A. Proof of Theorem 1

First we bound $|\hat{\epsilon}_\alpha(h) - \epsilon_\alpha(h)|$ with high probability and uniformly over $\mathcal{H}$. We adapt the classical proofs of generalization bounds in terms of the Rademacher complexity of a hypothesis class, e.g. (Bousquet et al., 2004).

Proposition 1. *Given the setup and assumptions described above, for any $\delta > 0$ with probability at least $1-\delta$ over the data, for any function $h \in \mathcal{H}$:*

$$|\epsilon_\alpha(h) - \hat{\epsilon}_\alpha(h)| \leq 2\sum_{i=1}^{N} \alpha_i \mathcal{R}_i(\mathcal{H}) + 3\sqrt{\frac{\log\left(\frac{4}{\delta}\right) M^2}{2}} \sqrt{\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}}, \tag{7}$$

where for each $i = 1, 2, \ldots, N$:

$$\mathcal{R}_i(\mathcal{H}) = \mathbb{E}_\sigma \left(\sup_{f \in \mathcal{H}} \left(\frac{1}{m_i} \sum_{j=1}^{m_i} \sigma_{i,j} L(f(x_{i,j}), y_{i,j}) \right) \right), \tag{8}$$

and where $\sigma_{i,j}$ are independent Rademacher random variables.

Proof. Write:

$$\epsilon_\alpha(h) \leq \hat{\epsilon}_\alpha(h) + \sup_{f \in \mathcal{H}} \left(\epsilon_\alpha(f) - \hat{\epsilon}_\alpha(f) \right) \tag{9}$$

To link the second term to its expectation, we prove the following:

Lemma 1. *Define the function $\phi : (\mathcal{X} \times \mathcal{Y})^m \to \mathbb{R}$ by:*

$$\phi(\{x_{1,1}, y_{1,1}\}, \ldots, \{x_{N,m_N}, y_{N,m_N}\}) = \sup_{f \in \mathcal{H}} \left(\epsilon_\alpha(f) - \hat{\epsilon}_\alpha(f) \right).$$

Denote for brevity $z_{i,j} = \{x_{i,j}, y_{i,j}\}$. Then, for any $i \in \{1, 2, \ldots, N\}, j \in \{1, 2, \ldots, m_i\}$:

$$\sup_{z_{1,1}, \ldots, z_{N,m_N}, z'_{i,j}} |\phi(z_{1,1}, \ldots, z_{i,j}, \ldots, z_{N,m_N}) - \phi\left(z_{1,1}, \ldots, z'_{i,j}, \ldots, z_{N,m_N}\right)| \leq \frac{\alpha_i}{m_i} M \tag{10}$$

Proof. Fix any i, j and any $z_{1,1}, \ldots, z_{N,m_N}, z'_{i,j}$. Denote the α-weighted empirical average of the loss with respect to the sample $z_{1,1}, \ldots, z'_{i,j}, \ldots, z_{N,m_N}$ by ϵ'_α. Then we have that:

$$\begin{aligned}
&|\phi(\ldots, z_{i,j}, \ldots) - \phi(\ldots, z'_{i,j}, \ldots)| \\
&\quad = |\sup_{f \in \mathcal{H}} \left(\epsilon_\alpha(f) - \hat{\epsilon}_\alpha(f) \right) - \sup_{f \in \mathcal{H}} (\epsilon_\alpha(f) - \hat{\epsilon}'_\alpha(f))| \\
&\quad \leq |\sup_{f \in \mathcal{H}} (\hat{\epsilon}'_\alpha(f) - \hat{\epsilon}_\alpha(f))| \\
&\quad = \frac{\alpha_i}{m_i} |\sup_{f \in \mathcal{H}} \left(L\left(f(x'_{i,j}), y'_{i,j}\right) - L\left(f(x_{i,j}), y_{i,j}\right) \right)| \\
&\quad \leq \frac{\alpha_i}{m_i} M
\end{aligned}$$

Note: the inequality we used above holds for bounded functions inside the supremum. □

Let S denote a random sample of size m drawn from a distribution as the one generating out data (i.e. m_i samples from $\mathcal{D}_i$ for each i). Now, using Lemma 1, McDiarmid's inequality gives:

$$\begin{aligned}
\mathbb{P}\left(\phi(S) - \mathbb{E}(\phi(S)) \geq t\right) &\leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{\alpha_i^2}{m_i^2} M^2} \right) \\
&= \exp\left(-\frac{2t^2}{M^2 \sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}} \right)
\end{aligned}$$

For any $\delta > 0$, setting the right-hand side above to be $\delta/4$ and using (9), we obtain that with probability at least $1 - \delta/4$:

$$\epsilon_\alpha(h) \leq \hat{\epsilon}_\alpha(h) + \mathbb{E}_S \left(\sup_{f \in \mathcal{H}} \left(\epsilon_\alpha(f) - \hat{\epsilon}_\alpha(f) \right) \right) + \sqrt{\frac{\log\left(\frac{4}{\delta}\right) M^2}{2}} \sqrt{\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}} \tag{11}$$

To deal with the expected loss inside the second term, introduce a ghost sample (denoted by S'), drawn from the same distributions as our original sample (denoted by S). Denoting the weighted empirical loss with respect to the ghost sample by ϵ'_α, $\beta_i = m_i/m$ for all i, and using the convexity of the supremum, we obtain:

$$\begin{aligned}
&\mathbb{E}_S \left(\sup_{f \in \mathcal{H}} \left(\epsilon_\alpha(f) - \hat{\epsilon}_\alpha(f) \right) \right) \\
&\quad = \mathbb{E}_S \left(\sup_{f \in \mathcal{H}} \left(\mathbb{E}_{S'} \left(\hat{\epsilon}'_\alpha(f) \right) - \hat{\epsilon}_\alpha(f) \right) \right) \\
&\quad \leq \mathbb{E}_{S,S'} \left(\sup_{f \in \mathcal{H}} \left(\hat{\epsilon}'_\alpha(f) - \hat{\epsilon}_\alpha(f) \right) \right) \\
&\quad = \mathbb{E}_{S,S'} \left(\sup_{f \in \mathcal{H}} \left(\frac{1}{m} \sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{\alpha_i}{\beta_i} \left(L(f(x'_{i,j}), y'_{i,j}) - L(f(x_{i,j}), y_{i,j}) \right) \right) \right)
\end{aligned}$$

Introducing m independent Rademacher random variables and noting that $L(f(x'), y') - L(f(x), y)$ and

$\sigma\left(L(f(x^{'}),y^{'})-L(f(x),y)\right)$ have the same distribution, as long as (x,y) and $(x^{'},y^{'})$ have the same distribution:

$$\begin{aligned}
&\mathbb{E}_S\left(\sup_{f\in\mathcal{H}}\left(\epsilon_\alpha(f)-\hat{\epsilon}_\alpha(f)\right)\right)\\
&\quad\leq \mathbb{E}_{S,S',\sigma}\left(\sup_{f\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{\alpha_i}{\beta_i}\sigma_{i,j}\left(L(f(x^{'}_{i,j}),y^{'}_{i,j})\right.\right.\right.\\
&\qquad\left.\left.\left.-L(f(x_{i,j}),y_{i,j})\right)\right)\right)\\
&\quad\leq \mathbb{E}_{S',\sigma}\left(\sup_{f\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{\alpha_i}{\beta_i}\sigma_{i,j}L(f(x^{'}_{i,j}),y^{'}_{i,j})\right)\right)\\
&\quad+\mathbb{E}_{S,\sigma}\left(\sup_{f\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{\alpha_i}{\beta_i}\left(-\sigma_{i,j}\right)L(f(x_{i,j}),y_{i,j})\right)\right)\\
&\quad=2\mathbb{E}_{S,\sigma}\left(\sup_{f\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{\alpha_i}{\beta_i}\sigma_{i,j}L(f(x_{i,j}),y_{i,j})\right)\right)
\end{aligned}$$

We can now link the last term to the empirical analog of the Rademacher complexity, by using the McDiarmid Inequality (with an observation similar to Lemma 1). Putting this together, we obtain that for any $\delta>0$ with probability at least $1-\delta/2$:

$$\begin{aligned}
\epsilon_\alpha(h)&\leq\hat{\epsilon}_\alpha(h)\\
&\quad+2\mathbb{E}_\sigma\left(\sup_{f\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{\alpha_i}{\beta_i}\sigma_{i,j}L(f(x_{i,j}),y_{i,j})\right)\right)\\
&\quad+3\sqrt{\frac{\log\left(\frac{4}{\delta}\right)M^2}{2}}\sqrt{\sum_{i=1}^{N}\frac{\alpha_i^2}{m_i}}
\end{aligned}\tag{12}$$

Finally, note that:

$$\begin{aligned}
&\mathbb{E}_\sigma\left(\sup_{f\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{\alpha_i}{\beta_i}\sigma_{i,j}L(f(x_{i,j}),y_{i,j})\right)\right)\\
&\quad\leq\mathbb{E}_\sigma\left(\sum_{i=1}^{N}\alpha_i\sup_{f\in\mathcal{H}}\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\sigma_{i,j}L(f(x_{i,j}),y_{i,j})\right)\right)\\
&\quad=\sum_{i=1}^{N}\alpha_i\mathbb{E}_\sigma\left(\sup_{f\in\mathcal{H}}\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\sigma_{i,j}L(f(x_{i,j}),y_{i,j})\right)\right)\\
&\quad=\sum_{i=1}^{N}\alpha_i\mathcal{R}_i\left(\mathcal{H}\right)
\end{aligned}$$

Bounding $\hat{\epsilon}_\alpha(h)-\epsilon_\alpha(h)$ with the same quantity and with probability at least $1-\delta/2$ follows by a similar argument. The result then follows by applying the union bound. □

Now we show:

Theorem 1. *Given the setup above, let* $\hat{h}_\alpha = argmin_{h\in\mathcal{H}}\hat{\epsilon}_\alpha(h)$ *and* $h_T^* = argmin_{h\in\mathcal{H}}\epsilon_T(h)$. *For any* $\delta>0$, *with probability at least* $1-\delta$ *over the data:*

$$\begin{aligned}
\epsilon_T(\hat{h}_\alpha)&\leq\epsilon_T(h_T^*)+4\sum_{i=1}^{N}\alpha_i\mathcal{R}_i\left(\mathcal{H}\right)+2\sum_{i=1}^{N}\alpha_i d_{\mathcal{H}}\left(\mathcal{D}_i,\mathcal{D}_T\right)\\
&\quad+6\sqrt{\frac{\log\left(\frac{4}{\delta}\right)M^2}{2}}\sqrt{\sum_{i=1}^{N}\frac{\alpha_i^2}{m_i}},
\end{aligned}\tag{4}$$

where, for each source $i=1,\ldots,N$,

$$\mathcal{R}_i\left(\mathcal{H}\right)=\mathbb{E}_\sigma\left(\sup_{f\in\mathcal{H}}\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\sigma_{i,j}L(f(x_{i,j}),y_{i,j})\right)\right)$$

and $\sigma_{i,j}$ *are independent Rademacher random variables.*

Proof. For any $h\in\mathcal{H}$:

$$\begin{aligned}
|\epsilon_\alpha(h)-\epsilon_T(h)|&=|\sum_{i=1}^{N}\alpha_i\epsilon_i(h)-\epsilon_T(h)|\\
&\leq\sum_{i=1}^{N}\alpha_i|\epsilon_i(h)-\epsilon_T(h)|\\
&\leq\sum_{i=1}^{N}\alpha_i d_{\mathcal{H}}\left(\mathcal{D}_i,\mathcal{D}_T\right).
\end{aligned}$$

Now applying this bound twice and using Proposition 1, we get that with probability at least $1-\delta$:

$$\begin{aligned}
\epsilon_T(\hat{h}_\alpha)&\leq\epsilon_\alpha(\hat{h}_\alpha)+\sum_{i=1}^{N}\alpha_i d_{\mathcal{H}}\left(\mathcal{D}_i,\mathcal{D}_T\right)\\
&\leq\hat{\epsilon}_\alpha(\hat{h}_\alpha)+2\sum_{i=1}^{N}\alpha_i\mathcal{R}_i\left(\mathcal{H}\right)\\
&\quad+3\sqrt{\frac{\log\left(\frac{4}{\delta}\right)M^2}{2}}\sqrt{\sum_{i=1}^{N}\frac{\alpha_i^2}{m_i}}+\sum_{i=1}^{N}\alpha_i d_{\mathcal{H}}\left(\mathcal{D}_i,\mathcal{D}_T\right)\\
&\leq\hat{\epsilon}_\alpha(h_T^*)+2\sum_{i=1}^{N}\alpha_i\mathcal{R}_i\left(\mathcal{H}\right)\\
&\quad+3\sqrt{\frac{\log\left(\frac{4}{\delta}\right)M^2}{2}}\sqrt{\sum_{i=1}^{N}\frac{\alpha_i^2}{m_i}}+\sum_{i=1}^{N}\alpha_i d_{\mathcal{H}}\left(\mathcal{D}_i,\mathcal{D}_T\right)\\
&\leq\epsilon_\alpha(h_T^*)+4\sum_{i=1}^{N}\alpha_i\mathcal{R}_i\left(\mathcal{H}\right)\\
&\quad+6\sqrt{\frac{\log\left(\frac{4}{\delta}\right)M^2}{2}}\sqrt{\sum_{i=1}^{N}\frac{\alpha_i^2}{m_i}}+\sum_{i=1}^{N}\alpha_i d_{\mathcal{H}}\left(\mathcal{D}_i,\mathcal{D}_T\right)
\end{aligned}$$

$$\leq \epsilon_T(h_T^*) + 4\sum_{i=1}^{N}\alpha_i\mathcal{R}_i\left(\mathcal{H}\right) + 6\sqrt{\frac{\log\left(\frac{4}{\delta}\right)M^2}{2}}\sqrt{\sum_{i=1}^{N}\frac{\alpha_i^2}{m_i}} + 2\sum_{i=1}^{N}\alpha_i d_{\mathcal{H}}\left(\mathcal{D}_i,\mathcal{D}_T\right)$$

□

B. Details about Algorithm 1

B.1. Distribution-independent upper bounds on the Rademacher complexity

Here we give examples of some well-known upper bounds on the Rademacher complexity of certain function classes, which are distribution-independent. Applying such a bound on the Rademacher terms in Theorem 1 will make the dependence of the second term in the bound on the weights disappear. Therefore, we focus on the remaining terms in our algorithm.

Throughout this section, we discuss the Rademacher complexity of a function class $\mathcal{H}$ with respect to a set of samples $\{x_1,\ldots,x_n\}\sim\mathcal{D}$, defined as:

$$\mathcal{R}\left(\mathcal{H}\right) = \mathbb{E}_\sigma\left(\sup_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right)\right) \tag{13}$$

In the case of bounded binary linear classifiers $\mathcal{H}=\{x\to\langle\mathbf{w},\mathbf{x}\rangle : \|\mathbf{w}\|_2\leq B\}$, acting on a bounded domain $\mathcal{X}$ (i.e. for all $x\in\mathcal{X}, \|x\|_2\leq D$), Lemma 26.10 in (Shalev-Shwartz & Ben-David, 2014) shows that:

$$\mathcal{R}\left(\mathcal{H}\right)\leq\frac{BD}{\sqrt{n}}.$$

More generally, the Rademacher complexity of a set of binary classifiers with a finite VC dimention h can be bounded by (Bousquet et al., 2004):

$$\mathcal{R}\left(\mathcal{H}\right)\leq C\sqrt{\frac{h}{n}},$$

for some constant C. The Rademacher complexity is also related to another popular complexity measure, the covering number, via Dudley's entropy bound (Bousquet et al., 2004):

$$\mathcal{R}\left(\mathcal{H}\right)\leq\frac{C}{\sqrt{n}}\int_0^\infty\sqrt{\log N\left(\mathcal{H},t,n\right)}\,dt,$$

where $N\left(\mathcal{H},t,n\right)$ is the size of the smallest t-cover of the space $\mathcal{H}$, under the metric:

$$d_n\left(h,h'\right)=\frac{1}{n}|\{h(x_i)\neq h'(x_i) : i=1,\ldots,n\}|.$$

Note that both the VC-dimension and the covering number are distribution-independent measures of complexity, so the corresponding upper bounds do not depend on $\mathcal{D}$ as well.

B.2. Computing the empirical discrepancies

As explained in Section 3.2, the discrepancies:

$$d_{\mathcal{H}}\left(\mathcal{D}_i,\mathcal{D}_T\right)=\sup_{h\in\mathcal{H}}\left(|\epsilon_i(h)-\epsilon_T(h)|\right). \tag{14}$$

are unknown in practice and therefore need to be estimated from their empirical counterparts:

$$\begin{aligned} d_{\mathcal{H}}\left(S_i,S_T\right) &= \sup_{h\in\mathcal{H}}\left(|\hat{\epsilon}_i(h)-\hat{\epsilon}_T(h)|\right) \\ &= \sup_{h\in\mathcal{H}}(|\frac{1}{m_i}\sum_{j=1}^{m_i}L\left(h\left(x_{i,j}\right),y_{i,j}\right) \\ &\quad -\frac{1}{m_T}\sum_{j=1}^{m_T}L\left(h\left(x_{T,j}\right),y_{T,j}\right)|). \end{aligned} \tag{15}$$

Here, we explain how these are computed in our experiments. Notice that for the $0/1$-loss, a symmetric hypothesis class $\mathcal{H}$ and whenever $\mathcal{Y}=\{-1,+1\}$, we have:

$$\begin{aligned} &d_{\mathcal{H}}\left(S_i,S_T\right) \\ &= \sup_{h\in\mathcal{H}}|\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbb{1}_{\{h(x_{i,j})y_{i,j}<0\}}-\frac{1}{m_T}\sum_{j=1}^{m_T}\mathbb{1}_{\{h(x_{T,j})y_{T,j}<0\}}| \\ &= \sup_{h\in\mathcal{H}}\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbb{1}_{\{h(x_{i,j})y_{i,j}<0\}}-\frac{1}{m_T}\sum_{j=1}^{m_T}\mathbb{1}_{\{h(x_{T,j})y_{T,j}<0\}}\right) \\ &= \sup_{h\in\mathcal{H}}\left(1-(\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbb{1}_{\{h(x_{i,j})\bar{y}_{i,j}<0\}}+\frac{1}{m_T}\sum_{j=1}^{m_T}\mathbb{1}_{\{h(x_{T,j})y_{T,j}<0\}})\right) \\ &= 1-\inf_{h\in\mathcal{H}}\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\mathbb{1}_{\{h(x_{i,j})\bar{y}_{i,j}<0\}}+\frac{1}{m_T}\sum_{j=1}^{m_T}\mathbb{1}_{\{h(x_{T,j})y_{T,j}<0\}}\right), \end{aligned} \tag{16}$$

where $\bar{y}_{i,j}=1-y_{i,j}$ is the flipped label of the j-th data point from the i-th source. Now notice that computing the infimum in equation (16) is equivalent to solving a (weighted) empirical risk minimization problem with the input data from the source and the target merged and the labels being the flipped labels from the source and the actual labels from the target.

Therefore, computing the empirical discrepancies is equivalent to solving an empirical risk minimization problem and standard convex upper bounds can be applied to make the problem tractable. In our experiments, we solve the ERM problem by using square loss.

C. Additional results from experiments

Over the next pages we present more detailed results from the experiments on the Animals with Attributes 2 dataset. The table from the main body of the paper (Table 1) is split according to the type of data corruption. In addition, we performed experiments in which a proportion p of the samples in the n corrupted sources are modified (instead of all of them). Apart from $p = 1$, we experimented with $p = 0.5$ and $p = 0.2$. We present the same type of results for these cases, together with a more detailed breakdown, depending on the type of corruption.

Table 3: Summary of the results for $p = 1$, over all 85 prediction tasks and all corruptions (same as Table 1).

Baseline \ n	$n = 0$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	505/5/0	497/13/0	487/23/0	475/35/0	442/68/0	325/185/0	0/510/0
All data	0/85/0	115/395/0	267/243/0	370/140/0	438/72/0	468/42/0	479/31/0	484/26/0
Median of probs.	9/76/0	47/463/0	172/338/0	336/174/0	469/41/0	504/6/0	502/8/0	499/11/0
(Feng et al., 2014)	8/77/0	32/478/0	110/400/0	338/172/0	457/53/0	504/6/0	502/8/0	497/13/0
(Yin et al., 2018)	14/71/0	179/331/0	390/120/0	432/78/0	472/38/0	502/8/0	503/7/0	497/13/0
(Pregibon, 1982)	55/30/0	308/202/0	361/149/0	416/94/0	437/73/0	455/55/0	470/40/0	485/25/0
Batch norm	0/85/0	107/403/0	317/193/0	416/94/0	446/63/1	478/32/0	487/23/0	482/28/0

Table 4: Summary of results for $p = 1$, split by the type of data corruption

(a) Summary of the results for $p = 1$ and label bias, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	85/0/0	84/1/0	82/3/0	80/5/0	76/9/0	55/30/0	0/85/0
All data	61/24/0	82/3/0	85/0/0	85/0/0	85/0/0	85/0/0	84/1/0
Median of probs.	23/62/0	81/4/0	85/0/0	85/0/0	85/0/0	85/0/0	84/1/0
(Feng et al., 2014)	4/81/0	19/66/0	85/0/0	85/0/0	85/0/0	85/0/0	84/1/0
(Yin et al., 2018)	50/35/0	85/0/0	84/1/0	85/0/0	85/0/0	85/0/0	84/1/0
(Pregibon, 1982)	51/34/0	64/21/0	84/1/0	84/1/0	84/1/0	83/2/0	83/2/0
Batch norm	53/32/0	81/4/0	85/0/0	85/0/0	85/0/0	85/0/0	84/1/0

(b) Summary of the results for $p = 1$ and shuffled labels, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	83/2/0	77/8/0	71/14/0	65/20/0	56/29/0	38/47/0	0/85/0
All data	4/81/0	51/34/0	69/16/0	74/11/0	82/3/0	79/6/0	79/6/0
Median of probs.	2/83/0	24/61/0	63/22/0	83/2/0	80/5/0	79/6/0	80/5/0
(Feng et al., 2014)	3/82/0	0/85/0	51/34/0	77/8/0	80/5/0	79/6/0	80/5/0
(Yin et al., 2018)	32/53/0	63/22/0	62/23/0	76/9/0	80/5/0	79/6/0	80/5/0
(Pregibon, 1982)	57/28/0	63/22/0	70/15/0	71/14/0	75/10/0	73/12/0	71/14/0
Batch norm	3/82/0	41/44/0	60/25/0	63/21/1	79/6/0	79/6/0	79/6/0

(c) Summary of the results for $p = 1$ and shuffled features, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	85/0/0	83/2/0	81/4/0	78/7/0	62/23/0	0/85/0
All data	39/46/0	60/25/0	68/17/0	73/12/0	77/8/0	80/5/0	85/0/0
Median of probs.	6/79/0	30/55/0	84/1/0	85/0/0	85/0/0	85/0/0	85/0/0
(Feng et al., 2014)	10/75/0	72/13/0	84/1/0	85/0/0	85/0/0	85/0/0	85/0/0
(Yin et al., 2018)	18/67/0	76/9/0	84/1/0	85/0/0	85/0/0	85/0/0	85/0/0
(Pregibon, 1982)	56/29/0	58/27/0	67/18/0	74/11/0	78/7/0	85/0/0	85/0/0
Batch norm	40/45/0	66/19/0	72/13/0	77/8/0	78/7/0	79/6/0	80/5/0

(d) Summary of the results for $p = 1$ and blurred images, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	84/1/0	83/2/0	82/3/0	77/8/0	59/26/0	0/85/0
All data	0/85/0	2/83/0	26/59/0	53/32/0	66/19/0	75/10/0	78/7/0
Median of probs.	5/80/0	6/79/0	19/66/0	72/13/0	85/0/0	85/0/0	85/0/0
(Feng et al., 2014)	5/80/0	6/79/0	30/55/0	70/15/0	85/0/0	85/0/0	84/1/0
(Yin et al., 2018)	26/59/0	47/38/0	61/24/0	73/12/0	84/1/0	85/0/0	84/1/0
(Pregibon, 1982)	61/24/0	71/14/0	75/10/0	81/4/0	83/2/0	85/0/0	85/0/0
Batch norm	8/77/0	44/41/0	65/20/0	72/13/0	78/7/0	82/3/0	81/4/0

(e) Summary of the results for $p = 1$ and dead pixels, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	85/0/0	83/2/0	84/1/0	84/1/0	77/8/0	58/27/0	0/85/0
All data	0/85/0	12/73/0	44/41/0	70/15/0	74/11/0	77/8/0	78/7/0
Median of probs.	6/79/0	6/79/0	14/71/0	61/24/0	85/0/0	85/0/0	85/0/0
(Feng et al., 2014)	6/79/0	6/79/0	25/60/0	58/27/0	85/0/0	85/0/0	84/1/0
(Yin et al., 2018)	23/62/0	51/34/0	68/17/0	70/15/0	84/1/0	85/0/0	84/1/0
(Pregibon, 1982)	28/57/0	38/47/0	51/34/0	52/33/0	56/29/0	69/16/0	85/0/0
Batch norm	1/84/0	28/57/0	59/26/0	69/16/0	74/11/0	79/6/0	79/6/0

(f) Summary of the results for $p = 1$ and RGB channels swapped, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	84/1/0	84/1/0	83/2/0	78/7/0	53/32/0	0/85/0
All data	11/74/0	60/25/0	78/7/0	83/2/0	84/1/0	83/2/0	80/5/0
Median of probs.	5/80/0	25/60/0	71/14/0	83/2/0	84/1/0	83/2/0	80/5/0
(Feng et al., 2014)	4/81/0	7/78/0	63/22/0	82/3/0	84/1/0	83/2/0	80/5/0
(Yin et al., 2018)	30/55/0	68/17/0	73/12/0	83/2/0	84/1/0	84/1/0	80/5/0
(Pregibon, 1982)	55/30/0	67/18/0	69/16/0	75/10/0	79/6/0	75/10/0	76/9/0
Batch norm	2/83/0	57/28/0	75/10/0	80/5/0	84/1/0	83/2/0	79/6/0

Table 5: Summary of the results for $p = 0.5$, over all 85 prediction tasks and all corruptions.

Baseline \ n	$n = 0$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	508/2/0	501/9/0	488/22/0	471/39/0	424/86/0	303/207/0	156/354/0
All data	0/85/0	0/510/0	82/428/0	158/352/0	215/295/0	241/269/0	223/287/0	168/342/0
Median of probs.	9/76/0	30/480/0	53/457/0	93/417/0	189/321/0	272/238/0	252/258/0	216/294/0
(Feng et al., 2014)	8/77/0	28/482/0	19/491/0	84/426/0	172/338/0	254/256/0	253/257/0	217/293/0
(Yin et al., 2018)	14/71/0	123/387/0	227/283/0	155/355/0	247/259/4	295/215/0	282/228/0	224/286/0
(Pregibon, 1982)	55/30/0	287/223/0	282/228/0	329/181/0	350/160/0	358/152/0	374/136/0	367/143/0
Batch norm	0/85/0	2/508/0	78/432/0	139/370/1	183/326/1	186/323/1	155/354/1	97/412/1

Table 6: Summary of results for $p = 0.5$, split by the type of data corruption

(a) Summary of the results for $p = 0.5$ and label bias, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	85/0/0	84/1/0	82/3/0	79/6/0	66/19/0	39/46/0	0/85/0
All data	0/85/0	52/33/0	73/12/0	82/3/0	84/1/0	84/1/0	84/1/0
Median of probs.	9/76/0	41/44/0	72/13/0	80/5/0	84/1/0	84/1/0	84/1/0
(Feng et al., 2014)	5/80/0	7/78/0	62/23/0	76/9/0	83/2/0	84/1/0	84/1/0
(Yin et al., 2018)	27/58/0	57/28/0	48/37/0	79/6/0	84/1/0	83/2/0	83/2/0
(Pregibon, 1982)	48/37/0	49/36/0	58/27/0	65/20/0	70/15/0	79/6/0	84/1/0
Batch norm	1/84/0	46/39/0	69/16/0	76/9/0	79/6/0	76/9/0	76/9/0

(b) Summary of the results for $p = 0.5$ and shuffled labels, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	82/3/0	72/13/0	59/26/0	42/43/0	28/57/0	17/68/0
All data	0/85/0	0/85/0	9/76/0	31/54/0	47/38/0	50/35/0	49/36/0
Median of probs.	3/82/0	0/85/0	0/85/0	13/72/0	40/45/0	47/38/0	49/36/0
(Feng et al., 2014)	4/81/0	0/85/0	0/85/0	4/81/0	29/56/0	44/41/0	48/37/0
(Yin et al., 2018)	25/60/0	44/41/0	17/68/0	14/67/4	19/66/0	28/57/0	39/46/0
(Pregibon, 1982)	56/29/0	44/41/0	59/26/0	59/26/0	60/25/0	67/18/0	64/21/0
Batch norm	0/85/0	0/85/0	1/83/1	10/74/1	4/80/1	2/82/1	2/82/1

(c) Summary of the results for $p = 0.5$ and shuffled features, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	85/0/0	83/2/0	82/3/0	76/9/0	44/41/0	0/85/0
All data	0/85/0	26/59/0	51/34/0	57/28/0	57/28/0	41/44/0	0/85/0
Median of probs.	6/79/0	4/81/0	10/75/0	56/29/0	69/16/0	50/35/0	6/79/0
(Feng et al., 2014)	6/79/0	4/81/0	14/71/0	61/24/0	72/13/0	56/29/0	6/79/0
(Yin et al., 2018)	9/76/0	18/67/0	49/36/0	73/12/0	77/8/0	67/18/0	6/79/0
(Pregibon, 1982)	49/36/0	48/37/0	54/31/0	60/25/0	60/25/0	56/29/0	50/35/0
Batch norm	1/84/0	32/53/0	53/32/0	60/25/0	61/24/0	55/30/0	14/71/0

(d) Summary of the results for $p = 0.5$ and blurred images, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	85/0/0	83/2/0	83/2/0	83/2/0	82/3/0	75/10/0	67/18/0
All data	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0
Median of probs.	4/81/0	3/82/0	4/81/0	6/79/0	17/68/0	14/71/0	24/61/0
(Feng et al., 2014)	4/81/0	3/82/0	3/82/0	5/80/0	16/69/0	16/69/0	26/59/0
(Yin et al., 2018)	19/66/0	31/54/0	10/75/0	29/56/0	34/51/0	32/53/0	38/47/0
(Pregibon, 1982)	58/27/0	58/27/0	63/22/0	67/18/0	70/15/0	72/13/0	73/12/0
Batch norm	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0

(e) Summary of the results for $p = 0.5$ and dead pixels, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	85/0/0	83/2/0	84/1/0	84/1/0	80/5/0	70/15/0	62/23/0
All data	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0
Median of probs.	4/81/0	3/82/0	4/81/0	4/81/0	10/75/0	9/76/0	18/67/0
(Feng et al., 2014)	4/81/0	3/82/0	3/82/0	4/81/0	9/76/0	10/75/0	16/69/0
(Yin et al., 2018)	20/65/0	32/53/0	11/74/0	21/64/0	29/56/0	20/65/0	21/64/0
(Pregibon, 1982)	25/60/0	24/61/0	37/48/0	37/48/0	38/47/0	40/45/0	37/48/0
Batch norm	0/85/0	0/85/0	0/85/0	2/83/0	4/81/0	4/81/0	1/84/0

(f) Summary of the results for $p = 0.5$ and RGB channels swapped, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	85/0/0	84/1/0	84/1/0	84/1/0	78/7/0	47/38/0	10/75/0
All data	0/85/0	4/81/0	25/60/0	45/40/0	53/32/0	48/37/0	35/50/0
Median of probs.	4/81/0	2/83/0	3/82/0	30/55/0	52/33/0	48/37/0	35/50/0
(Feng et al., 2014)	5/80/0	2/83/0	2/83/0	22/63/0	45/40/0	43/42/0	37/48/0
(Yin et al., 2018)	23/62/0	45/40/0	20/65/0	31/54/0	52/33/0	52/33/0	37/48/0
(Pregibon, 1982)	51/34/0	59/26/0	58/27/0	62/23/0	60/25/0	60/25/0	59/26/0
Batch norm	0/85/0	0/85/0	16/69/0	35/50/0	38/47/0	18/67/0	4/81/0

Table 7: Summary of the results for $p = 0.2$, over all 85 prediction tasks and all corruptions.

Baseline \ n	$n = 0$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	507/3/0	505/5/0	504/6/0	492/18/0	459/51/0	429/81/0	404/106/0
All data	0/85/0	0/510/0	0/510/0	0/510/0	0/510/0	1/509/0	2/508/0	1/509/0
Median of probs.	9/76/0	28/482/0	21/489/0	16/494/0	17/493/0	28/482/0	30/478/2	31/479/0
(Feng et al., 2014)	8/77/0	30/480/0	24/486/0	16/494/0	16/494/0	20/489/1	23/485/2	26/484/0
(Yin et al., 2018)	14/71/0	95/415/0	146/364/0	34/476/0	42/468/0	39/471/0	36/474/0	40/470/0
(Pregibon, 1982)	55/30/0	282/228/0	282/228/0	275/235/0	287/223/0	264/246/0	281/229/0	267/243/0
Batch norm	0/85/0	0/510/0	0/510/0	0/510/0	0/510/0	0/509/1	0/509/1	1/508/1

Table 8: Summary of results for $p = 0.2$, split by the type of data corruption

(a) Summary of the results for $p = 0.2$ and label bias, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	85/0/0	84/1/0	84/1/0	76/9/0	60/25/0	46/39/0	28/57/0
All data	0/85/0	0/85/0	0/85/0	0/85/0	1/84/0	2/83/0	1/84/0
Median of probs.	5/80/0	1/84/0	0/85/0	0/85/0	11/74/0	13/70/2	13/72/0
(Feng et al., 2014)	5/80/0	4/81/0	0/85/0	0/85/0	3/81/1	5/78/2	8/77/0
(Yin et al., 2018)	25/60/0	49/36/0	7/78/0	19/66/0	12/73/0	4/81/0	2/83/0
(Pregibon, 1982)	47/38/0	40/45/0	48/37/0	44/41/0	40/45/0	44/41/0	41/44/0
Batch norm	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0

(b) Summary of the results for $p = 0.2$ and shuffled labels, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	84/1/0	84/1/0	84/1/0	77/8/0	74/11/0	73/12/0
All data	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0
Median of probs.	4/81/0	4/81/0	1/84/0	1/84/0	1/84/0	0/85/0	0/85/0
(Feng et al., 2014)	5/80/0	4/81/0	1/84/0	1/84/0	1/84/0	0/85/0	0/85/0
(Yin et al., 2018)	17/68/0	31/54/0	4/81/0	1/84/0	1/84/0	0/85/0	2/83/0
(Pregibon, 1982)	47/38/0	54/31/0	49/36/0	53/32/0	52/33/0	54/31/0	51/34/0
Batch norm	0/85/0	0/85/0	0/85/0	0/85/0	0/84/1	0/84/1	0/84/1

(c) Summary of the results for $p = 0.2$ and shuffled features, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	85/0/0	83/2/0	82/3/0	73/12/0	66/19/0	61/24/0
All data	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0
Median of probs.	5/80/0	4/81/0	4/81/0	4/81/0	3/82/0	3/82/0	2/83/0
(Feng et al., 2014)	5/80/0	4/81/0	4/81/0	4/81/0	3/82/0	3/82/0	2/83/0
(Yin et al., 2018)	8/77/0	5/80/0	4/81/0	4/81/0	3/82/0	3/82/0	2/83/0
(Pregibon, 1982)	50/35/0	44/41/0	44/41/0	52/33/0	49/36/0	51/34/0	39/46/0
Batch norm	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0

(d) Summary of the results for $p = 0.2$ and blurred images, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	85/0/0	84/1/0	84/1/0	83/2/0	83/2/0	84/1/0	84/1/0
All data	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0
Median of probs.	4/81/0	3/82/0	4/81/0	6/79/0	7/78/0	10/75/0	10/75/0
(Feng et al., 2014)	4/81/0	3/82/0	4/81/0	6/79/0	7/78/0	10/75/0	11/74/0
(Yin et al., 2018)	15/70/0	15/70/0	6/79/0	9/76/0	14/71/0	18/67/0	22/63/0
(Pregibon, 1982)	57/28/0	56/29/0	56/29/0	55/30/0	55/30/0	55/30/0	58/27/0
Batch norm	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0

(e) Summary of the results for $p = 0.2$ and dead pixels, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	85/0/0	84/1/0	84/1/0	83/2/0	84/1/0	82/3/0	83/2/0
All data	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0
Median of probs.	6/79/0	5/80/0	4/81/0	4/81/0	4/81/0	3/82/0	5/80/0
(Feng et al., 2014)	6/79/0	5/80/0	4/81/0	3/82/0	4/81/0	4/81/0	4/81/0
(Yin et al., 2018)	12/73/0	14/71/0	7/78/0	6/79/0	6/79/0	9/76/0	10/75/0
(Pregibon, 1982)	30/55/0	36/49/0	26/59/0	27/58/0	23/62/0	25/60/0	26/59/0
Batch norm	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	1/84/0

(f) Summary of the results for $p = 0.2$ and RGB channels swapped, over all 85 prediction tasks

Baseline \ n	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 55$	$n = 59$
Reference only	84/1/0	84/1/0	85/0/0	84/1/0	82/3/0	77/8/0	75/10/0
All data	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0
Median of probs.	4/81/0	4/81/0	3/82/0	2/83/0	2/83/0	1/84/0	1/84/0
(Feng et al., 2014)	5/80/0	4/81/0	3/82/0	2/83/0	2/83/0	1/84/0	1/84/0
(Yin et al., 2018)	18/67/0	32/53/0	6/79/0	3/82/0	3/82/0	2/83/0	2/83/0
(Pregibon, 1982)	51/34/0	52/33/0	52/33/0	56/29/0	45/40/0	52/33/0	52/33/0
Batch norm	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0	0/85/0