How DMAT can be exploited for standard tasks/datasets? (R1): In this work, we consider the scenario when the manifold information is exact and show that this information can be very useful for improving robustness to novel attacks. For standard tasks / datasets, one possible pipeline may include the following steps: (1) train a generative model (e.g. StyleGAN) to capture the approximate manifold (low-dimensional representation) for the dataset, (2) project the data samples onto the learned manifold, and (3) train a robust classifier using the proposed DMAT. In Table 3 in the paper, we show that although the classifier is only trained using on-manifold samples, remarkably it demonstrates good generalization to natural off-manifold samples. To further boost the performance of DMAT for off-manifold samples, during inference time, one can project the input samples onto the manifold before feeding them to the robust classifier. In this case, the projection operation is *not* used as a defense mechanism, but as an approach to reduce the distribution shift between on-manifold samples and natural images. In Table A, we present evaluation results when the above pipeline is considered. We note that we do not consider end-to-end attacks in this setting since our main focus is to study the robustness of the classification model itself.

Table A: Evaluation of DMAT on natural images with and without projection against attacks on the classifier.

Method	Standard	PGD-50	Fog	Snow	Elastic	Gabor	JPEG	L_2
Normal Training (ERM)	67.21%	0.00%	0.38%	0.35%	0.69%	0.04%	0.00%	1.26%
DMAT	74.72%	34.63%	36.25%	50.56%	54.14%	45.39%	13.29%	48.42%
DMAT + Projection	**77.96%**	**64.39%**	**37.02%**	**65.15%**	**66.47%**	**70.27%**	**72.64%**	**70.77%**

PGD should not be viewed as the strongest attack for evaluation. (R2): We agree with the reviewer that considering a set of adaptive attacks would strengthen our evaluations. Upon your suggestion and for a feasible evaluation runtime, we now consider FGSM, PGD, and the Momentum Iterative Attacks [1] for the L_∞ threat model. Each test sample will be mis-classified if one of the attacks fools the classifier (i.e. the per-input worst case). Results are shown in Table B.

L_1 **attack should be evaluated. (R2):** Upon your suggestion, we now evaluate our proposed method (DMAT) and AT model (trained using L_∞) against unseen L_1 attacks. Results are presented in the last column of Table B. DMAT demonstrates improved generalization by around 9 percentage points compared to adversarial training.

Table B: Classification accuracy on OM-ImageNet test set under L_∞ and L_1 attacks.

Method	Standard	FGSM (L_∞)	PGD-50 (L_∞)	MI-PGD-50 (L_∞)	Worst Case (L_∞)	L_1
Normal Training (ERM)	74.72%	2.59%	0.00%	0.00%	0.00%	0.00%
AT against L_∞ [PGD-5]	73.31%	48.02%	38.88%	39.21%	38.80%	21.37%
DMAT [PGD-5, OM-PGD-5]	77.96%	49.12%	37.86%	37.65%	36.66%	30.70%

Other strong baselines such as TRADES should be included in the main paper. (R2): In the supplementary material (section C.2), we have presented the results of experiments using TRADES and discussed possible combinations of DMAT and TRADES. Results show that the generalization ability of TRADES to unseen attacks can also be improved by exploiting the learned manifold. We will move these results to the main paper and add more discussions on existing methods for unseen attacks.

The notion of "manifold" should be clarified. (R3 and R4): In our paper, manifold refers to the low-dimensional representation for the data samples. In particular, let the generator $G : \mathbb{R}^r \to \mathbb{R}^d$ where $r \ll d$. The range of the generator function G is referred to as "manifold". As we indicate in page 2 footnote, this is not the precise definition of "manifolds" used in topology. We adopted this term since it is commonly used in the generative model area to refer to the existence of lower-dimensional representations for natural images. We will explain this further in the paper.

The existence and uniqueness of the optimization solutions are not discussed .... it's unclear why the approximate image manifold is exact. (R4): For a given natural image x_i, we solve for w_i such that $g(w_i)$ is visually similar to x_i (see some sample results in the supplementary material; Figure 1, On-manifold). The objective we use is standard and proposed in prior works (e.g. [2]). Since the optimization step is solved by a gradient descent based method, the solution may not be unique but this is not an issue for training DMAT. We agree that on-manifold samples $\{g(w)|w \in \mathcal{W}\}$ are approximations to the data samples $\{x_i\}_{i=1}^N$. However, we note that classification model of DMAT, $\{g(w_i)\}_{i=1}^N$ is used as the training images *not* $\{x_i\}_{i=1}^N$, and therefore the manifold information g for the training set of DMAT is in fact *exact*. Remarkably, in Table 3 in the main text, we show that the trained classifier has also a very good generalization to natural images $\{x_j^{test}\}_{j=1}^M$.

Selection of training images. (R4): We partition the Mix-10 dataset into 90% training set and 10% test set since the original test set has a small size. We did not apply any additional curation process in the partition. We will make the training/test datasets, models and our code publicly available.

[1] Dong *et al.*, "Boosting Adversarial Attacks with Momentum", in CVPR 2018.
[2] Abdal *et al.*, "Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?", in ICCV 2019.