SLIM: Spuriousness Mitigation with Minimal Human Annotations

Xiwei Xuan[1], Ziquan Deng[1], Hsuan-Tien Lin[2], and Kwan-Liu Ma[1]

[1] University of California, Davis
[2] National Taiwan University
{xwxuan, ziqdeng, klma}@ucdavis.edu, htlin@csie.ntu.edu.tw

A Appendix

In this appendix, we use lowercase letters, lowercase boldface letters, and uppercase boldface letters to respectively denote scalars (a), vectors ($\mathbf{v}$), and matrices ($\mathbf{W}$). The appendix is organized as follows:

- Sec. A.1 presents the preliminary knowledge underpinning our interpretation in Sec. A.2.
- Sec. A.2 presents a theoretical interpretation to further support our choice of preserving data that receives correct attention to enhance core feature learning.
- Sec. A.3 presents the theorem that grounds our interpretation in Sec. A.2.
- Sec. A.4 provides a more detailed description of the datasets we used for the experiments.
- Sec. A.5 presents the training setting for our experiments.
- Sec. A.6 provides the interface and instructions for our crowdsourcing tasks.
- Sec. A.7 provides the definition of AIoU score.
- Sec. A.8 offers more examples to evaluate attention consistency in the original feature space, our constructed attention space, and the environmental feature spaces. Additionally, we provide further examples to validate *SLIM*'s enhanced attention accuracy.
- Sec. A.9 discusses the limitations of our solution.

A.1 Proof Preliminaries

To simplify the complex real-world issue of spurious correlations into a formal framework, in alignment with previous works [2, 4, 6], we adopt a two-layer nonlinear convolutional neural network (CNN) based on a data model that captures spurious correlations. The two-layer CNN is defined as follows:

$$f(\mathbf{x}; \mathbf{W}) = \sum_{j \in [J]} \sum_{p=1}^{P} \sigma(\langle \mathrm{w}_j, \mathrm{x}^{(p)} \rangle), \tag{A1}$$

where $\mathbf{w}_j \in \mathbb{R}^d$ is the weight vector of the j-th filter, J is the number of filters (neurons) of the network, and $\sigma(z) = z^3$ is the activation function.

$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_J] \in \mathbb{R}^{d \times J}$ denotes the weight matrix of the CNN. In [1, 4, 6], they assume a mild overparameterization of the CNN with $J = polylog(d)$ and initialize $\mathbf{W}^{(0)} \sim N(0, \sigma_0^2)$, where $= polylog(d)/d$.

To understand the underlying dynamics in feature learning, we introduce the following data model where the input consists of a core feature, a spurious feature, and noise patches.

Definition 1 (Data model. [4]). *A data point* $(\boldsymbol{x}, y, s) \in (\mathbb{R}^d)^P \times \{\pm 1\} \times \{\pm 1\}$ *is generated from the distribution* $\mathcal{D}$ *as follows.*

- *Randomly generate the true label* $y \in \{\pm 1\}$.
- *Generate spuriousness label* $s \in \{\pm y\}$, *where* $s = y$ *with probability* $\alpha > 0.5$.
- *Generate* $\boldsymbol{x}$ *as a collection of* P *patches:* $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(P)}) \in (\mathbb{R}^d)^P$, *where*
  - *Core Feature. One and only one patch is given by* $\beta_c \cdot y \cdot \boldsymbol{v}_c$ *with* $||\boldsymbol{v}_c||_2 = 1$. β_c *is the core feature strength.*
  - *Spurious Feature. One and only one patch is given by* $\beta_s \cdot s \cdot \boldsymbol{v}_s$ *with* $||\boldsymbol{v}_s||_2 = 1$ *and* $\langle \boldsymbol{v}_c, \boldsymbol{v}_c \rangle = 0$. β_s *is the spurious feature strength.*
  - *Random noise. The rest* P-2 *patches are Gaussian noises* ξ *independently drawn from* $N(0, (\sigma_p^2/d) \cdot \boldsymbol{I}_d)$ *with* σ_p *as an absolute constant.*

With the given data model, considering the training dataset $S = \{(\mathbf{x}_i, y_i, a_i)\}_{i=1}^N$ and let S be partitioned into large group S_1 and small group S_2 such that S_1 contains all the data that can be correctly classified by the spurious feature, i.e., $s_i = y_i$, and S_2 contains all the data that can only be correctly classified by the core feature, i.e., $s_i = -y_i$. Denote $\hat{\alpha} = \frac{|S_1|}{N}$ and therefore $1 - \hat{\alpha} = \frac{|S_2|}{N}$.
Remark. Different from the original definition in [4], we do not make assumptions about the relative strengths of β_c and β_s. Rather, our approach estimates the relative strengths of β_c and β_s through attention correctness annotations, as the saliency map can reflect the features learned by the model.

A.2 Theoretical Inspiration

In Sec. 5, we have demonstrated the robustness of annotating attention correctness and corroborated that decoupling core and environment features is crucial for learning core features. In building feature-balanced datasets, our approach primarily focuses on leveraging data that receives correct attention from the reference model. One reason for this is that environment features can be more accurately and efficiently isolated based on the identified core features. Adopting the **Theorem** 1 proposed in [4], we interpret it from a different perspective to further support and justify that preserving data with high attention scores guarantees the effective learning of core features in a more balanced dataset.

Lemma 1. *Under training dataset* S, *which follows the distribution described in* ***Definition*** *1, when the data is trained using gradient descent for* $T_0 = \tilde{\Theta}(\eta)(1/\eta\beta_s^3\sigma_0)$ *iterations on the model as introduced in Eqn. A1, instances receiving higher attention scores are more likely to have their core features learned in a new training scenario with a more balanced data distribution (i.e.,* $\hat{\alpha} \to 1/2$*).*

Proof. Based on **Theorem** 1, we are implied that in the early T_0 iterations

$$\beta_c \ll \beta_s \sqrt[3]{2\hat{\alpha}-1} \Rightarrow P_{lrc} \to 0, \tag{A2}$$

where P_{lrc} is the probability of the model learned the core feature. Thus, the converse-negative proposition of proposition (A2) is

$$P_{lrc} > 0 \Rightarrow \beta_c \not\ll \beta_s \sqrt[3]{2\hat{\alpha}-1}, \tag{A3}$$

in the initial T_0 iterations. For instance, $\mathbf{x}_i$, whose core feature has been effectively learned, there should exist a constant threshold, denoted as $Tr(\mathbf{x}_i)$. This threshold ensures that the core feature has the chance to be learned once $\beta_c - \beta_s \sqrt[3]{2\hat{\alpha}-1} > Tr(\mathbf{x}_i)$. Intuitively, $P_{lrc}(\mathbf{x}_i) \propto (\beta_c(\mathbf{x}_i) - \beta_s(\mathbf{x}_i)\sqrt[3]{2\hat{\alpha}-1})$. Since the strengths of the core feature $\beta_c(\mathbf{x}_i)$ and spurious feature $\beta_s(\mathbf{x}_i)$ are natures of the data itself and do not change, in a more balanced data distribution, as $\hat{\alpha} \to 1/2$, $(\beta_c(\mathbf{x}_i) - \beta_s(\mathbf{x}_i)\sqrt[3]{2\hat{\alpha}-1})$ is increasing, consequently $P_{lrc}(\mathbf{x}_i)$ is increasing. Since the learned feature of an instance can be interpreted via saliency maps, a higher attention score means that its core feature has been learned more accurately. Therefore, such instances have a higher probability of the core feature being continuously learned as the data distribution becomes more balanced. □

Although this theoretical insight is built on a simplified binary classification model, it provides an inspirational hint towards understanding the benefit of utilizing data with high attention scores in more complex scenarios.

A.3 Auxiliary Theorem

Theorem 1. *(Theorem 2.2 in [4].) Consider the training dataset S that follows the distribution in Definition 1. Consider the two-layer nonlinear CNN model as in Eqn. (A1) initialized with $\boldsymbol{W}^{(0)} \sim N(0, \sigma_0^2)$. After training with gradient decent for $T_0 = \tilde{\Theta}(1/\eta\beta_s^3\sigma_0)$ iterations, for all $j \in [J]$ and $t \in [0, T_0)$, we have*

$$\tilde{\Theta}(\eta)\beta_s^3(2\hat{\alpha}-1)\langle \boldsymbol{w}_j^{(t)}, \boldsymbol{v}_s\rangle^2 \le \langle \boldsymbol{w}_j^{(t+1)}, \boldsymbol{v}_s\rangle - \langle \boldsymbol{w}_j^{(t)}, \boldsymbol{v}_s\rangle \le \tilde{\Theta}(\eta)\beta_s^3\hat{\alpha}\langle \boldsymbol{w}_j^{(t)}, \boldsymbol{v}_s\rangle^2, \tag{A4}$$

$$\tilde{\Theta}(\eta)\beta_c^3\hat{\alpha}\langle \boldsymbol{w}_j^{(t)}, \boldsymbol{v}_c\rangle^2 \le \langle \boldsymbol{w}_j^{(t+1)}, \boldsymbol{v}_c\rangle - \langle \boldsymbol{w}_j^{(t)}, \boldsymbol{v}_c\rangle \le \tilde{\Theta}(\eta)\beta_c^3\langle \boldsymbol{w}_j^{(t)}, \boldsymbol{v}_c\rangle^2. \tag{A5}$$

With the updates of the spurious and core feature in the early iterations, **Theorem** 1 gives the condition-if $\beta_c^3 < \beta_s^3(2\hat{\alpha}-1)$-that GD will learn the spurious feature very quickly while hardly learning the core feature.

A.4 Datasets

Waterbirds [11]. It is constructed to study the spurious correlation between the image background and the object. To this end, bird images in Caltech-UCSD

Birds-200-2011 (CUB-200-2011) dataset [12] are grouped into waterbirds and landbirds. All birds are then cut and pasted onto new background images from the Places dataset [19], with waterbirds having a higher probability on water and landbirds having a higher probability on land. The training set contains 4,795 images in total, 3,498 for landbirds with land background, 184 for landbirds with water background, 56 for waterbirds with land background, and 1,057 for waterbirds with water background. The validation set contains 1,199 images in total, 467 for landbirds with land background, 466 for landbirds with water background, 133 for waterbirds with land background, and 133 for waterbirds with water background.

CelebA [8]. It is a large-scale face attribute dataset comprised of photos of celebrities. Each image is annotated with 40 binary attributes. Aligned with other works focusing on spuriousness mitigation, we chose "blond hair" or "non-blond hair" as the target attributes, and gender as the spurious feature for hair color classification. The training set contains 162,770 images in total, 71,629 for non-blond haired female, 66,874 for non-blond haired male, 22,880 for blond haired female, and 1,387 for blond haired male. The validation set contains 19,867 images in total, 8,535 for non-blond haired female, 8,276 for non-blond haired male, 2,874 for blond haired female, and 182 for blond haired male.

ISIC [3]. It contains images of a skin lesion, categorized into (1) benign lesions or (2) malignant lesions. In a real-life task, this would be done to determine whether a biopsy should be taken. Aligning with previous studies [10], we target colorful patches as spurious features, and also follow the same strategy to obtain data from its official platform.

NICO. Derived from NICO++ [18], this dataset features various object categories in shifted contexts to probe spurious correlations. It is a multi-class image dataset presenting a diverse set of objects in varied contextual scenarios, allowing convenient adjustment of the distributions of object and context labels. We randomly sample eight animal categories with eight different contextual labels. To challenge the model with spurious correlations, the training set distribution follows three rules: (a) each object class is distributed across various contexts; (b) there is one dominant context for each single class; (c) the dominant context for each class is unique. For instance, most "sheep" images have context "grass" and "grass" context is only dominant in "sheep". The detailed distribution is shown in Fig. A1.

ImagetNet9 (IN9) [14]. This dataset is a curated subset extracted from the larger ImageNet collection, specifically designed to scrutinize and address the model bias towards object backgrounds. To evaluate our framework for spurious correlation mitigation, we adopt the "Mixed-Rand" setup, which is a particular data arrangement where images are organized to have a randomized correlation between the object and its background. This setup aims to challenge models to focus on the object itself rather than the background, helping to test and improve the robustness of models against spurious correlations. We utilize the training and validation splits provided by ImageNet9, ensuring that our experiments are aligned with established benchmarks for consistency and comparability. More-

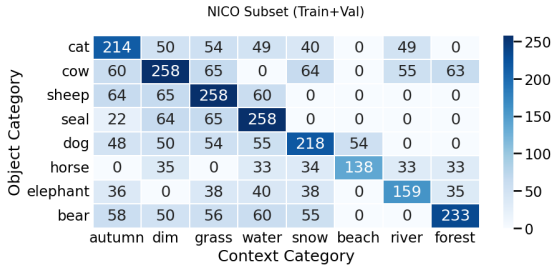


Fig. A1: Data distribution in the combination of training and validation set of NICO, with respect to object and context categories.

over, their provided model trained on IN-9L is used as our reference model. The outcomes of our experiments are then evaluated on the Mixed-Rand.

A.5 Training Setting

In Table 4, we present the amount of data required for attention annotation and the size of the constructed data used for model training. In this section, we provide additional details on the training settings. To maintain consistency with existing methods, we use SGD as the optimization algorithm. The hyper-parameter ranges, batch sizes, and training epochs used in our experiments are tuned according to these methods [4, 7, 11, 15, 17], as listed in Table A1. The training setups corresponding $SLIM_{Val}$ are following DFR [7]. The experiments were conducted on two NVIDIA RTX 4090 GPUs with 24GB memory.

Table A1: Hyperparameters used for the $SLIM_{Tr}$'s results in Sec. 5.2 on different datasets.

Dataset	Waterbirds	CelebA	ISIC	NICO	ImageNet 9
Initial lr	1E-3	1E-4	0.002	1E-6	1E-6
Weight Decay	0.1	0.1	0.1	0.5	0.1
Batch Size	128	128	128	128	128
Training Epochs	50	30	30	50	10
Core Cluster	2	2	2	8	9
Env Cluster	3	3	2	10	9

A.6 Crowdsourcing Instruction

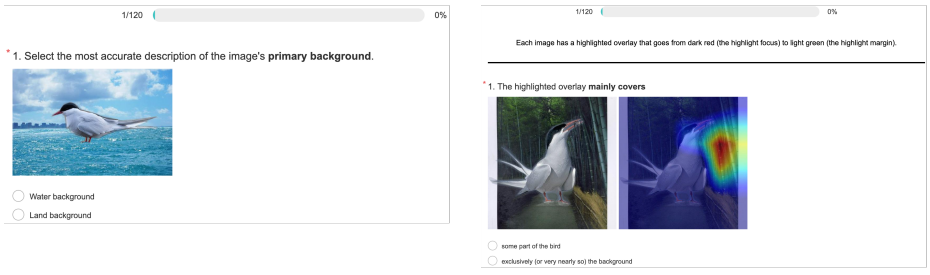
In Sec. 5.3, we employed crowdsourcing tasks with the Waterbirds and NICO datasets to compare the consistency of annotating spuriousness versus attention correctness. For this study, we selected a random set of 120 images from each dataset and established two separate tasks: one for spuriousness labeling

and another for attention correctness labeling. For each task, we recruited 60 participants who are native English speakers, independently from the Prolific platform, to prevent learning biases from cross-task participation. Participants received an hourly fee for their participation. In ensuring ethical research standards, our study refrained from collecting personally identifiable information and excluded any potentially offensive content.

We provided the following instruction to the spuriousness labeling task participants: "This study focuses on evaluating the image annotation tasks. Participants will be presented with a series of 120 images featuring different animals. For each image, the task involves selecting the most accurate description of the primary background from the provided options. There are no specific prerequisites for participation. Simply make selections based on your observation. Notice: This study ensures the confidentiality of your participation, as it neither collects personally identifiable information nor contains any offensive content. Your feedback will exclusively be utilized for academic research purposes."

We provided the following instruction to the participants for the attention correctness labeling task: "This study aims to evaluate a Machine Learning Model's attention correctness. Participants will review 120 image pairs. Each pair includes an original bird image and a version with a highlighted overlay indicating the model's focus area. The task is to choose the more accurate description of the highlighted region from two options. No special skills are required for participation. Simply make selections based on your observation. Notice: This study ensures the confidentiality of your participation, as it neither collects personally identifiable information nor contains any offensive content. Your feedback will exclusively be utilized for academic research purposes."

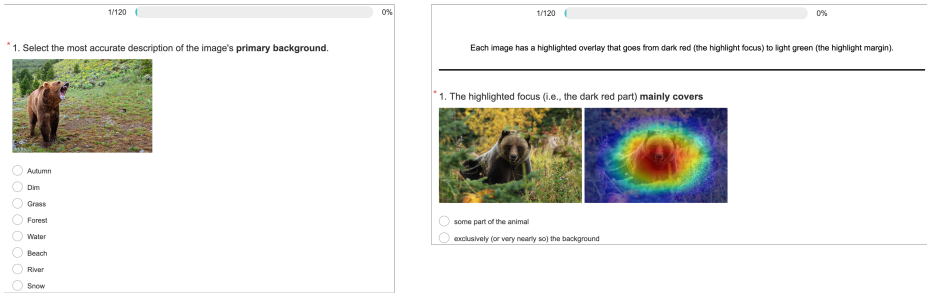
The interfaces for spuriousness labeling task are listed in Fig. A2.(a) and Fig. A3.(a). And the interfaces for attention correctness labeling task are listed in Fig. A2.(b) and Fig. A3.(b).



(a) (b)

Fig. A2: Crowdsourcing interface for (a) spuriousness labeling and (b) attention correctness annotation on Waterbirds dataset.

For the results provided in Secs. 5.2 and 5.4, we utilized a similar attention correctness labeling instruction and interface. The differences are as follows: (1)



(a) (b)

Fig. A3: Crowdsourcing interface for (a) spuriousness labeling and (b) attention correctness annotation on NICO dataset.

the instances selected for annotation are based on our proposed sampling strategy, as introduced in Sec. 4.2; (2) in the attention correctness labeling interface, option (a) is "some part of the {∗}," where {∗} represents the specific prediction corresponding to the image. In the case of ISIC dataset, we collaborated with domain experts to obtain annotation. For the other datasets, our participants were sourced from the Prolific platform.

A.7 AIoU

Previous research has often employed binary attribute maps to calculate the Intersection-over-Union (IoU) score against the ground-truth bounding box [9].

$$IoU(M, B) = \frac{\sum_{j,k} \min(M_{jk}, B_{jk})}{\sum_{j,k} \max(M_{jk}, B_{jk})}, \tag{A6}$$

However, the conventional IoU's reliability for assessing attribute map quality is compromised by its sensitivity to the chosen binarization threshold. To overcome this limitation, the revised approach [16] replaces the binary intersection with a minimum operator between a bounding box B_y and an explanation map M_y of ground truth calss y. AIoU employs a maximum operator in place of the binary union, facilitating a more consistent evaluation that is less susceptible to thresholding variations. Eqn.(A6) assesses the alignment between an explanation map and the ground-truth bounding box; however, it overlooks the possibility that, despite precise alignment for the correct class, explanation maps for alternative classes might overlap with the bounding box of the true class.

$$AIoU = \frac{IoU(M_y, B_y)}{IoU(M_y, B_y) + \max_{y\prime \in [C/y]} IoU(M_{y\prime}, B_y)}, \tag{A7}$$

Consequently, AIoU is a modified IoU metric that refines its denominator to account for the class with the explanation map exhibiting the maximum intersection with the ground-truth bounding box. In our evaluation, we use GradCAM as the explanation map.

Table A2: Ablation study on the dimensionality of attention space.

Annotation Amounts	`Dim` = 2	`Dim` = 3	`Dim` = 5	`Dim` = 10
60 (1.3%)	$78.21_{\pm 0.52}$	$78.43_{\pm 0.44}$	$78.91_{\pm 0.41}$	$79.73_{\pm 0.43}$
90 (1.9%)	$85.93_{\pm 0.92}$	$86.24_{\pm 0.96}$	$86.18_{\pm 0.91}$	$86.20_{\pm 0.94}$
120 (2.5%)	$89.12_{\pm 0.64}$	$89.12_{\pm 0.72}$	$89.13_{\pm 0.72}$	$89.12_{\pm 0.74}$

A.8 Additional Experiment Results

Table A3: Results with additional model architectures and datasets.

Method	Waterbirds (ViT)		MetaShift (ResNet50)		FMoW (DenseNet121)	
	Worst	Avg	Worst	Avg	Worst	Avg
ERM	$85.5_{\pm 1.2}$	$96.3_{\pm 0.5}$	$62.1_{\pm 4.8}$	$72.9_{\pm 1.4}$	$32.3_{\pm 1.3}$	$53.0_{\pm 0.6}$
JTT	$86.7_{\pm 1.5}$	$95.3_{\pm 0.7}$	$64.6_{\pm 2.3}$	$74.4_{\pm 0.6}$	$33.4_{\pm 0.9}$	$52.5_{\pm 0.3}$
DISC	$91.5_{\pm 1.3}$	$95.3_{\pm 1.1}$	$73.5_{\pm 1.4}$	$75.5_{\pm 1.1}$	$36.1_{\pm 1.8}$	$53.9_{\pm 0.4}$
SLIM$_{Tr}$	**92.1$_{\pm 0.6}$**	$96.4_{\pm 0.3}$	**75.7$_{\pm 1.0}$**	$76.4_{\pm 0.8}$	**37.4$_{\pm 1.1}$**	$54.1_{\pm 0.4}$
GDRO	$91.3_{\pm 0.8}$	$94.9_{\pm 0.3}$	$66.0_{\pm 3.8}$	$73.6_{\pm 2.1}$	$30.8_{\pm 0.8}$	$52.1_{\pm 0.5}$

Influence of the attention space's dimension. Table A2 shows *SLIM*$_{Tr}$'s ablation study results on Waterbirds, where for each annotation amount (N), we only vary the attention space's `dim` and measure `worst-group acc`. Results reveal a slight performance improvement with higher `dims` when N=60, but it is much less than the performance boosting caused by increasing N. When N=120 (same as Table 2 setting), we observe stable performance when varying `dims`. As 120 is a modest annotation amount, 2 `dim` is preferred.

Results with additional model architectures and datasets. Table A3 includes results: (1) on Waterbirds using a reference model matching ViT-S/16 in [5]; and (2) on MetaShift and FMoW, using reference models matching the corresponding ones in DISC [13]. Table A3 again confirms *SLIM*'s outstanding performance over the baselines.

Attention Consistency. In Sec. 5.3, we have quantitatively compared how similar neighbors are in the original feature space versus the attention space created by *SLIM*. In this section, we provide qualitative comparison by randomly selecting three points and examining the GradCAMs of their 10 nearest neighbors in the original representation space and our proposed attention space as showcased in Figs. A4 and A5. We can observe that, unlike the original space, the attention space aptly groups instances with coherent attributions. This facilitates the attention annotation and expansion with consistent attribution patterns.

Environment Feature Space. After disentangling core and environment features, we construct environment feature sets based on the inverse-attention-weighted features vectors $F_{\hat{A}}$. Here, we provide some intuitive examples to verify

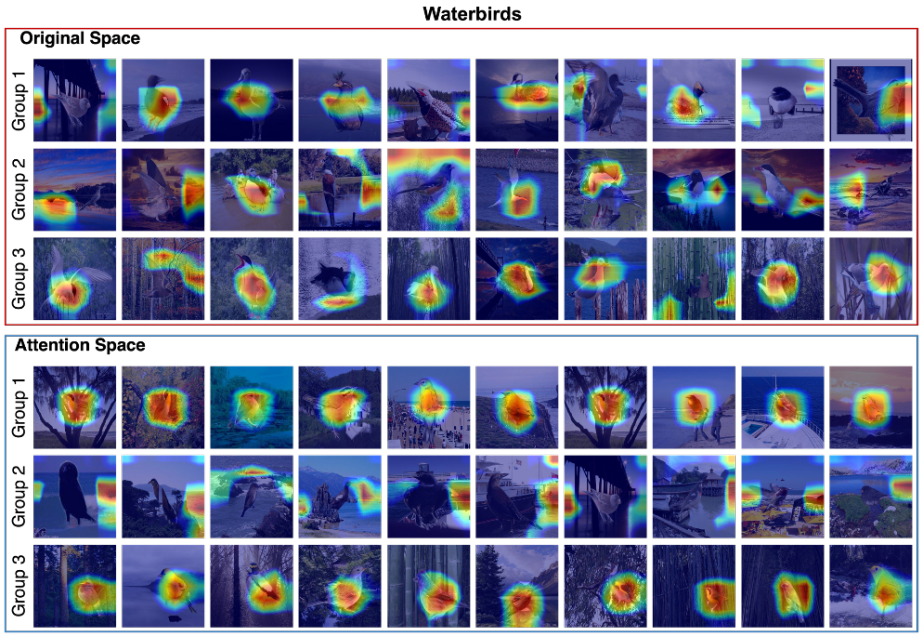


Fig. A4: Comparison of attention consistency between the original representation and attention spaces on the Waterbirds dataset. Examples in each group represent nearest neighbors within the corresponding space.

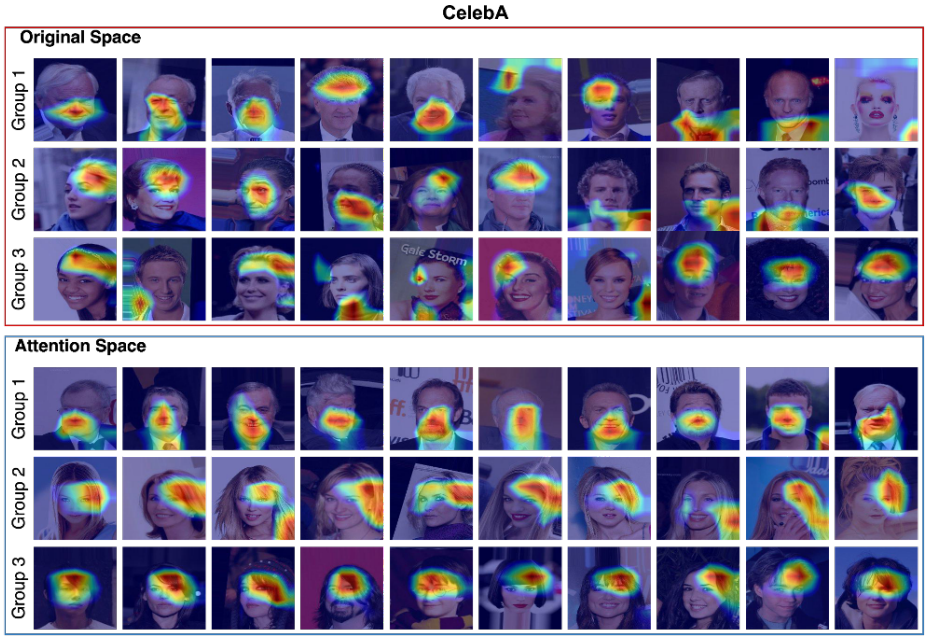


Fig. A5: Comparison of attention consistency between the original representation and attention spaces on the CelebA dataset. Examples in each group represent nearest neighbors within the corresponding space.

the consistency of the environment feature within clusters and the diversity of the environment feature between clusters in the environment feature space. We randomly select points from different clusters in environment feature space and examine the GradCAMs of their 10 nearest neighbors, the results as showcased in Figs. A6 and A7. As illustrated in Fig. A6, each group of data has a high consistency in environment features, such as land and sea backgrounds, ocean backgrounds, and forest backgrounds. This example demonstrates that we can effectively estimate the environment features by weighting F with inverse attention masks $\bar{A}$ (visualized as GradCAMs in Fig. A6) after identifying the core attention mask A. Furthermore, we find that compared to manually labeling spurious features, this proposed method allows us to identify different types of environment features more accurately and in greater detail. This paves the way for our ultimate goal: ensuring a balanced representation of core features across various environment features. In Fig. A7, we observe a similar situation: after isolating the core feature, namely hair, the first group exhibits a consistent environment feature, such as wearing glasses, while the second group consistently appears as white individuals.

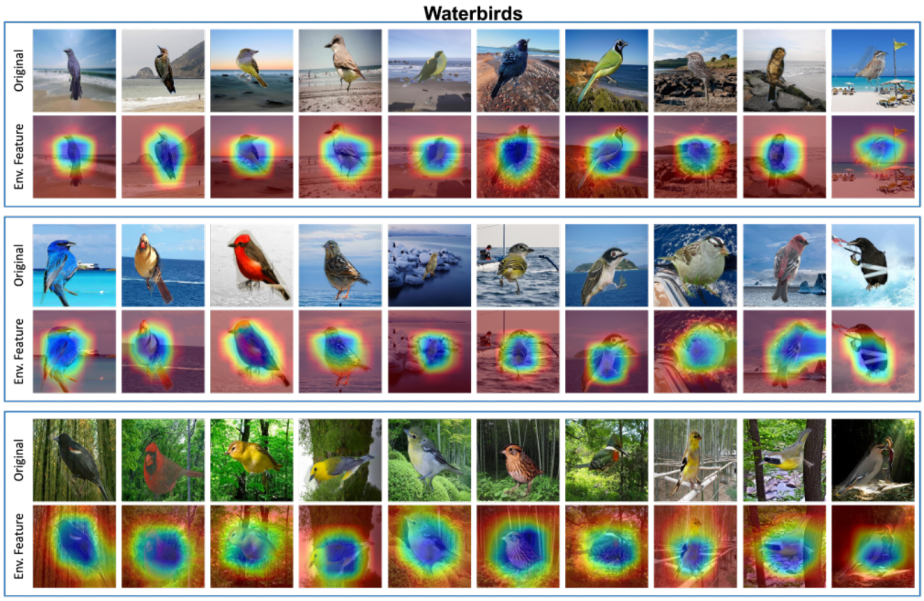


Fig. A6: Three groups of examples from the environment feature space on the Waterbird dataset.

Qualitative Evaluation of Enhanced Attention Accuracy. We provide more GradCAM examples showcase *SLIM*'s capability of correcting model's wrong attention in Fig. A8.

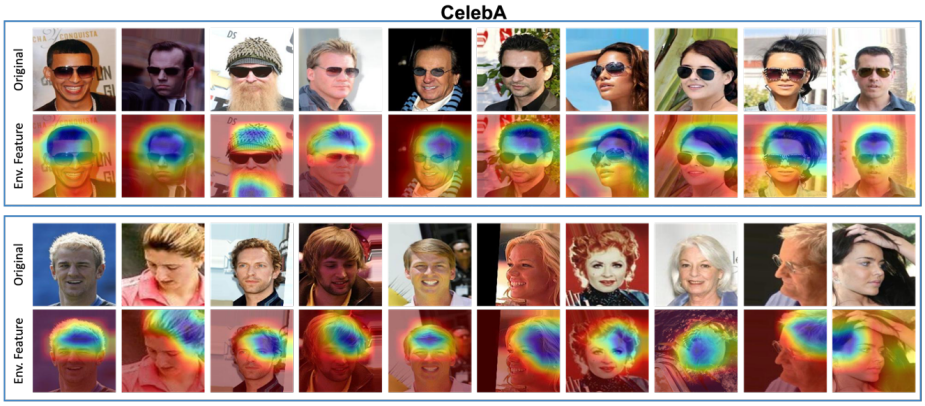


Fig. A7: Two groups of examples from the environment feature space on the CelebA dataset.

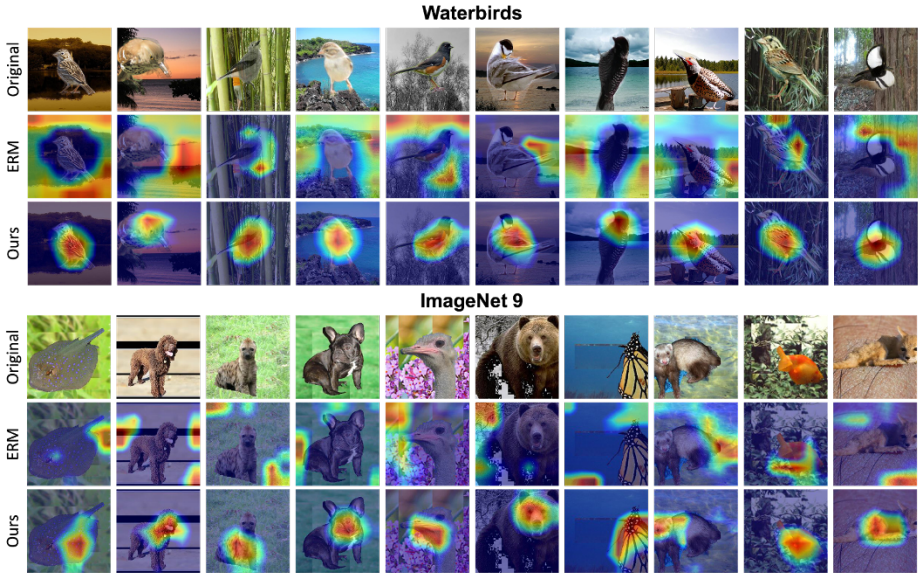


Fig. A8: GradCAM qualitative evaluation on Waterbirds and ImageNet-9. Dark red highlighted regions correspond to the attributions that are weighed more in the prediction. *SLIM* allows learning the core features instead of spuriousness.

A.9 Limitations

A limitation of our method is its reliance on attention-based spuriousness detection. Despite its effectiveness in handling spurious features that can be represented by a certain image region, it overlooks some types of spurious features, such as color or lighting. Such features are hard to be disentangled by attention-based model attributions. In the future, we plan to study how to mitigate other formats of spurious features.

References

1. Cao, Y., Chen, Z., Belkin, M., Gu, Q.: Benign overfitting in two-layer convolutional neural networks. Advances in neural information processing systems **35**, 25237–25250 (2022)
2. Chen, Z., Deng, Y., Wu, Y., Gu, Q., Li, Y.: Towards understanding the mixture-of-experts layer in deep learning. Advances in neural information processing systems **35**, 23049–23062 (2022)
3. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 168–172 (2018). https://doi.org/10.1109/ISBI.2018.8363547
4. Deng, Y., Yang, Y., Mirzasoleiman, B., Gu, Q.: Robust learning with progressive data expansion against spurious correlation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=9QEVJ9qm46
5. Ghosal, S.S., et al.: Are vision transformers robust to spurious correlations? IJCV **132**(3), 689–709 (2024)
6. Jelassi, S., Li, Y.: Towards understanding how momentum improves generalization in deep learning. In: International Conference on Machine Learning. pp. 9965–10040. PMLR (2022)
7. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last layer re-training is sufficient for robustness to spurious correlations. arXiv preprint arXiv:2204.02937 (2022)
8. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
9. Nguyen, G., Kim, D., Nguyen, A.: The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021), https://openreview.net/forum?id=0KPS9YdZ8Va
10. Rieger, L., Singh, C., Murdoch, W.J., Yu, B.: Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org (2020)
11. Sagawa*, S., Koh*, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=ryxGuJrFvS
12. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
13. Wu, S., Yuksekgonul, M., Zhang, L., Zou, J.: Discover and cure: Concept-aware mitigation of spurious correlation. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
14. Xiao, K., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition. ArXiv preprint arXiv:2006.09994 (2020)
15. Yang, Y., Gan, E., Karolina Dziugaite, G., Mirzasoleiman, B.: Identifying spurious biases early in training through the lens of simplicity bias. In: Proceedings of The 27th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 238, pp. 2953–2961. PMLR (02–04 May 2024)

16. Yang, Y., Nushi, B., Palangi, H., Mirzasoleiman, B.: Mitigating spurious correlations in multi-modal models during fine-tuning. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
17. Zhang, M., Sohoni, N.S., Zhang, H.R., Finn, C., Ré, C.: Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. arXiv preprint arXiv:2203.01517 (2022)
18. Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., Cui, P.: Nico++: Towards better benchmarking for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16036–16047 (2023)
19. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)