A ADDITIONAL ASSUMPTIONS

A4: We list two classical examples here:

- when **A4** is "Θ is finite, $l(\cdot,\cdot)$ is a zero-one loss, samples are *i.i.d*", $\phi(|\Theta|,n,\delta) = \sqrt{(\log(|\Theta|)+\log(1/\delta))/2n}$
- when **A4** is "samples are *i.i.d*", $\phi(|\Theta|,n,\delta) = 2\mathcal{R}(\mathcal{L}) + \sqrt{(\log 1/\delta)/2n}$, where $\mathcal{R}(\mathcal{L})$ stands for Rademacher complexity and $\mathcal{L} = \{l_\theta \,|\, \theta \in \Theta\}$, where l_θ is the loss function corresponding to θ.

For more information or more concrete examples of the generic term, one can refer to relevant textbooks such as (Bousquet et al., 2003).

A5: the worst distribution for expected risk equals the worst distribution for empirical risk, *i.e.*,

$$\arg\max_{\mathcal{P}'\in T(\mathcal{P},\mathcal{A})} r_{\mathcal{P}'}(\widehat{\theta}) = \arg\max_{\mathcal{P}'\in T(\mathcal{P},\mathcal{A})} \widehat{r}_{\mathcal{P}'}(\widehat{\theta})$$

where $T(\mathcal{P},\mathcal{A})$ is the collection of distributions created by elements in $\mathcal{A}$ over samples from $\mathcal{P}$.

Assumption **A5** appears very strong, however, the successes of methods like adversarial training (Madry et al., 2018) suggest that, in practice, **A5** might be much weaker than it appears.

A6: With $(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})$, the worst case sample in terms of maximizing cross-entropy loss and worst case sample in terms of maximizing classification error for model $\widehat{\theta}$ follows:

$$\forall \mathbf{x}, \quad \frac{\mathbf{y}^\top f(\mathbf{x};\widehat{\theta})}{\inf_{a\in\mathcal{A}} \mathbf{y}^\top f(a(\mathbf{x});\widehat{\theta})} \geq \exp\left(\mathbb{I}(g(f(\mathbf{x};\widehat{\theta})) \neq g(f(\mathbf{x}';\widehat{\theta})))\right) \tag{8}$$

where $\mathbf{x}'$ stands for the worst case sample in terms of maximizing classification error, *i.e.*,

$$\mathbf{x}' = \arg\min_{\mathbf{x}} \mathbf{y}^\top g(f(\mathbf{x};\widehat{\theta}))$$

Also,

$$\forall \mathbf{x}, \quad |\inf_{a\in\mathcal{A}} \mathbf{y}^\top f(a(\mathbf{x});\widehat{\theta})| \geq 1 \tag{9}$$

Although Assumption **A6** appears complicated, it describes simple situations that we will unveil in two scenarios:

- If $g(f(\mathbf{x};\widehat{\theta})) = g(f(\mathbf{x}';\widehat{\theta}))$, which means either the sample is misclassified by $\widehat{\theta}$ or the adversary is incompetent to find a worst case transformation that alters the prediction, the RHS of Eq. 8 is 1, thus Eq. 8 always holds (because $\mathcal{A}$ has the identity map as one of its elements).
- If $g(f(\mathbf{x};\widehat{\theta})) \neq g(f(\mathbf{x}';\widehat{\theta}))$, which means the adversary finds a transformation that alters the prediction. In this case, A2 intuitively states that the $\mathcal{A}$ is reasonably rich and the adversary is reasonably powerful to create a gap of the probability for the correct class between the original sample and the transformed sample. The ratio is described as the ratio of the prediction confidence from the original sample over the prediction confidence from the transformed sample is greater than e.

We inspect Assumption **A6** by directly calculating the frequencies out of all the samples when it holds. Given a vanilla model (Base), we notice that over 74% samples out of 50000 samples fit this assumption.

B PROOF OF THEORETICAL RESULTS

B.1 PROOF OF LEMMA 3.1

Lemma. *With Assumptions A1, A4, and A5, with probability at least* $1-\delta$*, we have*

$$\sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} r_{\mathcal{P}'}(\widehat{\theta}) \leq \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}} \sup_{a\in\mathcal{A}} \mathbb{I}(g(f(a(\mathbf{x});\widehat{\theta})) \neq \mathbf{y}) + \phi(|\Theta|, n, \delta) \tag{10}$$

Proof. With Assumption A5, we simply say

$$\arg\max_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} r_{\mathcal{P}'}(\widehat{\theta}) = \arg\max_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \widehat{r}_{\mathcal{P}'}(\widehat{\theta}) = \mathcal{P}_w$$

we can simply analyze the expected risk following the standard classical techniques since both expected risk and empirical risk are studied over distribution $\mathcal{P}_w$.

Now we only need to make sure the classical analyses (as discussed in A4) are still valid over distribution $\mathcal{P}_w$:

- when **A4** is "Θ is finite, $l(\cdot,\cdot)$ is a zero-one loss, samples are *i.i.d*", $\phi(|\Theta|, n, \delta) = \sqrt{\frac{\log(|\Theta|) + \log(1/\delta)}{2n}}$. The proof of this result uses Hoeffding's inequality, which only requires independence of random variables. One can refer to Section 3.6 in Liang (2016) for the detailed proof.
- when **A4** is "samples are *i.i.d*", $\phi(|\Theta|, n, \delta) = 2\mathcal{R}(\mathcal{L}) + \sqrt{\frac{\log 1/\delta}{2n}}$. The proof of this result relies on McDiarmid's inequality, which also only requires independence of random variables. One can refer to Section 3.8 in Liang (2016) for the detailed proof.

Assumption A1 guarantees the samples from distribution $\mathcal{P}_w$ are still independent, thus the generic term holds for at least these two concrete examples, thus the claim is proved.

□

B.2 PROOF OF PROPOSITION 3.2

Proposition. *With A2, and* $d_e(\cdot,\cdot)$ *in A2 chosen to be* ℓ_1 *norm, for any* $a \in \mathcal{A}$*, we have*

$$\sum_i ||f(\mathbf{x}_i;\widehat{\theta}) - f(a(\mathbf{x}_i);\widehat{\theta})||_1 = W_1(f(\mathbf{x};\widehat{\theta}), f(a(\mathbf{x});\widehat{\theta})) \tag{11}$$

Proof. We leverage the order statistics representation of Wasserstein metric over empirical distributions (*e.g.*, see Section 4 in Bobkov & Ledoux (2019))

$$W_1(f(\mathbf{x};\widehat{\theta}), f(a(\mathbf{x});\widehat{\theta})) = \inf_\sigma \sum_i ||f(\mathbf{x}_i;\widehat{\theta}) - f(a(\mathbf{x}_{\sigma(i)}),\widehat{\theta})||_1$$

where σ stands for a permutation of the index, thus the infimum is taken over all possible permutations. With Assumption A2, when $d_e(\cdot,\cdot)$ in A2 chosen to be ℓ_1 norm, we have:

$$||f(\mathbf{x}_i;\widehat{\theta}) - f(a(\mathbf{x}_i);\widehat{\theta})||_1 \leq \min_{j\neq i} ||f(\mathbf{x}_i;\widehat{\theta}) - f(a(\mathbf{x}_j),\widehat{\theta})||_1$$

Thus, the infimum is taken when σ is the natural order of the samples, which leads to the claim. □

B.3 PROOF OF THEOREM 3.3

Theorem. *With Assumptions A1, A2, A4, A5, and A6, and $d_e(\cdot,\cdot)$ in A2 is ℓ_1 norm, with probability at least $1-\delta$, the worst case generalization risk will be bounded as*

$$\sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} r_{\mathcal{P}'}(\widehat{\theta}) \leq \widehat{r}_{\mathcal{P}}(\widehat{\theta}) + \sum_i ||f(\mathbf{x}_i;\widehat{\theta}) - f(\mathbf{x}_i';\widehat{\theta})||_1 + \phi(|\Theta|, n, \delta) \tag{12}$$

and $\mathbf{x}' = a(\mathbf{x})$, where $a = \arg\max_{a \in \mathcal{A}} \mathbf{y}^\top f(a(\mathbf{x});\widehat{\theta})$.

Proof. First of all, in the context of multiclass classification, where $g(f(\mathbf{x},;\theta))$ predicts a label with one-hot representation, and $\mathbf{y}$ is also represented with one-hot representation, we can have the empirical risk written as:

$$\widehat{r}_{\mathcal{P}}(\mathbf{x};\widehat{\theta}) = 1 - \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \sim \mathcal{P}} \mathbf{y}^\top g(f(\mathbf{x};\widehat{\theta}))$$

Thus,

$$\begin{aligned}
\sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \widehat{r}_{\mathcal{P}'}(\mathbf{x};\widehat{\theta}) &= \widehat{r}_{\mathcal{P}}(\mathbf{x};\widehat{\theta}) + \sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \widehat{r}_{\mathcal{P}'}(\mathbf{x};\widehat{\theta}) - \widehat{r}_{\mathcal{P}}(\mathbf{x};\widehat{\theta}) \\
&= \widehat{r}_{\mathcal{P}}(\mathbf{x};\widehat{\theta}) + \frac{1}{n} \sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \Big(\sum_{(\mathbf{x},\mathbf{y}) \sim \mathcal{P}} \mathbf{y}^\top g(f(\mathbf{x};\widehat{\theta})) - \sum_{(\mathbf{x},\mathbf{y}) \sim \mathcal{P}'} \mathbf{y}^\top g(f(\mathbf{x};\widehat{\theta})) \Big)
\end{aligned}$$

With A6, we can continue with:

$$\sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \widehat{r}_{\mathcal{P}'}(\mathbf{x};\widehat{\theta}) \leq \widehat{r}_{\mathcal{P}}(\mathbf{x};\widehat{\theta}) + \frac{1}{n} \sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \Big(\sum_{(\mathbf{x},\mathbf{y}) \sim \mathcal{P}} \mathbf{y}^\top \log(f(\mathbf{x};\widehat{\theta})) - \sum_{(\mathbf{x},\mathbf{y}) \sim \mathcal{P}'} \mathbf{y}^\top \log(f(\mathbf{x};\widehat{\theta})) \Big)$$

If we use $e(\cdot) = -\mathbf{y}^\top \log(\cdot)$ to replace the cross-entropy loss, we simply have:

$$\sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \widehat{r}_{\mathcal{P}'}(\mathbf{x};\widehat{\theta}) \leq \widehat{r}_{\mathcal{P}}(\mathbf{x};\widehat{\theta}) + \frac{1}{n} \sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \Big(\sum_{(\mathbf{x},\mathbf{y}) \sim \mathcal{P}'} e(f(\mathbf{x};\widehat{\theta})) - \sum_{(\mathbf{x},\mathbf{y}) \sim \mathcal{P}} e((f(\mathbf{x};\widehat{\theta}))) \Big)$$

Since $e(\cdot)$ is a Lipschitz function with constant ≤ 1 (because of A6, Eq.(9)) and together with the dual representation of Wasserstein metric (See *e.g.*, Villani (2003)), we have

$$\sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \widehat{r}_{\mathcal{P}'}(\mathbf{x};\widehat{\theta}) \leq \widehat{r}_{\mathcal{P}}(\mathbf{x};\widehat{\theta}) + W_1(f(\mathbf{x},\widehat{\theta}), f(\mathbf{x}',\widehat{\theta}))$$

where $\mathbf{x}' = a(\mathbf{x})$, where $a = \arg\max_{a \in \mathcal{A}} \mathbf{y}^\top f(a(\mathbf{x});\widehat{\theta})$.

Further, we can use the help of Proposition 3.2 to replace Wassertein metric with ℓ_1 distance. Finally, we can conclude the proof with Assumption A5 as how we did in the proof of Lemma 3.1. □

B.4 PROOF OF LEMMA 3.4

Lemma. *With Assumptions A1-A6, and $d_e(\cdot,\cdot)$ in A2 chosen as ℓ_1 norm distance, $d_x(\cdot,\cdot)$ in A3 chosen as Wasserstein-1 metric, assuming there is a $a'() \in \mathcal{A}$ where $\widehat{r}_{\mathcal{P}_{a'}}(\widehat{\theta}) = \frac{1}{2}\big(\widehat{r}_{\mathcal{P}_{a^+}}(\widehat{\theta}) + \widehat{r}_{\mathcal{P}_{a^-}}(\widehat{\theta})\big)$, with probability at least $1-\delta$, we have:*

$$\sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} r_{\mathcal{P}'}(\widehat{\theta}) \leq \frac{1}{2}\big(\widehat{r}_{\mathcal{P}_{a^+}}(\widehat{\theta}) + \widehat{r}_{\mathcal{P}_{a^-}}(\widehat{\theta})\big) + \sum_i ||f(a^+(\mathbf{x}_i);\widehat{\theta}) - f(a^-(\mathbf{x}');\widehat{\theta})||_1 + \phi(|\Theta|, n, \delta) \tag{13}$$

Proof. We can continue with

$$\sup_{\mathcal{P}' \in T(\mathcal{P},\mathcal{A})} \widehat{r}_{\mathcal{P}'}(\mathbf{x};\widehat{\theta}) \leq \widehat{r}_{\mathcal{P}}(\mathbf{x};\widehat{\theta}) + W_1(f(\mathbf{x},\widehat{\theta}), f(\mathbf{x}',\widehat{\theta}))$$

from the proof of Lemma 3.3. With the help of Assumption A3, we have:

$$d_x(f(a^+(\mathbf{x}), \widehat{\theta}), f(a^-(\mathbf{x}), \widehat{\theta})) \geq d_x(f(\mathbf{x}, \widehat{\theta}), f(\mathbf{x}', \widehat{\theta}))$$

When $d_x(\cdot, \cdot)$ is chosen as Wasserstein-1 metric, we have:

$$\sup_{\mathcal{P}' \in T(\mathcal{P}, \mathcal{A})} \widehat{r}_{\mathcal{P}'}(\mathbf{x}; \widehat{\theta}) \leq \widehat{r}_{\mathcal{P}}(\mathbf{x}; \widehat{\theta}) + W_1(f(a^+(\mathbf{x}), \widehat{\theta}), f(a^-(\mathbf{x}), \widehat{\theta}))$$

Further, as the LHS is the worst case risk generated by the transformation functions within $\mathcal{A}$, and $\widehat{r}_{\mathcal{P}}(\mathbf{x}; \widehat{\theta})$ is independent of the term $W_1(f(a^+(\mathbf{x}), \widehat{\theta}), f(a^-(\mathbf{x}), \widehat{\theta}))$, WLOG, we can replace $\widehat{r}_{\mathcal{P}}(\mathbf{x}; \widehat{\theta})$ with the risk of an arbitrary distribution generated by the transformation function in $\mathcal{A}$. If we choose to use $\widehat{r}_{\mathcal{P}_{a'}}(\widehat{\theta}) = \frac{1}{2}\big(\widehat{r}_{\mathcal{P}_{a^+}}(\widehat{\theta}) + \widehat{r}_{\mathcal{P}_{a^-}}(\widehat{\theta})\big)$, we can conclude the proof, with help from Proposition 3.2 and Assumption A5 as how we did in the proof of Lemma 3.3. □

	Vanilla Scenario			Challenging Scenario		
	Vanilla	Augmented	Regularized	Vanilla	Augmented	Regularized
Frequency	0.005	0.152	0.999	0.001	0.021	0.711
Paired Distance	217968.06	42236.75	1084.4	66058.4	28122.45	4287.31
Wasserstein (greedy)	152736.47	38117.77	1084.4	37156.5	20886.7	4218.53
Paired/Wasserstein	1.42	1.10	1	1.77	1.34	1.02

Table 5: Empirical results from synthetic data for Assumption A2.

C SYNTHETIC RESULTS TO VALIDATE ASSUMPTIONS

We test the assumptions introduced in this paper over MNIST data and rotations as the variation of the data.

Assumption A2: We first inspect Assumption A2, which essentially states the distance $d_e(\cdot, \cdot)$ is the smaller between a sample and its augmented copy (60° rotation) than the sample and the augmented copy from any other samples. We take 1000 training examples and calculate the ℓ_1 pairwise distances between the samples and its augmented copies, then we calculated the frequencies when the A2 hold for one example. We repeat this for three different models, the vanilla model, the model trained with augmented data, and the model trained with regularized adversarial training. The results are shown in the Table 5 and suggest that, although the A2 does not hold in general, it holds for regularized adversarial training case, where A2 is used. Further, we test the assumption in a more challenging case, where half of the training samples are 15° rotations of the other half, thus we may expect the A2 violated for every sample. Finally, as A2 is essentially introduced to replace the empirical Wasserstein distance with ℓ_1 distances of the samples and the augmented copies, we directly compare these metrics. However, as the empirical Wasserstein distance is forbiddingly hard to calculate (as it involves permutation statistics), we use a greedy heuristic to calculate by iteratively picking the nearest neighbor of a sample and then remove the neighbor from the pool for the next sample. Our inspection suggests that, even in the challenging scenario, the paired distance is a reasonably good representative of Wasserstein distance for regularized adversarial training method.

Assumption A3: Whether Assumption A3 hold will depend on the application and the domain knowledge of vertices, thus here we only discuss the general performances if we assume A3 hold. Conveniently, this can be shown by comparing the performances of RA and the rest methods in the experiments reported in Section 4.1: out of six total scenarios ($\{$texture, rotation, contrast$\}$ $\times$ $\{$MNIST, CIFAR10$\}$), there are four scenarios where RA outperforms VWA, this suggests that the domain-knowledge of vertices can actually help in most cases, although not guaranteed in every case.

Assumption A6: We inspect Assumption A6 by directly calculating the frequencies out of all the samples when it holds. Given a vanilla model (Base), we notice that over 74% samples fit this assumption.

	Texture			Rotation			Contrast		
	C	R	I	C	R	I	C	R	I
Base	**0.7013**	0.3219	0.714	0.7013	0.0871	0.5016	0.7013	0.2079	0.34
VA	0.6601	0.5949	0.9996	0.7378	0.4399	0.6168	0.7452	0.6372	0.4406
RA	0.6571	0.6259	**1**	0.6815	0.5166	0.852	**0.7742**	0.6325	**0.535**
VWA	0.6049	0.5814	**1**	0.714	0.6009	0.9172	0.7387	**0.6708**	0.479
RWA	0.663	**0.6358**	**1**	**0.7606**	**0.6486**	**0.9244**	0.7489	0.6326	0.3736

Table 6: Results of CIFAR10 data. ("C" stands for clean accuracy, "R" stands for robustness, and "I" stands for invariance score): invariance score shows big differences while accuracy does not.

D Additional Details of Synthetic Experiments Setup

We consider three different sets of transformation functions:

- Texture: we use Fourier transform to perturb the texture of the data by discarding the high-frequency components of the given a radius r. The smaller r is, the less high-frequency components the image has. We consider $\mathcal{A} = \{a(), a_{12}(), a_{10}(), a_8(), a_6()\}$, where $a()$ is the identity map. Thus, vertexes are $a()$ and $a_6()$.
- Rotation: we rotate the images clockwise r degrees. We consider $\mathcal{A} = \{a(), a_{15}(), a_{30}(), a_{45}(), a_{60}()\}$, where $a()$ is the identity map. Thus, vertexes are $a()$ and $a_{60}()$.
- Contrast: we create the images depicting the same semantic information, but with different scales of the pixels, including the negative color representation. Therefore, we have $\mathcal{A} = \{a(\mathbf{x}) = \mathbf{x}, a_1(\mathbf{x}) = \mathbf{x}/2, a_2(\mathbf{x}) = \mathbf{x}/4, a_3(\mathbf{x}) = 1 - \mathbf{x}, a_4(\mathbf{x}) = (1 - \mathbf{x})/2, a_5(\mathbf{x}) = (1 - \mathbf{x})/4$, where $\mathbf{x}$ stands for the image whose pixel values have been normalized to be between 0 and 1. We consider $a()$ and $a_3()$ as vertexes.

We first train the baseline models to get reasonably high performance, and then train other augmented models with the same hyperparameters. VA and RA are augmented with vertexes, while VWA and RWA are augmented with $\mathcal{A}$. For methods with a regularizer, we run the experiments with 9 hyperparameters evenly split in the logspace from 10^{-4} to 10^4, and we report the methods with the best worst-case accuracy.

Results Discussion Table 6 tells roughly the same story with Table 1. The invariance score of the worst case methods in Table 6 behave lower than we expected, we conjecture this is mainly because some elements in $\mathcal{A}$ of "contrast" will transform the data into samples inherently hard to predict (*e.g.* $a(\mathbf{x}) = \mathbf{x}/4$ will squeeze the pixel values together, so the images look blurry in general and hard to recognize), the model repeatedly identifies these case as the worst case and ignores the others. As a result, RWA effectively degrades to RA yet is inferior to RA because it does not have the explicit vertex information. To verify the conjecture, we count how often each augmented sample to be considered as the worst case: for "texture" and "rotation", each augmented sample generated by $\mathcal{A}$ are picked up with an almost equal frequency, while for "constrast", $\mathbf{x}/2$ and $(1-\mathbf{x})/2$ are identified only 10%-15% of the time $\mathbf{x}/4$ and $(1 - \mathbf{x})/4$ are identified as the worst case.

	Worst	Clean	Vertex	All	Beyond	Invariance
Base	0.9860	0.9921		0.9911	0.9463	0.9236
VA	0.9906	**0.9928**	**0.9925**	**0.9927**	0.9650	0.9876
RA	0.9904	0.9909	0.9910	0.9909	0.9747	1
VWA	0.9903	0.9922		0.9923	0.9696	0.9940
RWA	**0.9911**	0.9915		0.9915	**0.9773**	**1**
RA-ℓ_1	0.9897	0.9904	0.9901	0.9903	0.9728	**1**
RA-W	0.9858	0.9888	0.9902	0.9893	0.9433	0.6428
RA-D	0.9892	0.9921	0.9912	0.9919	0.9373	0.2588
RA-KL	0.0980	0.0980	0.0980	0.0980	0.0980	0.2800
$\text{RA}_{\text{softmax}}$	0.9898	0.9917	0.9919	0.9920	0.9633	0.9928
$\text{RA}_{\text{softmax}}$-$\ell_1$	0.9904	0.9925	0.9918	0.9925	0.9672	0.9960

Table 7: More methods tested with more comprehensive metrics over MNIST on texture

	Worst	Clean	Vertex	All	Beyond	Invariance
Base	0.2960	0.9921		0.7410	0.8914	0.2056
VA	0.9336	0.9884	0.9886	0.9775	0.8711	0.5628
RA	0.9525	0.9930	0.9919	0.9829	0.9201	0.6044
VWA	0.9408	0.9466		0.9827	0.5979	0.6284
RWA	**0.9882**	**0.9934**		**0.9934**	**0.9417**	**0.8856**
RA-ℓ_1	0.9532	0.9913	0.9916	0.9824	0.9145	0.5912
RA-W	0.9274	0.9882	0.9875	0.9757	0.8514	0.4600
RA-D	0.9368	0.9895	0.989	0.9782	0.8431	0.4132
RA-KL	0.9424	0.9875	0.9872	0.9762	0.9194	0.6800
$\text{RA}_{\text{softmax}}$	0.9389	0.9900	0.9901	0.9792	0.8631	0.6060
$\text{RA}_{\text{softmax}}$-$\ell_1$	0.9424	0.9913	0.9901	0.9804	0.8663	0.5864

Table 8: More methods tested with more comprehensive metrics over MNIST on rotation.

	Worst	Clean	Vertex	All	Beyond	Invariance
Base	0.2699	0.9921		0.6377	0.2988	0.2003
VA	0.9837	0.9922	0.9917	0.9913	0.6044	0.4153
RA	0.9823	0.9936	0.9930	0.9911	0.6512	0.4166
VWA	0.4470	0.5360		0.7515	0.4649	0.2210
RWA	**0.9893**	**0.9940**		**0.9930**	0.4841	**0.8786**
RA-ℓ_1	0.9776	0.9935	0.9932	0.9902	0.6251	0.4176
RA-W	0.7357	0.9867	0.9865	0.9361	**0.6547**	0.2960
RA-D	0.9833	0.9913	0.9921	0.9909	0.6199	0.2000
RA-KL	0.9105	0.9894	0.9882	0.9677	0.6001	0.4153
$\text{RA}_{\text{softmax}}$	0.9839	0.9916	0.9910	0.9906	0.6221	0.4273
$\text{RA}_{\text{softmax}}$-$\ell_1$	0.9844	0.9920	0.9918	0.9909	0.5843	0.4236

Table 9: More methods tested with more comprehensive metrics over MNIST on contrast.

E More Synthetic Results

E.1 Experiment Setup

To understand these methods, we introduce a more comprehensive test of these methods, including the five methods discussed in the main paper, and multiple ablation test methods, including

- RA-ℓ_1: when squared ℓ_2 norm of RA is replaced by ℓ_1 norm.
- RA-W: when the norm distance of RA is replaced by Wasserstein distance, enabled by the implementation of Wasserstein GAN Arjovsky et al. (2017); Gulrajani et al. (2017).
- RA-D: when the norm distance of RA is replaced by a discriminator. Our implementation uses a one-layer neural network.

	Worst	Clean	Vertex	All	Beyond	Invariance
Base	0.3219	0.7013		0.5997	0.3084	0.7140
VA	0.5949	0.6601	0.6394	0.6530	0.5583	0.9996
RA	0.6259	0.6571	0.6485	0.6553	0.5826	**1**
VWA	0.5814	0.6049		0.6024	0.5213	**1**
RWA	**0.6358**	0.6630		0.6612	**0.5892**	**1**
RA-ℓ_1	0.6230	0.6609	0.6511	0.6578	0.5775	**1**
RA-W	0.6140	0.6860	0.6578	0.6783	0.5801	**1**
RA-D	0.5794	**0.7663**	**0.6734**	**0.7288**	0.5632	0.3220
RA-KL	0.5866	0.5873	0.5868	0.5870	0.5804	**1**
$\text{RA}_{\text{softmax}}$	0.6197	0.6263	0.6268	0.6266	0.5831	**1**
$\text{RA}_{\text{softmax}}$-$\ell_1$	0.6319	0.653	0.6480	0.6516	0.5830	**1**

Table 10: More methods tested with more comprehensive metrics over CIFAR10 on texture

	Worst	Clean	Vertex	All	Beyond	Invariance
Base	0.0871	0.7013		0.4061	0.4634	0.5016
VA	0.4399	0.7378	0.7199	0.6835	**0.5096**	0.6168
RA	0.5166	0.6815	0.6741	0.6452	0.4408	0.8520
VWA	0.6009	0.7140		0.7406	0.4446	0.9172
RWA	**0.6486**	**0.7606**		**0.7507**	0.4614	**0.9244**
RA-ℓ_1	0.4685	0.7505	0.7290	0.6852	0.4878	0.6248
RA-W	0.4228	0.7468	0.7287	0.6822	0.4753	0.6072
RA-D	0.4298	**0.7752**	0.7456	0.6941	0.4662	0.2664
RA-KL	0.5848	0.4241	0.4221	0.4211	0.3946	0.9200
$\text{RA}_{\text{softmax}}$	0.5143	0.7187	0.7175	0.6851	0.4694	0.8188
$\text{RA}_{\text{softmax}}$-$\ell_1$	0.4779	0.7341	0.725	0.6911	0.4944	0.7288

Table 11: More methods tested with more comprehensive metrics over CIFAR10 on rotation.

- RA-KL: when the norm distance of RA is replaced by KL divergence.
- $\text{RA}_{\text{softmax}}$: when the regularization of RA is applied to softmax instead of logits.
- $\text{RA}_{\text{softmax}}$-$\ell_1$: when the regularization of RA is applied to softmax instead of logits, and the squared ℓ_2 norm is replaced by ℓ_1 norm. This is the method suggested by pure theoretical discussion if we do not concern with the difficulties of passing gradient through backpropagation.

And we test these methods in the three scenarios mentioned in the previous section: texture, rotation, and contrast. The overall test follows the same regime as the one reported in the main manuscript, with additional tests:

- Vertex: average test performance on the perturbed samples with the vertex function from $\mathcal{A}$. Models with worst case augmentation are not tested with vertex as these models do not have the specific concept of vertex.
- All: average test performance on all the samples perturbed by all the elements in $\mathcal{A}$.
- Beyond: To have some sense of how well the methods can perform in the setting that follows the same concept, but not considered in $\mathcal{A}$, and not (intuitively) limited by the verteics of $\mathcal{A}$, we also test the accuracy of the models with some transformations related to the elements in $\mathcal{A}$, but not in $\mathcal{A}$, To be specific:
  - Texture: $\mathcal{A}_{\text{beyond}} = \{a_5(), a_4()\}$.
  - Rotation: $\mathcal{A}_{\text{beyond}} = \{a_3 30(), a_3 45()\}$.
  - Contrast: $\mathcal{A}_{\text{beyond}} = \{a(\mathbf{x}) = \mathbf{x}/2 + 0.5, a(\mathbf{x}) = \mathbf{x}/4 + 0.75, a(\mathbf{x}) = (1 - \mathbf{x})/2 + 0.5, a(\mathbf{x}) = (1 - \mathbf{x})/4 + 0.75\}$

  We report the average test accuracy of the samples tested all the elements in $\mathcal{A}_{\text{beyond}}$

	Worst	Clean	Vertex	All	Beyond	Invariance
Base	0.2079	0.7013		0.4793	0.2605	0.3400
VA	0.6372	0.7452	0.7243	0.7365	0.3733	0.4406
RA	0.6867	**0.7742**	**0.7702**	**0.7722**	0.5527	0.5350
VWA	0.6708	0.7387		0.7375	0.5539	0.4790
RWA	0.6326	0.7489		0.7246	0.4789	0.3736
RA-ℓ_1	**0.7096**	0.7688	0.7634	0.7666	**0.7330**	**0.6260**
RA-W	0.6325	0.7442	0.7303	0.7364	0.4994	0.4396
RA-D	0.6451	0.7515	0.7392	0.7479	0.4820	0.2393
RA-KL	0.1137	0.4515	0.4517	0.3317	0.2648	0.5026
$\text{RA}_{\text{softmax}}$	0.6856	0.7618	0.7558	0.7609	0.6531	0.4833
$\text{RA}_{\text{softmax}}$-$\ell_1$	0.6895	0.7585	0.7533	0.7581	0.7000	0.4946

Table 12: More methods tested with more comprehensive metrics over CIFAR10 on contrast.

E.2 RESULTS

We report the results in Table 7-12.

Ablation Study First we consider the ablation study to validate our choice as the squared ℓ_2 norm regularization, particularly because our choice considers both the theoretical arguments and practical arguments regarding gradients. In case of worst-case prediction, we can see the other RA variants can barely outperform RA, even not the one that our theoretical arguments directly suggest ($\text{RA}_{\text{softmax}}$-$\ell_1$ or RA-W). We believe this is mostly due to the challenges of passing the gradient with ℓ_1 norm and softmax, or through a classifier.

We also test the performances of other regularizations that are irrelevant to our theoretical studies, but are popular choices in general (RA-D and RA-KL). These methods in general perform badly, can barely match RA in terms of the worst-case performance. Further, when some cases when RA-D and RA-KL can outperform RA in other accuracy-wise testing, these methods tend to behave terribly in invariance test, which suggests these regularizations are not effective. In the cases when RA-D and RA-KL can match RA in invariance test, these methods can barely compete with RA.

Broader Test We also test our methods in the broader test. As we can see, RWA behaves the best in most of the cases. In three out of these six test scenarios, RWA lost to three other different methods in the "beyond" case. However, we believe, in general, this is still a strong evidence to show that RWA is a generally preferable method.

Also, comparing the methods of RA vs. VA, and RWA vs. VWA, we can see that regularization helps mostly in the cases of "beyond" in addition to "invariance" test. This result again suggests the importance of regularizations, as in practice, training phase is not always aware of all the transformation functions during test phase.

	300	315	330	345	0	15	30	45	60	avg.
Base	0.2196	0.2573	0.3873	0.6502	0.8360	0.6938	0.4557	0.3281	0.2578	0.4539
ST	0.2391	0.2748	0.4214	0.7049	0.8251	0.7147	0.4398	0.2838	0.2300	0.4593
GC	0.1540	0.1891	0.2460	0.3919	0.5859	0.4145	0.2534	0.1827	0.1507	0.2853
ETN	0.3855	**0.4844**	**0.6324**	**0.7576**	0.8276	0.7730	0.7324	0.6245	0.5060	0.6358
VA	0.2233	0.2832	0.4318	0.6364	0.8124	0.6926	0.5973	0.7152	0.7923	0.5761
RA	0.3198	0.3901	0.5489	0.7170	0.8487	0.7904	0.7455	0.8005	0.8282	0.6655
VWA	0.3383	0.3484	0.3835	0.4569	0.7474	0.866	0.8776	0.8738	0.8629	0.6394
RWA	**0.4012**	0.4251	0.4852	0.6765	**0.8708**	**0.8871**	**0.8869**	**0.8870**	**0.8818**	**0.7113**

Table 13: Comparison to advanced rotation-invariant models. We report the test accuracy on the test sets clockwise rotated, $0°$-$60°$ and $300°$-$360°$. Average accuracy is also reported. Augmentation methods only consider $0°$-$60°$ clockwise rotations during training.

	Clean	Noise			Blur				Weather				Digital				mCE
		Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	
Base	23.9	79	80	82	82	90	84	80	86	81	75	65	79	91	77	80	80.6
VA	23.7	79	80	79	75	87	80	79	78	76	69	58	70	86	73	75	76.3
RA	23.6	78	78	79	74	87	79	76	78	75	69	58	68	85	75	75	75.6
RWA	23.1	76	77	78	71	86	76	75	75	73	66	55	68	83	76	73	73.9
VWA	22.4	61	63	63	68	75	65	66	70	69	64	56	55	70	61	63	64.6
SU	24.5	67	68	70	74	83	81	77	80	74	75	62	77	84	71	71	74.3
AA	22.8	69	68	72	77	83	80	81	79	75	64	56	70	88	57	71	72.7
MBP	23	73	74	76	74	86	78	77	77	72	63	56	68	86	71	71	73.4
SIN	27.2	69	70	70	77	84	76	82	74	75	69	65	69	80	64	77	73.3
AM	22.4	65	66	67	70	80	66	66	75	72	67	58	58	79	69	69	68.4
AMS	25.2	61	62	61	69	77	63	72	66	68	63	59	52	74	60	67	64.9

Table 14: Comparison to advanced models over ImageNet-C data. Performance reported (mCE) follows the standard in ImageNet-C data: mCE is the smaller the better.

F Additional Discussions for Comparisons with Advanced Methods

Rotation-invariant Image Classification We compare our results with specifically designed rotation-invariant models, mainly Spatial Transformer (ST) (Jaderberg et al., 2015), Group Convolution (GC) (Cohen & Welling, 2016), and Equivariant Transformer Network (ETN) (Tai et al., 2019). We also attempted to run CGNet (Kondor et al., 2018), but the procedure does not scale to the CIFAR10 and ResNet level. The results are reported in Table 13, where most methods use the same architecture (ResNet34 with most performance boosting heuristics enabled), except that GC uses ResNet18 because ResNet34 with GC runs 100 times slowly than others, thus not practical. We test the models with nine different rotations including $0°$ degree rotation. Augmentation related methods are using the $\mathcal{A}$ of "rotation" in synthetic experiments (Appendix D), so the testing scenario goes beyond what the augmentation methods have seen during training. The results in Table 13 strongly endorses the efficacy of augmentation-based methods. Interestingly, regularized augmentation methods, with the benefit of learning the concept of invariance, tend to behave well in the transformations not considered during training. As we can see, RA outperforms VWA on average.

Texture-perturbed ImageNet classification We also test the performance on the image classification over multiple perturbations. We train the model over standard ImageNet training set and test the model with ImageNet-C data Hendrycks & Dietterich (2019), which is a perturbed version of ImageNet by corrupting the original ImageNet validation set with a collection of noises. Following the standard, the reported performance is mCE, which is the smaller the better. We compare with several methods tested on this dataset, including Patch Uniform (PU) Lopes et al. (2019), AutoAugment (AA) Cubuk et al. (2019), MaxBlur pool (MBP) Zhang (2019), Stylized ImageNet (SIN) Hendrycks & Dietterich (2019), AugMix (AM) Hendrycks et al. (2020), AugMix w. SIN (AMS) Hendrycks et al. (2020). We use the performance reported in Hendrycks et al. (2020). Again, our augmention only uses the generic texture with perturbation (the $\mathcal{A}$ in our texture synthetic experiments with radius changed to $20, 25, 30, 35, 40$). The results are reported in Table 14, which shows that our generic method outperform the current SOTA methods after a continued finetuning process with reducing learning rates.

	Base	InfoDrop	HEX	PAR	VA	RA	VWA	RWA
Top-1	0.1204	0.1224	0.1292	0.1306	0.1362	0.1405	0.1432	**0.1486**
Top-5	0.2408	0.256	0.2564	0.2627	0.2715	0.2793	0.2846	**0.2933**

Table 15: Comparison to advanced cross-domain image classification models, over ImageNet-Sketch dataset. We report top-1 and top-5 accuracy following standards on ImageNet related experiments.

Cross-domain ImageNet-Sketch Classification We also compare to the methods used for cross-domain evaluation. We follow the set-up advocated by (Wang et al., 2019b) for domain-agnostic cross-domain prediction, which is training the model on one or multiple domains without domain identifiers and test the model on an unseen domain. We use the most challenging setup in this scenario: train the models with standard ImageNet training data, and test the model over ImageNet-Sketch data (Wang et al., 2019a), which is a collection of sketches following the structure ImageNet validation set. We compare with previous methods with reported performance on this dataset, such as InfoDrop (Achille & Soatto, 2018), HEX (Wang et al., 2019b), and PAR (Wang et al., 2019a), and report the performances in Table 15. Notice that, our data augmentation also follows the requirement that the characteristics of the test domain cannot be utilized during training. Thus, we only augment the samples with a generic augmentation set ($\mathcal{A}$ of "contrast" in synthetic experiments, Appendix D). The results again support the strength of the correct usage of data augmentation.