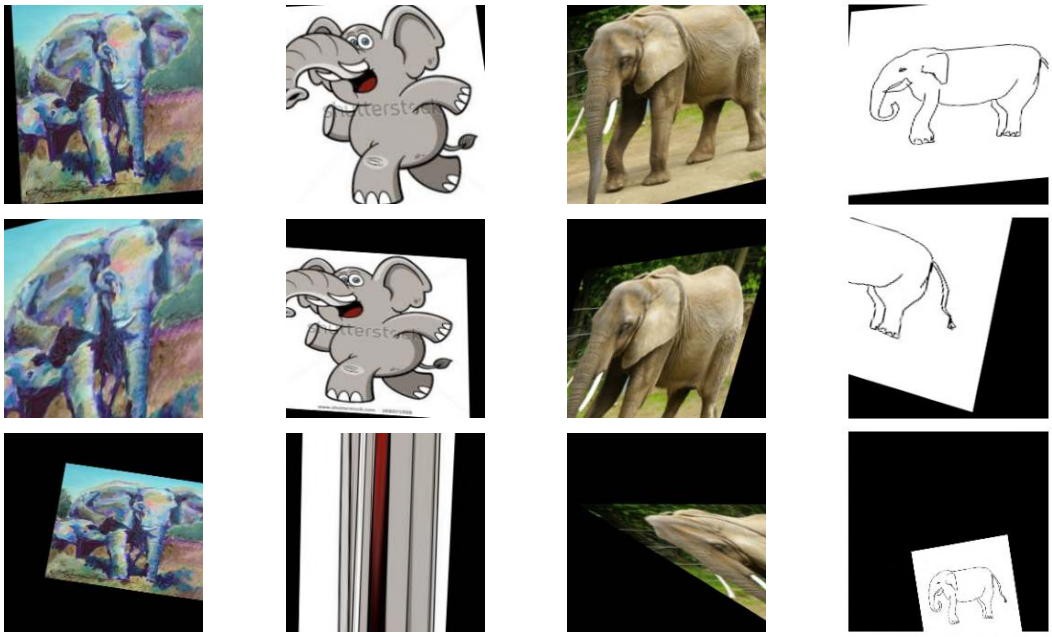
D VISUALIZING THE DOMAIN GENERALIZATION DATASETS

We visualize a few samples from each of the domain generalization datasets we used in the paper. We note that the datasets are diverse in terms of the nature of domain shifts and real-world applicability. Along with the clean samples, we visualize deformed versions of the samples under various values of σ for all the studied deformations.

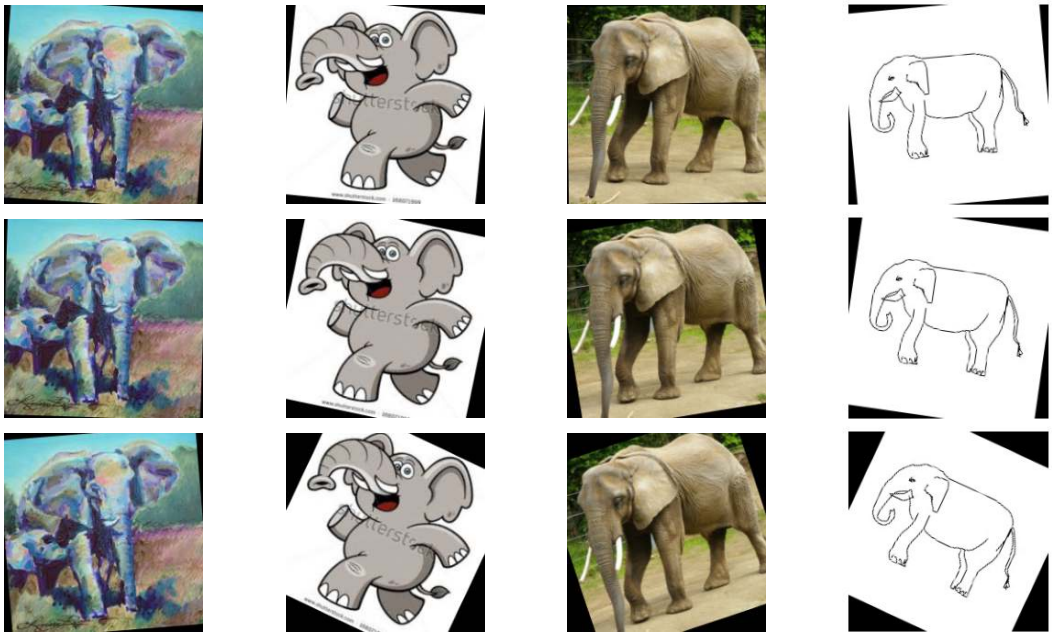
Clean Samples



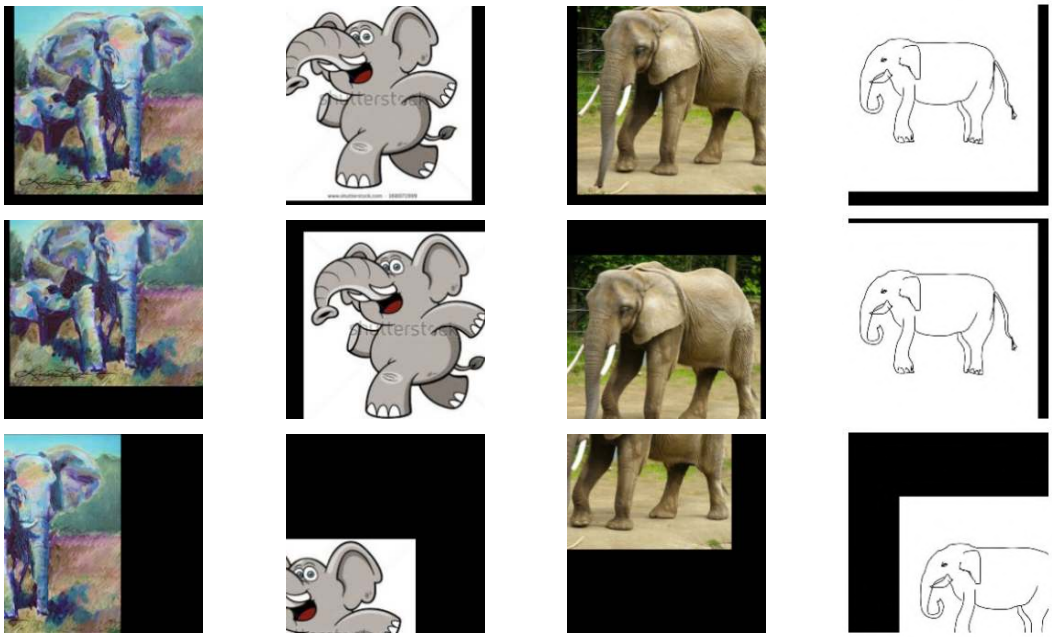
Affine, $\sigma = 0.1, 0.3, 0.5$



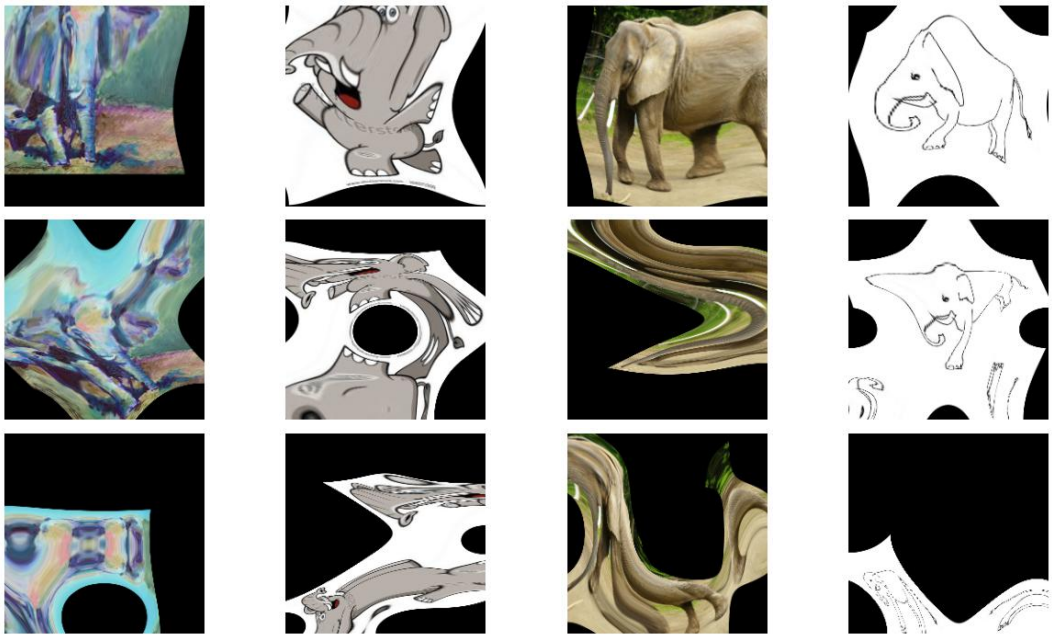
Rotation, $\sigma = 0.1, 0.3, 0.5$



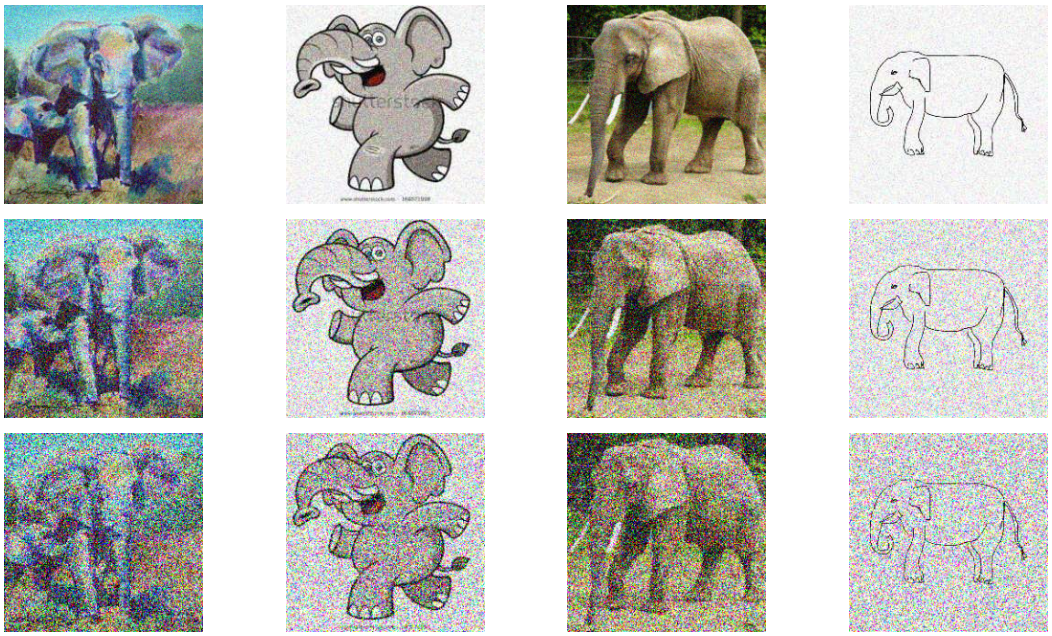
Translation, $\sigma = 0.1, 0.3, 0.5$



DCT, $\sigma = 0.1, 0.3, 0.5$



Pixel Perturbation, $\sigma = 0.1, 0.3, 0.5$



Scaling, $\lambda = 0.1, 0.3, 0.5$