Appendix

A1 Assumptions

A1: Θ is finite; $l(\cdot, \cdot)$ is zero-one loss for binary classification.

The assumption leads to classical discussions on the *i.i.d* setting in multiple textbooks (*e.g.*, [16]). However, modern machine learning concerns more than the *i.i.d* setting, therefore, we need to quantify the variations between train and test distributions. Analysis of domain adaptation is discussed [2], but still relies on the explicit knowledge of the target distribution to quantify the bound with an alignment of the distributions. The following discussion is devoted to the scenario when we do not have the target distribution to align.

Since we are interested in the $\theta^\star$ instead of the $\theta^\star(\mathcal{D})$, we first assume Θ is large enough and we can find a global optimum hypothesis that is applicable to any distribution, or in formal words:

A2: $L(\theta^\star; \mathcal{D}) = L(\theta^\star(\mathcal{D}); \mathcal{D})$ for any $\mathcal{D}$.

This assumption can be met when the conditional distribution $\mathbb{P}(\mathcal{Y}(\mathcal{D})|\mathcal{Z}(\mathcal{D}))$ is the same for any $\mathcal{D}$.

e.g., The true concept of "cat" is the same for any collection of images.

The challenge of cross-domain evaluation comes in when there exists multiple optimal hypothesis that are equivalently good for one distribution, but not every optimal hypothesis can be applied to other distributions.

e.g., For the distribution of picture book, "cats have chubby faces" can predict the true concept of "cat". A model only needs to learn one of these signals to reduce training error, although the other signal also exists in the data.

The follow-up discussion aims to show that RSC can force the model to learn multiple signals, so that it helps in cross-domain generalization.

Further, Assumption **A2** can be interpreted as there is at least some features $\mathbf{z}$ that appear in every distributions we consider. We use i to index this set of features. Assumption **A2** also suggests that $\mathbf{z}_i$ is *i.i.d.* (otherwise there will not exist $\theta^\star$) across all the distributions of interest (but $\mathbf{z}$ is not *i.i.d.* because $\mathbf{z}_{-i}$, where $-i$ denotes the indices other than i, can be sampled from arbitrary distributions).

e.g., $\mathbf{z}$ is the image; $\mathbf{z}_i$ is the ingredients of the true concept of a "cat", such as ears, paws, and furs; $\mathbf{z}_{-i}$ is other features such as "sitting by the window".

We use $\mathcal{O}$ to specify the distribution that has values on the i^{th}, but 0s elsewhere. We introduce the next assumption:

A3: Samples of any distribution of interest (denoted as $\mathcal{A}$) are perturbed version of samples from $\mathcal{O}$ by sampling arbitrary features for $\mathbf{z}_{-i}$: $\mathbb{E}_{\mathcal{A}}[\mathbb{E}_{\mathcal{S}}[\mathbf{z}]] = \mathbb{E}_{\mathcal{O}}[\mathbf{z}]$

Notice that this does not contradict with our cross-domain set-up: while Assumption **A3** implies that data from any distribution of interest is *i.i.d* (otherwise the operation $\mathbb{E}_{\mathcal{A}}[]$ is not valid), the cross-domain difficulty is raised when only different subsets of $\mathcal{A}$ are used for train and test. For example, considering $\mathcal{A}$ to be a uniform distribution of $[0, 1]$, while the train set is uniformly sampled from $[0, 0.5]$ and the test set is uniformly sampled from $(0.5, 1]$.

A2 Proof of Theoretical Results

A2.1 Corollary 1

Proof. We first study the convergence part, where we consider a fixed hypothesis. We first expand

$$\begin{aligned}
&|L(\widehat{\theta}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D})| \\
&= |L(\widehat{\theta}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S}) - L(\widehat{\theta}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D}) + L(\widehat{\theta}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D})| \\
&\leq |L(\widehat{\theta}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D})| + |L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D})|
\end{aligned}$$

We first consider the term $|L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D})|$, where we can expand

$$|L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D})| \leq 2|L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{O})|$$

because of Assumption **A4**.

Also, because of Assumption **A4**, if samples in $\mathcal{S}$ are perturbed versions of samples in $\mathcal{O}$, then samples in $\mathcal{O}$ can also be seen as perturbed versions of samples in $\mathcal{S}$, thus, Condition 6 can be directly re-written into:

$$|L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{O})| \leq \xi(p),$$

which directly leads us to the fact that $|L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{D})|$ has the expectation 0 (**A4**) and bounded by $[0, \xi(p)]$.

For $|L(\widehat{\theta}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S}) - L(\theta^{\star}_{\mathrm{RSC}}(\mathcal{S}); \mathcal{S})|$, the strategy is relatively standard. We first consider the convergence of a fixed hypothesis θ_{RSC}, then over n *i.i.d* samples, the empirical risk ($\widehat{L}(\theta_{\mathrm{RSC}})$) will be bounded within $[0, 1]$ with the expectation $L(\theta_{\mathrm{RSC}})$.

Before we consider the uniform convergence step, we first put the two terms together and apply the Hoeffding's inequality. When the random variable is with expectation $L(\theta_{\mathrm{RSC}})$ and bound $[0, 1 + 2\xi(p)]$, we have:

$$\mathbb{P}(|\widehat{L}(\theta_{\mathrm{RSC}}; \mathcal{S}) - L(\theta_{\mathrm{RSC}}; \mathcal{D})| \geq \epsilon) \leq 2\exp(-\frac{2n\epsilon^2}{(2\xi(p)+1)^2})$$

Now, we consider the uniform convergence case, where we have:

$$\mathbb{P}(\sup_{\theta_{\mathrm{RSC}} \in \Theta_{\mathrm{RSC}}} |\widehat{L}(\theta_{\mathrm{RSC}}; \mathcal{S}) - L(\theta_{\mathrm{RSC}}; \mathcal{D})| \geq \epsilon) \leq 2|\Theta_{\mathrm{RSC}}| \exp(-\frac{2n\epsilon^2}{(2\xi(p)+1)^2})$$

Rearranging these terms following standard tricks will lead to the conclusion.

A2.2 Corollary 2

Proof. Since we only concern with iteration t, we drop the subscript of $\mathbf{z}_t$ and $\tilde{\mathbf{z}}_t$. We first introduce another shorthand notation

$$h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \mathbf{z}) := \sum_{\langle \mathbf{z}_t, \mathbf{y} \rangle} l(f(\mathbf{z}; \widehat{\theta}_{\mathrm{RSC}}); \mathbf{y})$$

We expand

$$\begin{aligned}
\Gamma(\widehat{\theta}_{\mathrm{RSC}}(t+1)) =& |h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \mathbf{z}) - h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \tilde{\mathbf{z}})| \\
=& |h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \mathbf{z}) - h(\widehat{\theta}_{\mathrm{RSC}}(t), \tilde{\mathbf{z}}) + h(\widehat{\theta}_{\mathrm{RSC}}(t), \tilde{\mathbf{z}}) - h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \tilde{\mathbf{z}})| \\
=& |h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \mathbf{z}) - h(\widehat{\theta}_{\mathrm{RSC}}(t), \mathbf{z}) + h(\widehat{\theta}_{\mathrm{RSC}}(t), \mathbf{z}) - h(\widehat{\theta}_{\mathrm{RSC}}(t), \tilde{\mathbf{z}}) \\
& + h(\widehat{\theta}_{\mathrm{RSC}}(t), \tilde{\mathbf{z}}) - h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \tilde{\mathbf{z}})| \\
=& |h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \mathbf{z}) - h(\widehat{\theta}_{\mathrm{RSC}}(t), \mathbf{z}) + h(\widehat{\theta}_{\mathrm{RSC}}(t), \tilde{\mathbf{z}}) - h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \tilde{\mathbf{z}}) + \Gamma(\widehat{\theta}_{\mathrm{RSC}}(t))|
\end{aligned}$$

Recall that, by the definition of RSC, we have:

$$\widehat{\theta}_{\mathrm{RSC}}(t+1) = \widehat{\theta}_{\mathrm{RSC}}(t) - \frac{\partial h(\widehat{\theta}_{\mathrm{RSC}}(t), \tilde{\mathbf{z}})}{\partial \widehat{\theta}_{\mathrm{RSC}}(t)} \eta = \widehat{\theta}_{\mathrm{RSC}}(t) - \tilde{\mathbf{g}}\eta$$

We apply Taylor expansion over $h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \cdot)$ with respect to $\widehat{\theta}_{\mathrm{RSC}}(t)$ and have:

$$\begin{aligned}
h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \cdot) =& h(\widehat{\theta}_{\mathrm{RSC}}(t), \cdot) + \frac{\partial h(\widehat{\theta}_{\mathrm{RSC}}(t), \cdot)}{\partial \widehat{\theta}_{\mathrm{RSC}}(t)} (\widehat{\theta}_{\mathrm{RSC}}(t+1) - \widehat{\theta}_{\mathrm{RSC}}(t)) \\
& + \frac{1}{2} \frac{\partial^2 h(\widehat{\theta}_{\mathrm{RSC}}(t), \cdot)}{\partial^2 \widehat{\theta}_{\mathrm{RSC}}(t)} ||\widehat{\theta}_{\mathrm{RSC}}(t+1) - \widehat{\theta}_{\mathrm{RSC}}(t)||_2^2 + \sigma \\
=& h(\widehat{\theta}_{\mathrm{RSC}}(t), \cdot) - \frac{\partial h(\widehat{\theta}_{\mathrm{RSC}}(t), \cdot)}{\partial \widehat{\theta}_{\mathrm{RSC}}(t)} \tilde{\mathbf{g}}\eta + \frac{1}{2} \frac{\partial^2 h(\widehat{\theta}_{\mathrm{RSC}}(t), \cdot)}{\partial^2 \widehat{\theta}_{\mathrm{RSC}}(t)} ||\tilde{\mathbf{g}}\eta||_2^2 + \sigma,
\end{aligned}$$

where σ denotes the higher order terms.

Assumption **A6** conveniently allows us to drop terms regarding η^2 or higher orders, so we have:

$$h(\widehat{\theta}_{\mathrm{RSC}}(t), \cdot) - h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \cdot) = \frac{\partial h(\widehat{\theta}_{\mathrm{RSC}}(t), \cdot)}{\partial \widehat{\theta}_{\mathrm{RSC}}(t)} \tilde{\mathbf{g}}\eta \tag{8}$$

Finally, when $\cdot$ is replaced by $\mathbf{z}$ and $\tilde{\mathbf{z}}$,

we have:

$$h(\widehat{\theta}_{\mathrm{RSC}}(t), \tilde{\mathbf{z}}) - h(\widehat{\theta}_{\mathrm{RSC}}(t+1), \tilde{\mathbf{z}}) = \frac{\partial h(\widehat{\theta}_{\mathrm{RSC}}(t), \tilde{\mathbf{z}})}{\partial \widehat{\theta}_{\mathrm{RSC}}(t)} \tilde{\mathbf{g}}\eta = ||\tilde{\mathbf{g}}||_2^2 \eta$$

and

$$h(\widehat{\theta}_{\mathrm{RSC}}(t),\mathbf{z}) - h(\widehat{\theta}_{\mathrm{RSC}}(t+1),\mathbf{z}) = \frac{1}{\gamma_t(p)} \frac{\partial h(\widehat{\theta}_{\mathrm{RSC}}(t),\tilde{\mathbf{z}})}{\partial \widehat{\theta}_{\mathrm{RSC}}(t)} \tilde{\mathbf{g}}\eta = \frac{1}{\gamma_t(p)} ||\tilde{\mathbf{g}}||_2^2 \eta$$

We write these terms back and get

$$\Gamma(\widehat{\theta}_{\mathrm{RSC}}(t+1)) = \left| (\frac{1}{\gamma_t(p)} - 1)||\tilde{\mathbf{g}}||_2^2 \eta + \Gamma(\widehat{\theta}_{\mathrm{RSC}}(t)) \right|$$

We can simply drop the absolute value sign because all these terms are greater than zero. Finally, we rearrange these terms and prove the conclusion.

Additional Experiment

We tested strategies including applying RSC to top layers (conv5), to internal layers (conv4), and to top + internal layers (conv4 + conv5), and found that the top layer works the best. We conjecture this behavior is because operating on high-level feature maps can help the classifier train more effectively.

Layers	backbone	artpaint	cartoon	sketch	photo	Avg↑
conv4	ResNet18	79.91	74.56	74.73	**96.13**	81.33
conv4+conv5	ResNet18	81.61	76.93	78.48	95.72	83.19
conv5	ResNet18	**83.43**	**80.31**	**80.85**	95.99	**85.15**

Table A1. Ablation study of applying RSC to internal layers. RSC used the hyperparameters selected in above ablation studies:"Top-Gradient", Feature Dropping Percentage (33.3%) and Batch Percentage (33.3%).