

## A PRELIMINARIES AND PROOFS

### A.1 Preliminaries on Important Notations

We first define the following set of symbols:

Data	Row Space	Basis	Coordinate vector	Singular value matrix
$X_{id} \in \mathbb{R}^{p \times d}$	$S \in \mathbb{R}^d, \dim(S) = p$	$F \in \mathbb{R}^{d \times p}$	$z \in \mathbb{R}^p$	$\mathbf{I}_{F^\top F}$
$X_{ood} \in \mathbb{R}^{q \times d}$	$U \in \mathbb{R}^d, \dim(U) = q$	$G \in \mathbb{R}^{d \times q}$	$v \in \mathbb{R}^q$	$\mathbf{I}_{G^\top G}$
$\mathcal{W} \in \mathbb{R}^{m \times d}$	$R \in \mathbb{R}^d, \dim(R) = m$	$E \in \mathbb{R}^{d \times m}$	$r \in \mathbb{R}^m$	$\mathbf{I}_{E^\top E}$
$f(X_{id}) \in \mathbb{R}^{p \times m}$	$A \in \mathbb{R}^m, \dim(A) = p$	$E^\top F \in \mathbb{R}^{m \times p}$	$a \in \mathbb{R}^p$	$\Sigma_{E^\top F}$
$f(X_{ood}) \in \mathbb{R}^{q \times m}$	$B \in \mathbb{R}^m, \dim(B) = q$	$E^\top G \in \mathbb{R}^{m \times q}$	$b \in \mathbb{R}^q$	$\Sigma_{E^\top G}$

wherein  $p$  and  $q$  are the sample numbers of the in-distribution data and out-of-distribution data in the training set, respectively.  $d$  and  $m$  are the dimensions of the data and the model, respectively. Based on the above table, for each sample  $x_{id} \in X_{id}$ , it can be denoted as  $x_{id} = (Fz)^\top$ . Additionally,  $X_{id}$  is represented for the in-distribution training set as  $(Fz)^\top$ , where  $z$  stands for the set of  $z$ . Similarly,  $v, r, a$  and  $b$  are the set of  $v, r, a$  and  $b$ , respectively.

**Definition A.1** (Row space). Give a matrix  $\mathcal{W} \in \mathbb{R}^{m \times d}$ , the rowspace  $R$  of  $\mathcal{W}$  is the span of the row vectors of  $\mathcal{W}$ , which can be denoted as  $C(\mathcal{W}^\top)$ .

**Definition A.2** (Basis and Coordinate). Let  $E \in \mathbb{R}^{d \times \dim(R)}$  have orthonormal columns that span  $R$ . For arbitrary vector  $w \in \mathcal{W}$ , there exist a  $r \in \mathbb{R}^{\dim(R)}$  that satisfies  $w = Er$ , wherein  $E$  is called the orthonormal basis of space  $R$  and  $r$  is the coordinate vector of  $w$  under  $E$ .

#### PROOF OF EQUATION. 4.

$$\begin{aligned}
\sum_{t=1}^{\infty} \frac{\partial \mathcal{L}_t}{\partial \mathcal{W}_t} &= \sum_{t=1}^{\infty} 2 (\mathcal{W}_t - \mathcal{W}^*) \mathbb{E}[X^\top X] \\
&= 2 \sum_{t=1}^{\infty} (Er_t - Er^*)^\top \{p_{id} Fz(Fz)^\top + p_{ood} Gv(Gv)^\top\} \\
&= 2 \sum_{t=1}^{\infty} \{EE^\top (p_{id} Fa_t^\top + p_{ood} Gb_t^\top) - Er^{*\top}\}^\top \times \{p_{id} Fz(Fz)^\top + p_{ood} Gv(Gv)^\top\} \\
&= 2 \sum_{t=1}^{\infty} \{(p_{id} a_t F^\top + p_{ood} b_t G^\top) EE^\top - r^{*\top} E^\top\} \times \{p_{id} Fz(Fz)^\top + p_{ood} Gv(Gv)^\top\} \\
&= 2 \sum_{t=1}^{\infty} (p_{id} a_t F^\top + p_{ood} b_t G^\top) EE^\top \{p_{id} Fz(Fz)^\top + p_{ood} Gv(Gv)^\top\} - r^{*\top} E^\top \mathbb{E}[X^\top X] \\
&= 2 \sum_{t=1}^{\infty} \{p_i^2 a_t F^\top EE^\top Fz(Fz)^\top + p_o^2 b_t G^\top EE^\top Gv(Gv)^\top\} \\
&\quad + p_i p_o a_t F^\top EE^\top Gv(Gv)^\top + p_i p_o b_t G^\top EE^\top Fz(Fz)^\top - r^{*\top} E^\top \mathbb{E}[X^\top X]
\end{aligned} \tag{20}$$

Given that  $E^\top F$  and  $E^\top G$  represent the feature projections of the ID data basis and OOD data basis in the model space, respectively, it follows that  $F^\top EE^\top G \ll F^\top EE^\top F$  and  $F^\top EE^\top G \ll G^\top EE^\top G$ . Consequently, Equation 20 can be further simplified as:

$$\begin{aligned}
\sum_{t=1}^{\infty} \frac{\partial \mathcal{L}_t}{\partial \mathcal{W}_t} &\approx 2 \sum_{t=1}^{\infty} p_i^2 a_t F^\top EE^\top Fz(Fz)^\top + p_o^2 b_t G^\top EE^\top Gv(Gv)^\top - r^{*\top} E^\top \mathbb{E}[X^\top X] \\
&\approx 2 \sum_{t=1}^{\infty} \left\{ p_i^2 \tilde{a}_t \Sigma_{E^\top F, t}^2 X_{id} + p_o^2 \tilde{b}_t \Sigma_{E^\top G, t}^2 X_{ood} \right\}
\end{aligned} \tag{21}$$

where  $\tilde{a}_t = a_t - a^*$  and  $\tilde{b}_t = b_t - b^*$ . Note that in order to make the expression clearer, we omit the representation of some coordinate vectors ( $z$ ) in Eq. 2, so as to highlight the transformation represented by singular matrix  $\Sigma_{E^\top F, t}$  and  $\Sigma_{E^\top G, t}$ .  $\square$

## A.2 Proofs For Biased Performance on OOD and ID Data

**Definition A.3 (Projector).** Give a subspace  $S$  of  $\mathbb{R}^d$ , and  $P$  is the projection matrix which projects a vector  $x \in \mathbb{R}^d$  into the subspace  $S$ . If subspace  $S$  has a orthonormal basis  $E$ , we have:

$$\begin{aligned} P^2 &= P^\top = P \\ P(x) &= E^\top x \end{aligned} \quad (22)$$

**Lemma A.4.** *There exists householder matrix  $H = I - 2uu^H$  satisfying  $\det(H) = -1$ .*

**Lemma A.5.** *If  $A^*$  is the conjugate transpose of  $A$ , then  $A^*$  has the same nonzero singular values with  $A$ .*

PROOF OF LEMMA A.5. Given  $A \in \mathbb{C}^{m \times n}$  and  $A$  has the rank of  $r(A) = \min(m, n)$ ,  $A^* \in \mathbb{C}^{n \times m}$ . Let  $A = C_r^{m \times n}$ . Then we have  $A^*A$  and  $AA^*$  are both non-negative definite Hermite matrices. It can be obtained for Lemma A.4 that, for all  $\lambda \in \mathbb{R}$ , we have:

$$\lambda^m |I_n - AA^*| = \lambda^n |I_m - A^*A| \quad (23)$$

□

PROOF OF EQUATION. 7.

$$\begin{aligned} \mathcal{L}_{ood} &= (W_\infty - W^*)X_{ood}^\top \\ &= \left\{ W_0 - 2lr \lim_{t \rightarrow \infty} \sum_{t=1}^{\infty} \frac{\partial \mathcal{L}_t}{\partial W_t} - W^* \right\} X_{ood}^\top \\ &= \epsilon_{ood} - 2lr \sum_{t=1}^{\infty} \left\{ p_i^2 \Sigma_{E^\top F, t}^2 X_{id} + p_o^2 \Sigma_{E^\top G, t}^2 X_{ood} \right\} X_{ood}^\top \end{aligned} \quad (24)$$

$$\begin{aligned} &\approx \epsilon_{ood} - 2lr \sum_{t=1}^{\infty} p_i^2 \Sigma_{E^\top F, t}^2 \Sigma_{F^\top G} + p_o^2 \Sigma_{E^\top G, t}^2 \mathbf{I}_{G^\top G} \\ \mathcal{L}_{id} &= (W_\infty - W^*)X_{id}^\top \\ &\approx \epsilon_{id} - 2lr \sum_{t=1}^{\infty} p_i^2 \Sigma_{E^\top F, t}^2 \mathbf{I}_{F^\top F} + p_o^2 \Sigma_{E^\top G, t}^2 \Sigma_{G^\top F} \end{aligned} \quad (25)$$

From Lemma. 2, we have  $\Sigma_{G^\top F} = \Sigma_{F^\top G}$ . And since  $\dim(U) = q \ll \dim(S) = p$ , the smallest singular value in singular value matrix  $\min \Sigma_{F^\top G} = \min \Sigma_{G^\top F} = \sigma_{G^\top F}^q$ , wherein  $\sigma_{G^\top F}^q$  represents the  $q$ -th largest value in the singular value matrix. Ignoring terms representing data, it can be derived that:

$$\mathcal{L}_{ood} - \mathcal{L}_{id} \approx (p_i^2 - p_o^2)(1 - \Sigma_{F^\top G}) + \epsilon > 0, \quad (26)$$

□

**Discussion (Performance difference):** The result intuitively shows that the undirectly learned model performs better on feature distributions with larger sample numbers. As shown in Eq. 7, the difference in model performance between OOD and ID data is linearly related to the proportion of the corresponding samples and the correlation degree between the different feature distributions. What's more, when the out-of-domain data has the same proportion as in-domain data in the training dataset ( $p_i = p_o$ ), or the data distributions of OOD are consistent with ID, the task loss difference between OOD and ID data could be reduced to zero.

## A.3 Proofs for ID-targeted Model Sparse

**Lemma A.6.** (3.7) *Spurious features targeted model sparse can effectively reduce the performance deviation of the learned model between in-domain data and out-domain data.*

$$\left| \frac{\mathcal{R}(X_{ood}) - \mathcal{R}(X_{id})}{\mathcal{R}(X_{ood})^{sparse} - \mathcal{R}(X_{id})^{sparse}} \right| \approx \left| \frac{\sum_{j=1}^m p_o \tilde{\sigma}_j \xi_j \gamma_j X_{ood} - \sum_{i=1}^m p_i \sigma_i \xi_i \lambda_i X_{id}}{\sum_{j=1}^m p_o \tilde{\sigma}_j \xi_j \gamma_j X_{ood} - \sum_{i=1}^g p_i \sigma_i \xi_i \lambda_i X_{id}} \right| \geq 1, \quad (27)$$

where  $\sigma_i, \tilde{\sigma}_i$  is the  $i$ -th maximums in  $\Sigma_{E^\top F}$  and  $\Sigma_{E^\top G}$ . And we have  $\sigma > 0$  since the singular values are non-negative.  $m$  and  $g$  are the rank of the singular value matrix after performing compact singular decomposition and truncated singular value decomposition on the projections, respectively.

PROOF OF LEMMA 3.7. As mentioned before, the projection space before the model sparse could be represented as:

$$Er = \sum_{i=1}^m (p_{id} \sigma_i \xi_i \lambda_i^\top + p_{ood} \tilde{\sigma}_i \xi_i \gamma_i^\top) \quad (28)$$

SFP prunes the model by trimming the smallest singular values in  $\Sigma$  as well as their corresponding left and right singular vectors. In this way, SFP could remove the spurious features in ID data space and substructures in the model space simultaneously in a targeted manner

along the directions with weaker actions for projection. The projection space after sparse with only the most important  $\vartheta$  singular values can be formalized as:

$$Er^{sparse} = \sum_{i=1}^{\vartheta} p_{id} \sigma_i \xi_i \lambda_i^\top + \sum_{j=1}^m p_{ood} \tilde{\sigma}_j \xi_j \gamma_j^\top. \quad (29)$$

Based on the representation of the projection spaces, the model response to data features  $\mathcal{R}(X) = ErX$  can be calculated as:

$$\mathcal{R}(X) = \{p_i \Xi \Sigma_{E^\top F} \Lambda^{-1} + p_o \Xi \Sigma_{E^\top G} \Gamma^{-1}\}^\top X^\top \quad (30)$$

□

#### A.4 Proofs for the correspondence between model substructure and spurious features

Specifically, we define  $f^l(x)$  as the feature maps output of  $x$  at layer  $l$ . It represents the projection of  $x$  onto the model space defined over the spanning set  $E$  to be learned. We abbreviate the final probabilities as  $f(x)$  for simplification. Referring to Sec. 3.2.1, we have  $x \in X_{id}$  if  $\mathcal{L}_{ce}(x) \leq \Delta$ . Thus, the optimization target of SFP can be formulated as:

$$\min_E \mathbb{E}_{x \sim X} \mathcal{L}_{ce}(x, \mathcal{W}) + \eta \sum_{l=1}^L \mathbb{E}_{x \sim X_{id}} \|f^l(x)\|_1, \quad (31)$$

where  $\eta$  is the sparsity factor imposed on the feature projections for the identified ID data. It serves as an adjustable weight to calibrate the feature response of ID data, as well as sparse the corresponding substructures.

**Lemma A.7.** (3.8) Define  $e = |f^*(x) - f(x)|$  as the  $l_1$ -norm between the true distribution  $f^*(x)$  and  $f(x)$ . When  $\eta < 2e$ , SFP could effectively reduce the learning of the model to spurious features but keep the performance on the same features.

PROOF OF LEMMA A.7. The prediction errors of feature projections  $L_f$  can be defined as:

$$\begin{aligned} L_{ce} &= |f^*(x) - f(x)|^2 \\ &= \sum_{i,j=j_1 \cup j_2} (f^*(x) - \sigma_{i,j_1} \xi_i \top \lambda_{j_1} - \sigma_{i,j_2} \xi_i \top \gamma_{j_2})^2, \end{aligned} \quad (32)$$

and the corresponding gradient is:

$$\begin{aligned} \frac{\partial L_{ce}}{\partial \sigma_{i,j_1} \xi_i} &= \frac{\partial e^2}{\partial \sigma_{i,j_1} \xi_i} = 2e \frac{\partial e}{\partial \sigma_{i,j_1} \xi_i} \\ &= 2e \frac{|f^*(x) - \sigma_{i,j_1} \xi_i \top \lambda_{j_1} - \sigma_{i,j_2} \xi_i \top \gamma_{j_2}|}{\partial \sigma_{i,j_1} \xi_i} \\ &= -2e \lambda_{j_1}, \end{aligned} \quad (33)$$

where  $i$  and  $j$  are the index of column vectors in the orthogonal basis for model space and feature space, respectively. For OOD data, the gradient of the column vectors in the OOD projection matrix interacting with the  $j_{th}$  feature vector is  $-2e \gamma_{j_2}$ .

Therefore, for all data samples in the training set, the update of the  $i_{th}$  direction vector of the projection matrix at round  $t$  is:

$$\sigma_{i,j} \xi_i^t = \sigma_{i,j} \xi_i^{t-1} - p_i (-2e \lambda_{j_1}) - p_o (-2e \gamma_{j_2}) + p_i \eta \lambda_{j_1} \quad (34)$$

Split the in-domain features into the spurious features  $F'$  and the invariant features  $IN$ , and split the out-of-domain features into the unknown features  $G'$  and the invariant features  $IN$ . Since the environment features in-domain and out-domain are different with high probability under the OOD setting, we suppose  $F'$  and  $G'$  are orthogonal and define  $a, b \in \Xi$  as the column vectors interact with  $F'$  and  $G'$  respectively. The updates of  $a, b$  could be formulated as:

$$\begin{aligned} \sigma_{a,F'} \xi_a^t &= \sigma_{a,F'} \xi_a^{t-1} - p_i (-2e \lambda_{F'}) - p_i \eta \lambda_{F'} \\ \sigma_{b,G'} \xi_b^t &= \sigma_{b,G'} \xi_b^{t-1} - p_o (-2e \gamma_{G'}) \end{aligned} \quad (35)$$

Also, define  $c \in \Xi$  to be the set of the column vectors in the projection matrix that interacts with invariant features that are consistent in domain and out of domain, and the updates of  $c$  can be computed as:

$$\sigma_{c,IN} \xi_c^t = \sigma_{c,IN} \xi_c^{t-1} + 2e p_i \lambda_{IN} + 2e p_o \gamma_{IN} - p_i \eta \lambda_{IN} \quad (36)$$

To achieve spurious features-targeted unlearning and invariant features-targeted learning of the model, the following constraints need to be satisfied:

$$\begin{aligned} 2e p_i \lambda_{IN} + 2e p_o \gamma_{IN} - p_i \eta \lambda_{IN} &> 2e p_o \gamma_{G'} \\ \Rightarrow \eta &\leq \frac{2e p_i \lambda_{IN} + 2e p_o \gamma_{IN} - 2e p_o \gamma_{G'}}{p_i \lambda_{IN}} \approx 2e \end{aligned} \quad (37)$$

Since the de-learning rate of the spurious feature is positively correlated with  $\eta$ , the upper bound  $\eta = 2e$  is taken in this work. □

## B ABLATION EXPERIMENTS

In this section, we mainly focus on two aspects: the initialization of the dense model, and the mapping relation versatility: background-label mapping relation in the biased samples' setting.

### B.1 The initialization of the dense model

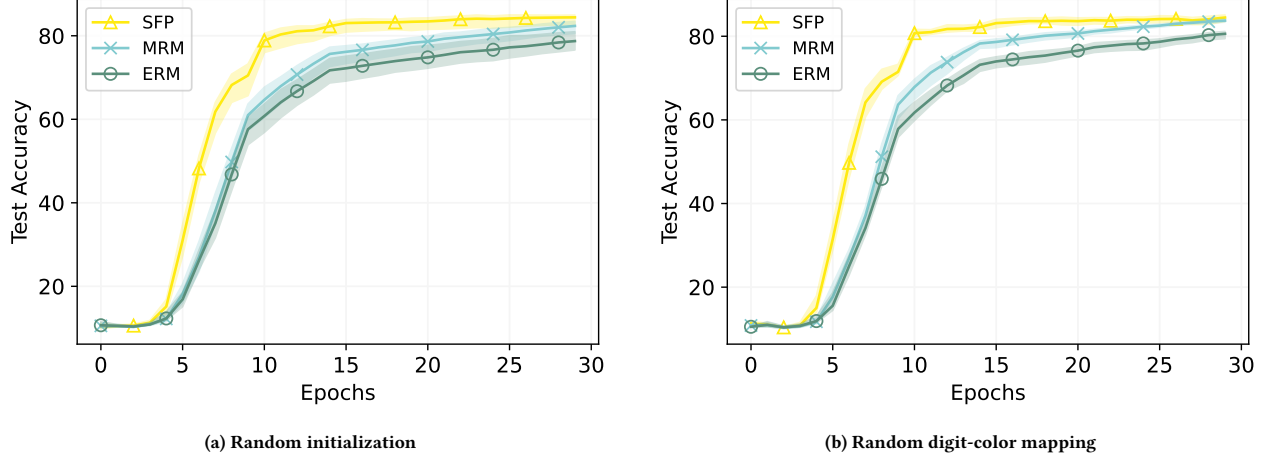


Figure 8: The effect of some random settings on the performance of SFP.

First, to demonstrate the consistent performance of the proposed SFP regardless of the randomness of the experiment environments, we used ten different random seeds to initialize the deep learning model and record the final accuracy on FullColoredMNIST (FCMNIST). We use (0.9, 0.7, 0.0) as the biased ratio coefficient in this experiment. Fig. 8a illustrates the mean accuracy of 10 different experiment settings. We notice that the proposed SFP outperforms MRM and ERM in all datasets. This indicates that SFP can successfully sparse the spurious feature-associated network structure regardless of different model initializations.

### B.2 The mappings relations of OOD samples

In our experiment setting, we use the biased data samples, which have a static one-to-one digit-color relationship, as the ID data samples. On the contrary, the OOD data samples have randomly assigned colored backgrounds. To further demonstrate that our method can successfully prune spurious features regardless of the mapping relations, we evaluate the test accuracy in 5 different mapping relation settings and draw the mean accuracy in Fig. 8b. As the experiment result shows, the average accuracy of SFP is relatively higher, and the variance of the accuracy is relatively lower, which shows the superiority of the proposed SFP is stable and robust. This suggests that SFP successfully prunes the sub-network associated with any spurious feature.

### B.3 The feature responses of spurious correlations

Furthermore, to validate the effectiveness of SFP in suppressing the learning of spurious features, we examine the progression of the network's feature responses to in-domain samples across the entire training trajectory. The response values are measured by the average attention across all feature channels at each layer. Specifically, we introduced a channel attention mechanism, named Squeeze-and-Excitation (SE) module [14], to score the channel saliency of feature maps for input  $x_i$ . The computed channel saliencies, denoted as  $\pi_l(x_i)$ , are numerical values produced by a Sigmoid function, ranging from 0 to 1. For models trained with ERM on unbiased data, the expected average attention values for the feature channels at each layer are 0.5. These values represent the relative importance of the corresponding feature channel, with smaller values denoting reduced importance. In summary, the sparsity of channel saliency determines the number of effective filters for structures. For inputs, the mean of these channel attentions indicates the models' fitting degree to the current samples. We conducted the experiments on ResNet-18 and ColoredMNIST.

The results are shown in Fig. 9. As the training progresses, SFP gradually weakens the feature responses to spurious correlated data, while under ERM and MOD methods, this response shows no significant changes. The failure of ERM is attributed to its inclination to learn all correlations indiscriminately to enhance predictive accuracy. On the other hand, the failure of the MOD method, as a structured OOD approach, lies in its utilization of existing pruning techniques without specific enhancements for OOD attributes. These pruning methods often lack feature specificity, meaning they do not consider the correspondence between the structure and features. Consequently, they



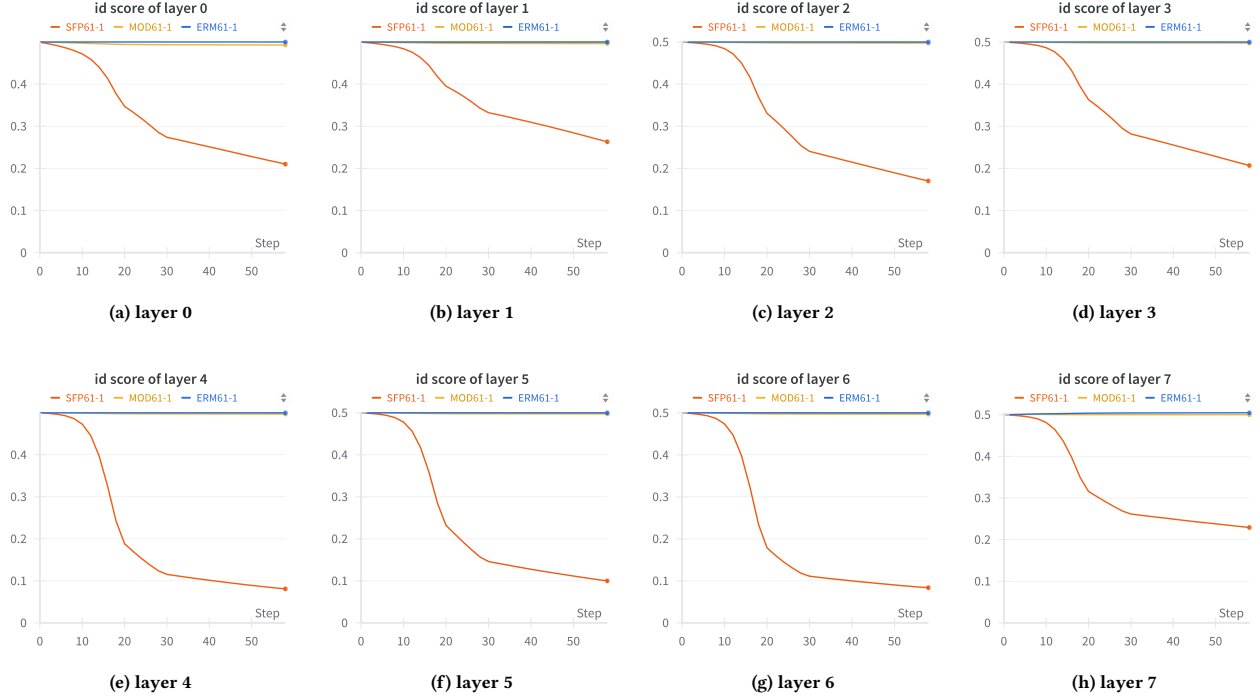


Figure 9: The feature response intensity to in-domain samples at different layers of the model.

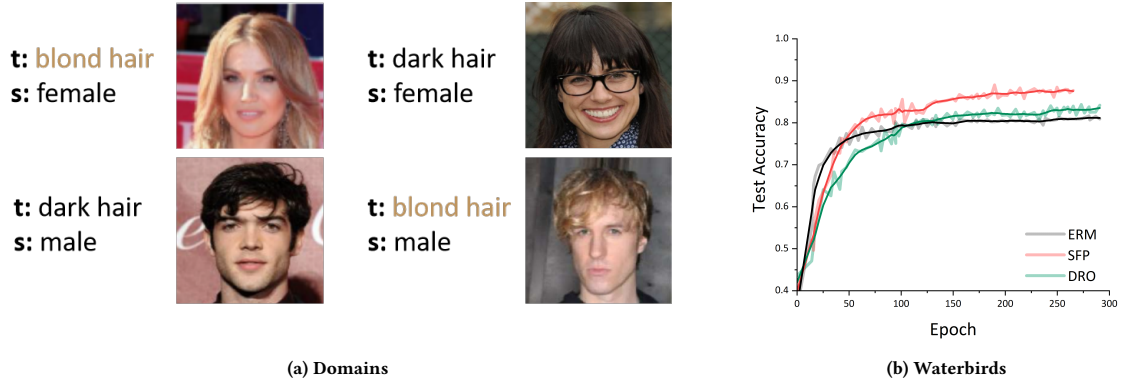


Figure 10: The average testing accuracy over different domains on CelebA. The hair color blond, dark is used as the target, and the gender male, female is used as the spurious attribute. The smallest combination group is blond-haired males.

apply the same sparse penalty to branches responding to invariant and spurious features simultaneously. In contrast, SFP, designed with OOD attributes in mind, employs feature-specific network pruning. Consequently, it sidesteps the above-mentioned issues.

## C ADDITIONAL EXPERIMENTAL RESULTS

### C.1 Dataset details

In this section, we will provide a clear description of the non-domainbed datasets in the main paper, including three synthetic dataset - FullColoredMNIST, ColoredObject, and SceneObject, and two real-world datasets - CelebA and Waterbirds.

- **FullColoredMNIST** is a ten-class biased variant of the original MNIST dataset [42]. The digit shapes serve as invariant features while colors as spurious ones. Ten different colors were selected to define a one-to-one corresponding relationship with ten-digit

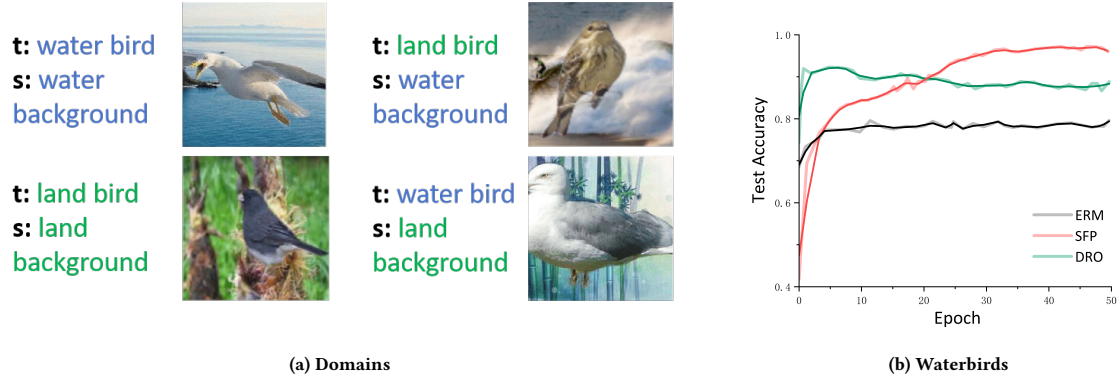


Figure 11: The average testing accuracy over different domains on Waterbirds. The bird species waterbird, landbird are used as the label, while the bird's locations water background, land background are used as a spurious attribute. The smallest domain is waterbirds on land.

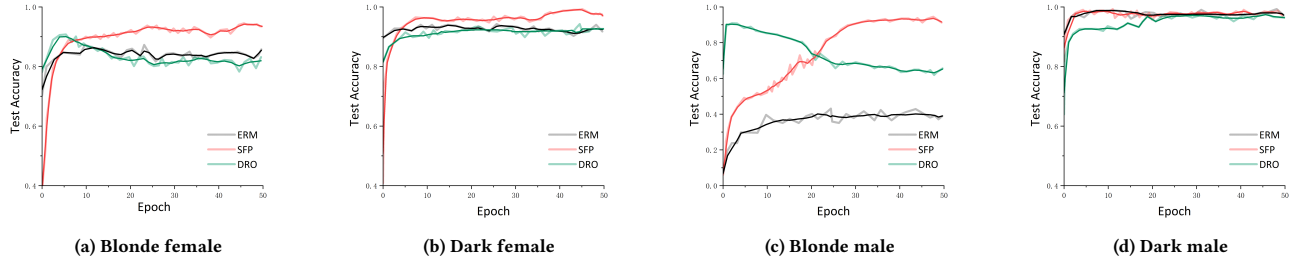


Figure 12: Testing accuracy of different domains on CelebA.

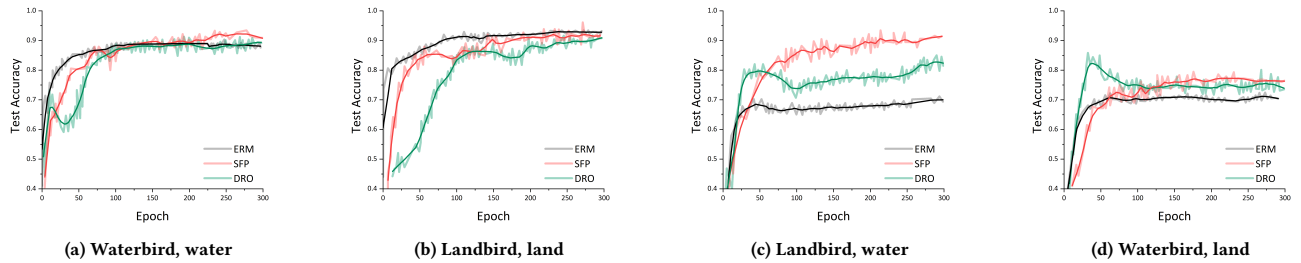


Figure 13: Testing accuracy of different domains on Waterbirds.

classes (e.g.,  $2 \leftrightarrow \text{green}$ ,  $4 \leftrightarrow \text{yellow}$ ). For each domain, a bias coefficient is defined to represent the ratio of images adhering to this specific relationship, with non-conforming images randomly colored.

- **ColoredObject** is constructed by superimposing ten classes of objects extracted from the MSCOCO dataset [23] onto backgrounds of ten distinct colors [42]. These ten classes of objects include boats, airplanes, trucks, dogs, zebras, horses, birds, trains, buses, and motorcycles. The spurious correlation is defined as the one-to-one correspondence between objects and colors.
- **SceneObject** [42] consists of ten classes of objects extracted from the MSCOCO dataset, which are placed into ten scenic backgrounds from the Places dataset [45]. These scenic backgrounds render the task more complex compared to ColoredObject. Similar to FULLCOLOREDMNIST, SceneObject establishes a one-to-one object-scene relationship, making it more biased and consequently more challenging than previous tasks.
- **CelebA** dataset is a widely-used celebrity face dataset with 162770 training examples [26]. It contains 40 attribute labels (like "Smiling", "Wearing Hat", etc.) Following previous OOD works [24, 32, 35], we classify hair color as either blonde or non-blond, a

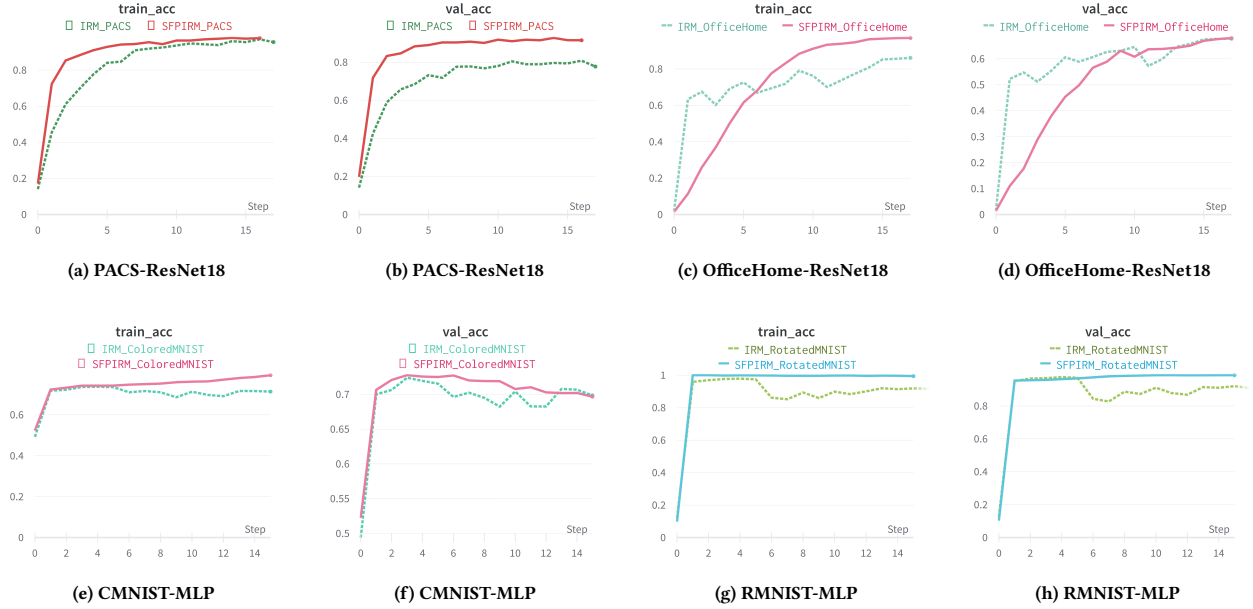


Figure 14: The accuracy comparison of SFP+IRM and IRM.

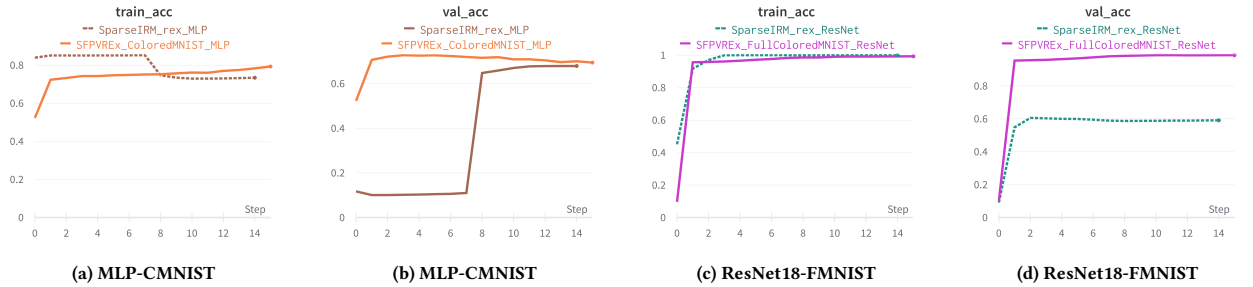


Figure 15: The accuracy comparison of SFP+VREx and SparseIRM+VREx.

feature spuriously associated with the gender binary of the celebrities (male or female). The training set is divided into four domains, including drak-haired females, blond-haired females, dark-haired males, and blond-haired males, with 1387 in the smallest group (blond-haired males).

- **WaterBirds** is a subset of the Caltech-UCSD Birds-200-2011 dataset [37] with 4795 training examples, specifically constructed for studying image recognition with spurious correlations of backgrounds [32]. It incorporates images of waterbirds and landbirds from the Caltech-UCSD Birds-200-2011 (CUB) dataset as the foreground, paired with either water or land backgrounds obtained from the Places dataset. The training set is divided into four domains, including landbirds on land, waterbirds on water, landbirds on water, and waterbirds on land, with 56 in the smallest group (waterbirds on land).

## C.2 Evaluation on more datasets

We further expanded the evaluation scope of SFP to two real-world datasets: WaterBirds and CelebA. The experiment is divided into two groups, including comparisons of average accuracy across all domains and comparisons of the accuracy on each individual domain.

Fig. 10 and Fig. 11 illustrated the comparative results of cross-domain average testing accuracy based on the CelebA and Waterbirds datasets, respectively. First, we visualized each domain of CelebA on Fig. 10a, and Waterbirds on Fig. 11a. Subsequently, we compare the cross-domain average accuracies of different methods in Fig. 10b (CelebA) and 11b (Waterbirds). The results demonstrate the superior performance of our proposed method, which reaches a remarkable accuracy of 96.41% on CelebA and 88.13% on Waterbirds. Specifically,

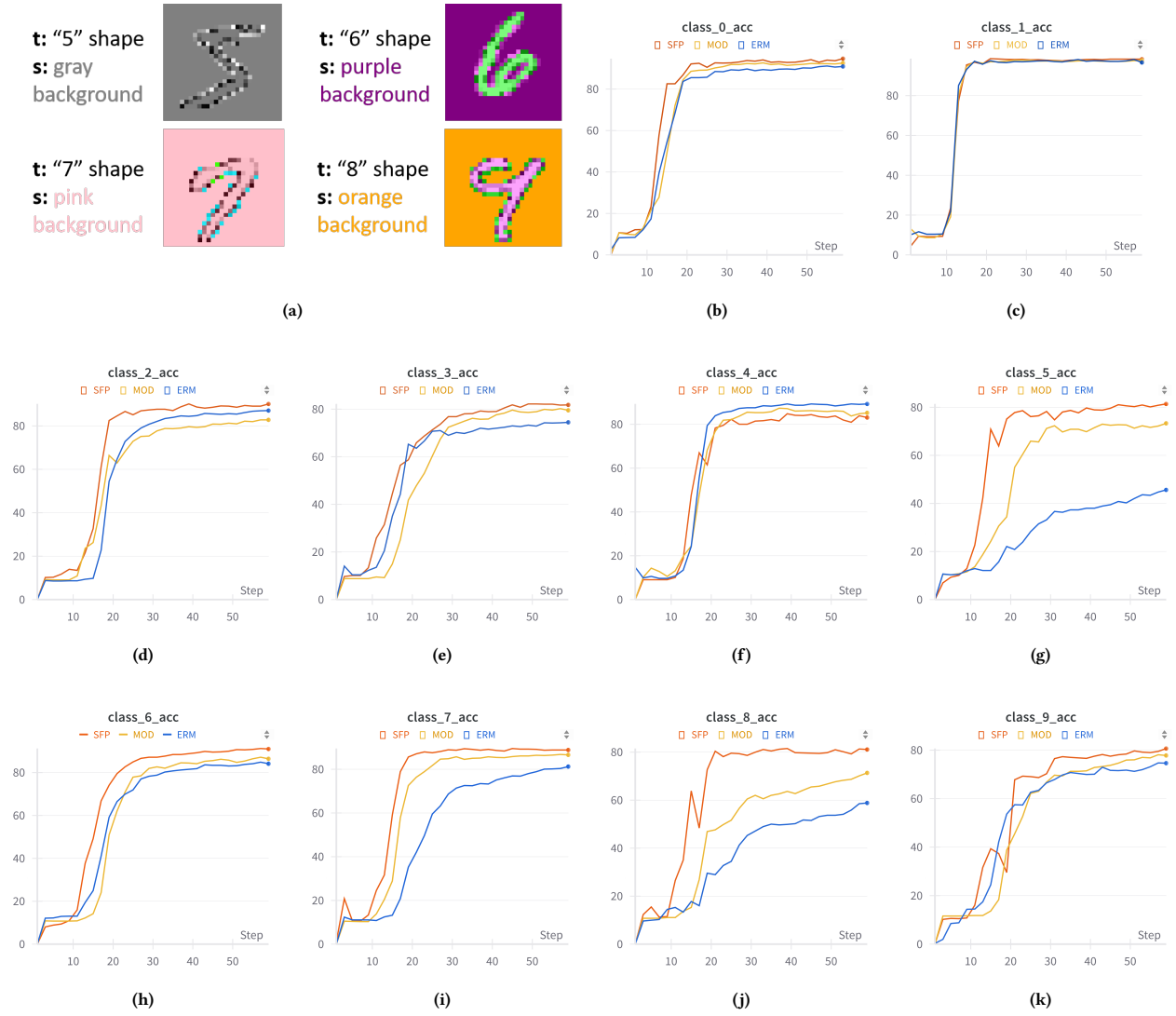


Figure 16: The performance of SFP across various domains with distinct classes.

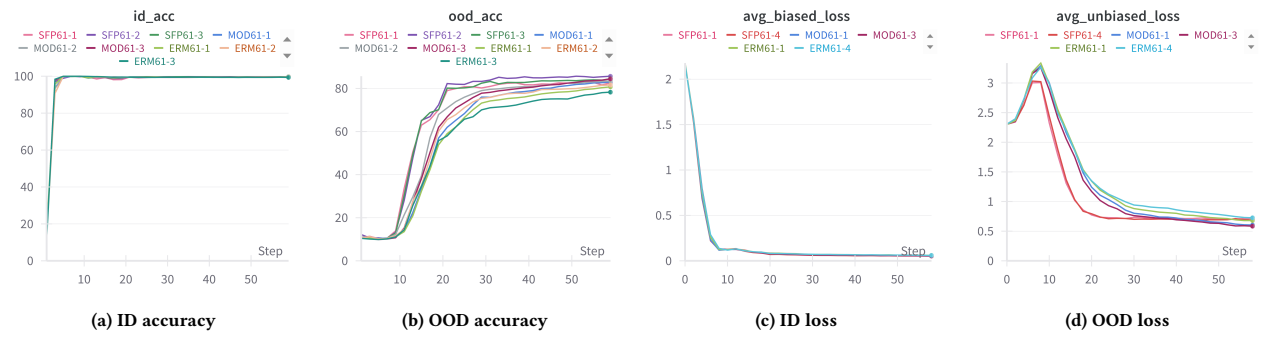


Figure 17: The evaluation of SFP on different domains.

SFP’s cross-domain average accuracy on the CelebA dataset surpasses ERM by 16.8% and DRO by 6.45% (A.b), while on the Waterbirds dataset, it exceeds ERM by 7.23% and DRO by 4.15%.

Further insights into the testing accuracy of individual domains based on the CelebA and Waterbirds datasets are provided in Figures C and D, respectively. Within this framework, the models were trained across all domains while tested on individual domains, with inter-domain sample quantities varied. Specifically, the CelebA dataset encompasses four domains: blonde-haired female, blond-haired male, dark-haired female, and dark-haired male, with the fewest samples found in the blond-haired male domain. Similarly, the Waterbirds dataset consists of four domains: waterbirds on water background, waterbirds on land background, landbirds on water background, and landbirds on land background, with the fewest samples observed in the waterbirds on land background domain. In this context, SFP consistently demonstrates a satisfied performance. As depicted in Fig. 12c and Fig. 12c, SFP achieves a substantial accuracy improvement within the smallest domain, exhibiting a remarkable 42.89% increase on CelebA (with blond male, ERM) and 20.28% increase on Waterbirds (with landbird on water background, ERM). Across the remaining three domains, test accuracies are comparably high. SFP improves the test accuracy by a minor 3.9% over DRO on the blond-haired female domain and 7.02% over ERM on the waterbirds-on-land domain.

**Discussion:** It can be seen that while ERM demonstrates satisfactory performance across multiple domains, it exhibits subpar performance within the smallest domain. What’s more, despite DRO achieving an enhancement in average accuracy, it sacrifices performance within specific groups to bolster accuracy within others, exemplifying instances of trade-offs. In contrast, SFP achieves the most robust generalization accuracy by iteratively learning invariant features through feature-oriented model pruning, thereby outperforming the other methods.

### C.3 Evaluation on combined methods

To demonstrate the orthogonal effect of SFP to non-structure OOD methods, we evaluate the performance of SFP combined with other non-structure OOD methods. The experiments are conducted on DomainBed, including PACs, COLOREDMNIST, OfficeHome, FullCOL-OREDMNIST, and RotatedMNIST. It can be seen that SFP achieves superior performance among all the competitors, and the improvements are significant in some cases. To be specific, Fig. 14 illustrates the performance comparison between SFP+IRM and the original IRM. It can be seen that SFP outperforms IRM with 9.52% on ‘PACS-ResNet18’, 5.41% on ‘COLOREDMNIST-MLP’, 0.92% on ‘OfficeHome-ResNet18’, and 4.45 % on ‘RotatedMNIST-MLP’. The superior results on IRM and VREx demonstrate the orthogonal enhancement effect of SFP.

Fig. 15 presents the performance comparison between SFP+VREx and SparseIRM+VREx. SFP outperforms SparseIRM with 3.41% higher test accuracy on MLP and even 29.12% on ResNet18. An interesting phenomenon is that, on small MLP, SparseIRM first shows overfitting during the training stage, and then, after 7 ( $\times$  300 iterations) steps, there is a significant increase in test accuracy. The training process of SparseIRM exhibits an obvious two-stage trend, which is the same as regular non-feature-targeted model pruning. Differently, SFP consistently shows a stable learning curve and achieves higher performance in both ID (train) and OOD (test) environments. In summary, the comparison results demonstrate that SFP has achieved preminent performance across most structure-based OOD generalization methods.

### C.4 Evaluation on different domains

We further evaluated the performance of the SFP on each domain to explore whether varying invariant targets under the same intensity of spurious features affect OOD generalization. The results are presented in the Fig. 16. We first illustrate several examples from the training and testing sets. In the training set, most samples align with the previously described in-domain data settings, establishing a one-to-one relationship between the background and the label for each digit. Each testing set contains digits of only one category, with the background of that digit category differing from the training domain.

In this setup, we evaluated the performance of SFP on different test domains and compared the results with the ERM method and another popular structured method MRM. Firstly, vertically within the same graph, SFP achieved better generalization performance than MRM in almost all cases. Secondly, horizontally across different graphs, there was a significant difference in the accuracy improvement of SFP over ERM across different classes. In examples where invariant features such as digits 0 and 1 are relatively easy to learn, the three methods showed comparable accuracy. This indicates that ERM and MOD also learned invariant features well. However, in examples where invariant features such as digits 5 and 8 are more challenging to learn, SFP outperformed MOD and ERM significantly. This not only highlights the strong OOD generalization performance of SFP but also reflects the poor mastery of ERM in complex invariant feature scenarios. The experimental conclusions in this section align with previous works [5, 27]: neural networks trained with the ERM inherently learn both invariant and spurious features, but they tend to prioritize shortcuts at the early stage.

Similarly, we also split the dataset as in-domain and out-of-domain and tracked the performance separately during training. The results are shown in Fig. 17. We conducted two evaluations with different initializations for each method. It can be observed that in all cases, SFP achieved outstanding OOD generalization performance, surpassing the baseline methods by approximately 5% during convergence. Moreover, we compared the task loss of SFP and other methods across different data domains, and the results further validated the superior performance of the proposed approach.