A. Proofs

This appendix includes the proofs of Propositions 1 and 2.

Proof of Proposition 1. Let us fix a label value $y \in \mathcal{Y}$ and two distinct domains $p_1(x, y)$ and $p_2(x, y)$ from the union of test and training domains. Let $X_1 \sim p_1(x, y), X_2 \sim p_2(x, y), Z_1 = h(X_1)$, and $Z_2 = h(X_2)$. Let w be the parameters of the common classifier for all training and test domains (i.e., the classifier that makes $e'_2 = 0$). Then for all z,

$$p_1(z|y) = \frac{p_1(z)p_1(y|z)}{\int p_1(z')p_1(y|z')dz'} \tag{3}$$

$$= \frac{p_2(z)p_1(y|z)}{\int p_2(z')p_1(y|z')dz'} \qquad \text{(by domain inv. of } p(z)\text{)} \tag{4}$$

$$= \frac{p_2(z)\delta_y\left(\text{argmax}_k f_w(z)_k\right)}{\int p_2(z')\delta_y\left(\text{argmax}_k f_w(z')_k\right)dz'} \tag{5}$$

$$= \frac{p_2(z)p_2(y|z)}{\int p_2(z')p_2(y|z')dz'} = p_2(z|y). \tag{6}$$

Above $\delta_y(y')$ is the Krokecker's delta function and the last two transitions use the fact the classifier with parameters w correctly classifies representations of both domains. □

Proof of Proposition 2. Let $X_1 \sim p_1^1(x, y), Z_1 = h(X_1)$ be a random sample from the first training domain and $X_i \sim p_i^3(x, y), Z_i = h(X_i)$ be a random sample from the i-th test domain. Let w be the parameters of the classifier found by the training algorithm that classifies correctly representations of all training domains. If we fix a label $y \in \mathcal{Y}$, then

$$\begin{aligned}&\mathbb{P}(\text{argmax}_k f_w(Z_i)_k = y \mid Y_i = y) = \\ &= \mathbb{P}(\text{argmax}_k f_w(Z_1)_k = y \mid Y_1 = y),\end{aligned} \tag{7}$$

as $p_i^3(z|y) = p_1^1(z|y)$ by the assumptions. As $e'_0 = 0$ implies that the classifier f_w perfectly classifies representations from the first training domain, the right-hand-side of (7) will be equal to 1. Therefore, the f_w will correctly classify also the representations of test domain $p_i^3(x, y)$. As i was arbitrary, this concludes the proof. □

B. Decompositions

This appendix describes additional properties of the decompositions presented in the main text (see Sec. 4.4).

Let $\mathscr{D} = \{p_i(x, y)\}_{i=1}^{n_1+n_3}$ be a family of $n_1 + n_3$ domains. Let S be a subset of $\{1, 2, \ldots, n_1 + n_3\}$ of size n_1, chosen uniformly at random, and let $\bar{S} \triangleq (\{1, 2, \ldots, n_1 + n_3\} \setminus S)$ be the complement of S. This subset S defines a partition of $\mathscr{D}$ into training and test domains. Let $w(S)$ and $\theta(S)$ denote the parameters of the classification head and the feature extractor after training on domains specified by S. Let $(X_1, Y_1), \ldots, (X_{n_1+n_3}, Y_{n_1+n_3})$ be random variables drawn from distributions $p_1(x, y), \ldots, p_{n_1+n_3}(x, y)$, respectively. To simplify the derivations below, we define

$$F(A; w, \theta) = \frac{1}{|A|}\sum_{i \in A} \mathbb{E}_{X_i, Y_i}\left[\ell\left(f_w(h_\theta(X_i)), Y_i\right)\right], \tag{8}$$

for any subset A of $\{1, 2, \ldots, n_1 + n_3\}$. To avoid unnecessary technical complications in statements and proofs below, we assume that the feature extractor parameters θ, label classifier parameters w, and domain classifier parameters θ belong to compact domains. This will allow us to replace infimum operators in the definitions of generalization and invariance metrics by minimum operators.

With these conventions, the generalization metrics can be written the following way

$$e'_0(S) = F(S; w(S), \theta(S)), \tag{9}$$

$$e'_1(S) = \min_{w' \in \mathcal{W}} F(\bar{S}; w', \theta(S)) \tag{10}$$

$$e'_2(S) = F(\bar{S}; \tilde{w}(S), \theta(S)), \text{ where } \tilde{w}(S) \in \arg\min_{w' \in \mathcal{W}} F(S \cup \bar{S}; w', \theta(S)), \tag{11}$$

$$e'_3(S) = F(\bar{S}; w(S), \theta(S)). \tag{12}$$

Next, we prove several statements to establish the relationship between the metrics.

Proposition 3. *Assume that* $n_1 = n_3$ *and* $(w(S), \theta(S)) \in \arg\min_{w',\theta'} F(w', S, \theta')$. *Then* $\mathbb{E}_S[e'_0(S)] \leq \mathbb{E}_S[e'_1(S)]$.

Proof.

$$\begin{aligned}
\mathbb{E}_S[e'_0(S)] &= \mathbb{E}_S F(S; w(S), \theta(S)) \\
&\leq \mathbb{E}_S \min_{w' \in \mathcal{W}} F(S; w', \theta(\bar{S})) && \text{(as } (w(S), \theta(S)) \text{ is a global minimum of } F(S; w', \theta')) \\
&= \mathbb{E}_{\bar{S}} \min_{w' \in \mathcal{W}} F(\bar{S}; w', \theta(S)) && \text{(replacing } S \text{ by } \bar{S}, \text{ as } n_1 = n_3 \Rightarrow S \overset{d}{=} \bar{S}) \\
&= \mathbb{E}_S[e'_1(S)]. && \text{(by definition)}
\end{aligned}$$

□

Note that the assumptions of the proposition are critical. If $n_3 < n_1$, then the classification on the test domains might work well simply because there are fewer domains in the test set. In an extreme case, when $n_3 = 1$, the representations of that single domain might be easily separable, while finding a universal good classifier for all training domains can be hard. As the training is performed on the training domains, it is reasonable to assume the learned parameters $(w(S), \theta(S))$ are optimal with respect to the training domains. If this assumption is violated, one cannot exclude that the model can end up with a better classifier for the test domains ($\bar{S}$) while being trained on the training domains (S). Even with these assumptions, the partition $\mathscr{D}$ of into training and test domains can be "adversarial" in a sense that the test domains are much easier, which will cause smaller e_1. However, such scenario cannot happen for every partition of $\mathscr{D}$. Hence, the Proposition 3 proves the desired relation between $e_0(S)$ and $e_1(S)$ only in expectation over S.

The next two inequalities can be proved with weaker assumptions.

Proposition 4. *Let the generalization metrics* $e'_1(S)$, $e'_2(S)$ *and* $e'_3(S)$ *be defined as above. Then, (i)* $e'_1(S) \leq e'_2(S)$. *Furthermore, if* $w(S) \in \arg\min_{w'} F(S; w', \theta(S))$ *then (ii)* $e'_2(S) \leq e'_3(S)$.

Proof. **(i)** We have that

$$e'_1(S) = \min_{w' \in \mathcal{W}} F(\bar{S}; w', \theta(S)) \leq F(\bar{S}; \tilde{w}(S), \theta(S)) = e'_2(S). \tag{13}$$

(ii) Let $\alpha \triangleq n_1/(n_1 + n_3)$. As $\tilde{w}(S)$ is a global minimum of $F(S \cup \bar{S}; w', \theta(S))$, we have that

$$F(S \cup \bar{S}; \tilde{w}(S), \theta(S)) \leq F(S \cup \bar{S}; w(S), \theta(S)), \tag{14}$$

which is equivalent to

$$\alpha F(S; \tilde{w}(S), \theta(S)) + (1-\alpha) F(\bar{S}; \tilde{w}(S), \theta(S)) \leq \alpha F(S; w(S), \theta(S)) + (1-\alpha) F(\bar{S}; w(S), \theta(S)). \tag{15}$$

This simplifies to

$$F(\bar{S}; \tilde{w}(S), \theta(S)) \leq F(\bar{S}; w(S), \theta(S)) + \frac{\alpha}{1-\alpha} \left(F(S; w(S), \theta(S)) - F(S; \tilde{w}(S), \theta(S)) \right). \tag{16}$$

The additional assumption that $w(S) \in \arg\min_{w'} F(S; w', \theta(S))$ implies that $F(S; w(S), \theta(S)) \leq F(S; \tilde{w}(S), \theta(S))$. Connecting this with (16) we get

$$F(\bar{S}; \tilde{w}(S), \theta(S)) \leq F(\bar{S}; w(S), \theta(S)), \tag{17}$$

which is the same as $e'_2(S) \leq e'_3(S)$. □

Next we investigate the relationship between invariance metrics. Similar to the function $F(A; w, \theta)$ defined above, we define

$$G(A; \omega, \theta) = \frac{1}{|A|} \sum_{i \in A} \mathbb{E}_{X_i} \left[\ell \left(g_\omega(h_\theta(X_i)), i \right) \right], \tag{18}$$

for any subset A of $\{1, 2, \ldots, n_1 + n_3\}$. With these conventions, the invariance metrics can be written as follows:

$$d'_0(S) = 1 - \min_{\omega \in \Omega} G(S; \omega, \theta(S)) - \frac{1}{n_1}, \tag{19}$$

$$d'_1(S) = 1 - \min_{\omega \in \Omega} G(S \cup \bar{S}; \omega, \theta(S)) - \frac{1}{n_1 + n_3}, \tag{20}$$

$$d'_2(S) = 1 - \frac{1}{C} \sum_{y=1}^{C} \min_{\omega \in \Omega} \mathbb{E}\left[\frac{1}{n_1 + n_3} \sum_{i \in S \cup \bar{S}} \ell(g_\omega(h_\theta(X_i)), i) \,\middle|\, E_y \right] - \frac{1}{n_1 + n_3}, \tag{21}$$

where E_y denotes the event $(Y_1 = y \wedge \cdots \wedge Y_{n_1+n_3} = y)$.

The relationship between $d'_0(S)$ and $d'_1(S)$ is significantly more complicated as the sets of domains are different, and the accuracy scores are not directly comparable. Furthermore, the training algorithm can be "adversarial" in the sense that its produced representations of training domains are more distinguishable compared to that of testing domains. The relationship between $d'_1(S)$ and $d'_2(S)$ is simpler, and the desired inequality can be proved with mild assumptions.

Proposition 5. *Let $d_1(S)$ and $d_2(S)$ be defined as above. Assuming that $P(Y_i = y) = 1/C$ for all $i \in \{1, 2, \ldots, n_1 + n_3\}$ and $y \in \{1, 2, \ldots, C\}$, we have that $d'_1(S) \leq d'_2(S)$.*

Proof.

$$\begin{aligned}
d'_2(S) &= 1 - \frac{1}{C} \sum_{y=1}^{C} \min_{\omega \in \Omega} \mathbb{E}\left[\frac{1}{n_1 + n_3} \sum_{i \in S \cup \bar{S}} \ell(g_\omega(h_\theta(X_i)), i) \,\middle|\, E_y \right] - \frac{1}{n_1 + n_3} \\
&= 1 - \frac{1}{C} \sum_{y=1}^{C} \min_{\omega \in \Omega} \left(\frac{1}{n_1 + n_3} \sum_{i \in S \cup \bar{S}} \mathbb{E}_{X_i}\left[\ell(g_\omega(h_\theta(X_i)), i) \,\middle|\, Y_i = y \right] \right) - \frac{1}{n_1 + n_3} \\
&\geq 1 - \min_{\omega \in \Omega} \left(\frac{1}{C} \sum_{y=1}^{C} \frac{1}{n_1 + n_3} \sum_{i \in S \cup \bar{S}} \mathbb{E}_{X_i}\left[\ell(g_\omega(h_\theta(X_i)), i) \,\middle|\, Y_i = y \right] \right) - \frac{1}{n_1 + n_3} \\
&= 1 - \min_{\omega \in \Omega} \left(\frac{1}{n_1 + n_3} \sum_{i \in S \cup \bar{S}} \mathbb{E}_{X_i}\left[\ell(g_\omega(h_\theta(X_i)), i) \right] \right) - \frac{1}{n_1 + n_3} \\
&= 1 - \min_{\omega \in \Omega} G(S \cup \bar{S}; \omega, \theta(S)) - \frac{1}{n_1 + n_3} \\
&= d'_1(S).
\end{aligned}$$

□

C. Dataset Samples

Fig. 4 shows one sample per class per domain for Colored MNIST and Camelyon17 datasets used in our experiments. Note that Camelyon17 images are originally four-channel RGBA images, while Colored MNIST images have 50 "channels", the first three of which are interpreted as RGB in the figure.

D. Hyperparameters of the algorithms

For all algorithms we fixed the regularization strength hyperparameter space to $\{0.1, 0.5, 1, 2, 5, 10, 15\}$. We used SGD optimizer in all cases. For Colored MNIST dataset we used learning rate of 0.01 and for Camelyon17 we used 0.001. All methods used 0.01 weight decay unless explicitly noted (*e.g.* in Tab. 1). We trained all our algorithms for fixed 10 epochs. One epoch of training on Camelyon17 took about 30 minutes (some algorithms, *e.g.* DANN, IRM, take a bit longer because of more complicated computations) on our machine with two NVIDIA Titan V GPUs, while the algorithms on Colored MNIST took only several seconds.

To produce the plots used in this paper we extracted learned representations by forwarding the data through the learned network and trained multiple logistic regression functions. The duration of this process strongly depends on the size of the

representations and the number of samples in the dataset. Processing the entire dataset for one model on a single Titan V GPU takes 20 minutes for Camelyon17 and less than a minute for Colored MNIST. To produce a single plot we process 10 models (after each epoch) or 8 models (for each value of the hyperparameter β). For some plots, we processed 10 models per epoch epochs to see the behavior in more details in the first three epochs (*e.g.* Fig. 3f).

E. More Results

In this section we provide more details about the trained models. We had trouble training IRM on both datasets. For many combinations of hyperparameters the loss was getting NaN at early stages of the training. One of the successful attempts was on Colored MNIST with $\beta = 10$. The biggest contributor to the error was changing during the training, as shown on Fig. 6a.

Fig. 6 shows generalization errors and domain-distinguishability on Colored MNIST for SD and GroupDRO algorithms. Still, e_2 and e_3 are the main contributors to the generalization error. Fig. 6d shows that more complex networks (i.e. ResNets with more parameters) have smaller errors on the training set and less invariance across training and test domains.

Fig. 7 shows the performance of three more algorithms on Camelyon17. With strong enough regularization, the models underfit the training set. The biggest contributor to the generalization error for other models is the test set inseparability. Fig. 8 shows the impact of the random seed when DeepCORAL is trained on Camelyon17. Some variance is visible in both generalization error and domain distringuishability. In all three cases the algorithm increases domain invariance by the end of the training, but it does not translate into better generalization.

F. Notes on Label Shift

Although label shift is common in real-world applications, it makes error analysis extremely complicated. First, when the number of domains is large, characterizing types of label shift is a challenge on its own. Label shift can appear between training domains, between individual training and test domains, and between the union of training domains and the union of test domains. iWildCam dataset from WILDS benchmark has all of these shifts. Second, the accuracy of the model can change in both directions on the test domains under label shift. In case when the test set contains more examples from "easier" classes, the error on the test set might be low, and it can hide the performance drop due to the domain shift. We also observe that most of the current domain generalization algorithms simply ignore the label shift issue. In fact, the proof of Proposition 1 shows that if $e'_2 = 0$ and the representations are domain invariant with respect to the union of training and test domains, label distribution $p(y)$ is also invariant. It implies that enforcing invariance of representations $p(z)$ when $p(y)$ is not invariant is not desirable, as a common classifier will not exist. Finally, to the best of our knowledge, robustness of algorithms with respect to unknown label shifts is not explored even if $p(x|y)$ is constant across domains. Current label shift literature (*e.g.* [22]) assumes access to the test domain and discusses adaptation strategies.

G. Visualization of Representation Spaces

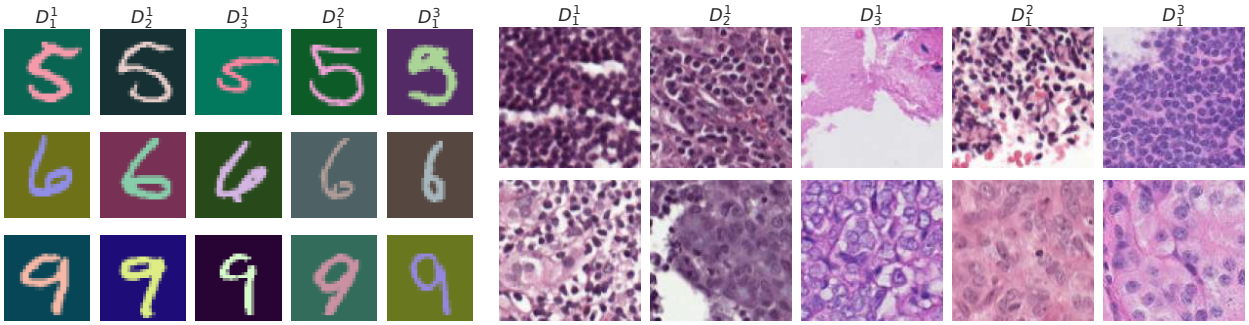
To visualize the learned representations, we forward pass the datasets (all domains) through the networks, take the representations $z = h_\theta(x)$, perform a single two-dimensional Principal Component Analysis for each model (on the combination of all domains), and plot the results. Colors encode the labels, while the marker type encodes the domain. Fig. 9a shows a typical failure on Colored MNIST dataset. For example, the representations of digits 1 are grouped in multiple clusters. Moreover, the representations of the samples from the validation domain V_1^2 (indicated by star markers) are quite far from the other clusters. This shows that a linear classifier that successfully works on the training domains might not be able to classify the digits in the validation domain. Fig. 9b shows the more successful case obtained with ERM+HSIC algorithm where the representations of digits are grouped regardless of the domains.

Fig. 9c shows the space learned by ERM baseline on Camelyon17 dataset. It demonstrates a case when the dots are clearly separable according to the label (color) for the training domains, but there are a few dots from the validation and test domains (V_1^2 and V_1^3) that are in the neighborhood of the wrong color. On the 2D space it can be seen that the representations of the samples from validation and test domains are not fully separable. Fig. 9d demonstrates the case when the regularization is too strong and the representations of most samples have collapsed near $(0, 0)$, leading to large e_0.

H. Correlation Plots

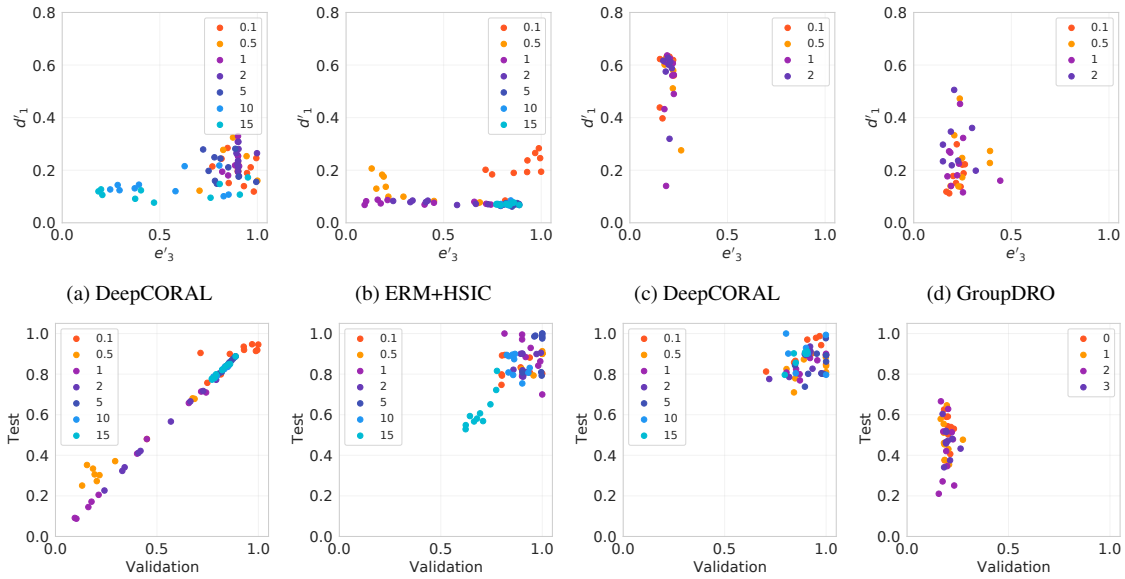
The first row of Fig. 5 shows correlation plots between generalization error e'_3 and training-validation domain distinguishability d'_1 for various algorithms. Every dot corresponds to a single model saved at each epoch. We expect to see positive correlation between these metrics. For Colored MNIST, DeepCORAL shows 0.52 correlation (Fig. 5a), while ERM+HSIC

has 0.05 correlation (Fig. 5b). The latter is explained by large e_0 when regularization is too strong. We do not see any correlation between e'_3 and d'_1 on Camelyon 17. The second row shows the relationship between the errors of many models on validation and test sets. High correlation implies that the accuracy on the validation set can serve as a good metric for model selection. ERM+HSIC, the only algorithm that performed well on Colored MNIST, demonstrates high correlation (Fig. 5e. The correlation for other algorithms is lower. Fig. 5h shows no correlation for DeepCORAL trained on Camelyon17 with various seeds.



(a) Colored MNIST (b) Camelyon17

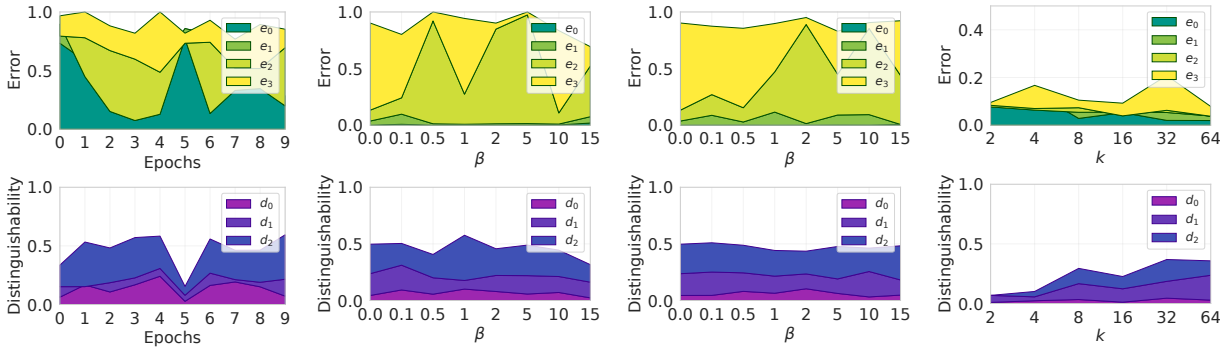
Figure 4. Samples of Colored MNIST and Camelyon17 datasets. Both datasets have three training domains D_1^1, D_2^1, D_3^1, one validation domain D_1^2 and one test domain D_1^3. The first row of (b) shows normal tissue samples, while the samples of the second row contain tumor.



(a) DeepCORAL (b) ERM+HSIC (c) DeepCORAL (d) GroupDRO

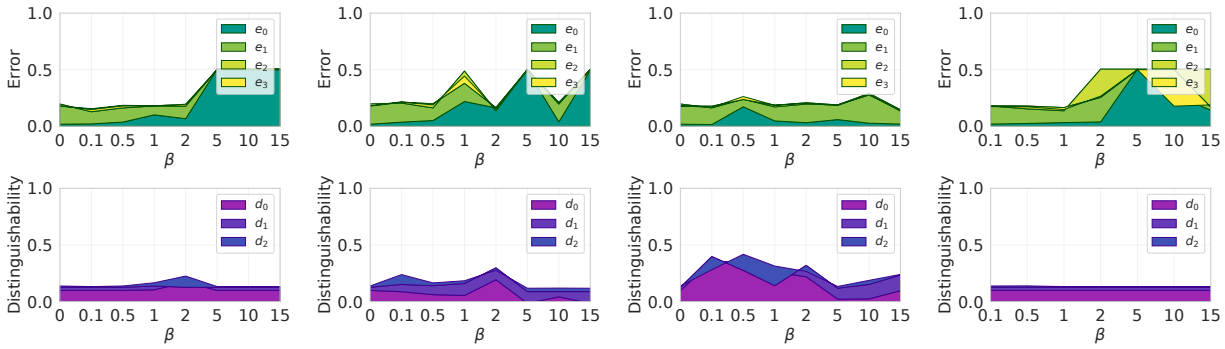
(e) ERM+HSIC (f) SD (g) DANN (h) DeepCORAL

Figure 5. First row: relationship between generalization error e'_3 and training-validation domain distinguishability d'_1 for various algorithms on (a-b) Colored MNIST and (c-d) Camelyon 17 datasets. Colors of the dots indicate the strength of regularization. Second row: correlation plots between validation and test errors. Colors of the dots indicate the strength of regularization for the models trained on Colored MNIST (e-g) or random seed for the models trained on Camelyon17 (h).



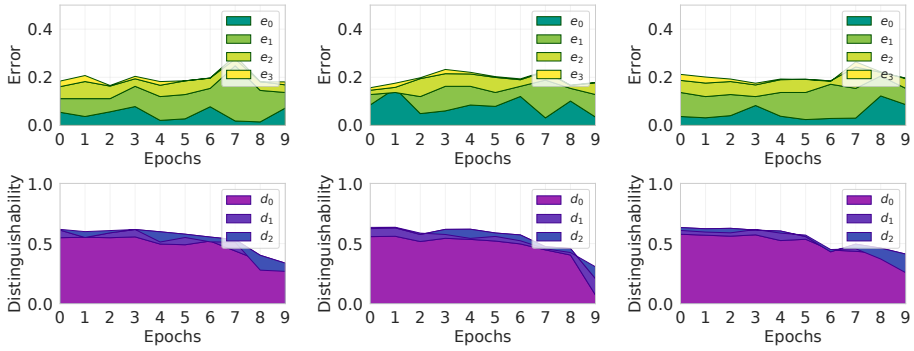
(a) IRM ($\beta = 10$) (b) SD (c) GroupDRO (d) ERM+HSIC ($\beta = 1$)

Figure 6. Decomposition of generalization errors and domain-distinguishability of three more algorithms on Colored MNIST dataset measured on the validation domains. Horizontal axis corresponds to regularization strength of the algorithms.



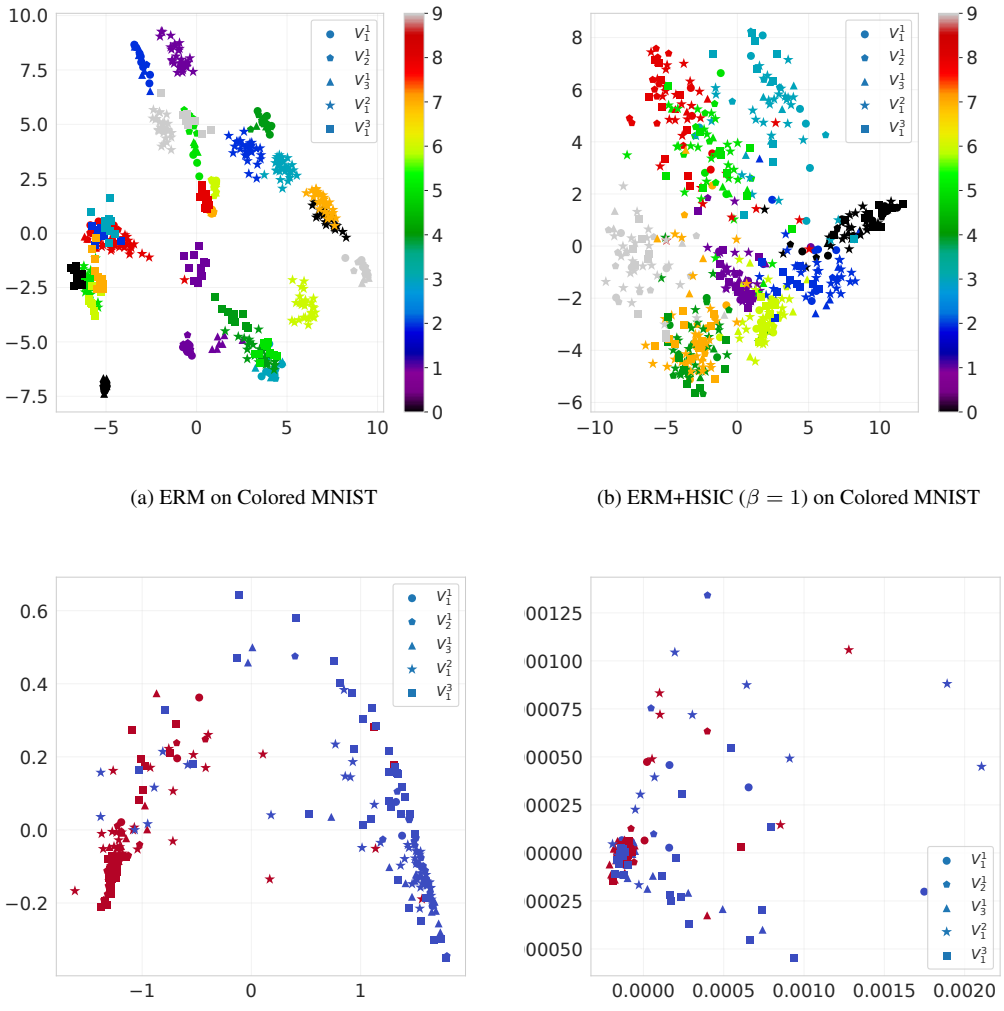
(a) ERM+HSIC (b) GroupDRO (c) DeepCORAL (d) SD

Figure 7. Decomposition of generalization errors and domain-distinguishability of three more algorithms on Camelyon17 dataset measured on the validation domains. Horizontal axis corresponds to regularization strength of the algorithms.



(a) Camelyon17 (seed=1) (b) Camelyon17 (seed=2) (c) Camelyon17 (seed=3)

Figure 8. The decomposition of generalization errors and domain-distinguishability of three DeepCORAL models trained on Camelyon17 ($\beta = 5$) with different random seeds, measured on the validation domain.



(a) ERM on Colored MNIST

(b) ERM+HSIC ($\beta = 1$) on Colored MNIST

(c) ERM on Camelyon17

(d) ERM+HSIC ($\beta = 15$) on Camelyon17

Figure 9. 2D PCA transformations of the learned representations of several models (from the last epoch). Colors indicate the labels, while marker shapes indicate the domains.