SELF-ENSEMBLE POST LEARNING FOR NOISY DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution shift and label noise remain major challenges to model robustness. In domain generalization (DG), noisy source labels can impair performance on unseen domains. We observe that intermediate representations retain discriminative information and attend to different image regions. Based on these observations, we propose **S**elf-**E**nsemble **P**ost **L**earning (SEPL) to improve the robustness of existing DG methods. SEPL consists of two phases: feature probing training and prediction ensemble inference. It trains multiple classifiers on intermediate features from a frozen pretrained model and uses semi-supervised learning to mitigate label noise. Their predictions are then combined through a crowdsourcing inference approach that accounts for differences in probe reliability. Extensive experiments across image classification benchmarks and medical-image settings show that SEPL improves robustness under label noise across DG methods.

1 INTRODUCTION

Distribution shift and label noise often coexist in practice (Figure 1(a)–(b)). Models must generalize beyond their training environments even when source labels are unreliable. In medical imaging, acquisition conditions and patient populations vary (Tschandl et al., 2018), and annotations can be imperfect (Karimi et al., 2020). Domain generalization (DG) and noisy-label learning have largely developed separately, with recent advances in domain adapters and debiased sample selection (Ji et al., 2025; Pan et al., 2026). Their intersection is now receiving direct attention through language-guided reweighting, adaptive proxy debiasing, and cross-domain noise detection (Dai et al., 2025; Wang et al., 2025b; 2026b). Strong DG performance with clean labels need not persist under label corruption (Qiao & Low, 2024; Sanyal et al., 2025): Figure 1(c) shows increasing noise fitting and falling OOD accuracy despite roughly stable ID accuracy.

Our starting point is the behavior of representations from a trained model. A layer-wise probe is a lightweight classifier trained on the features of a particular layer while the backbone remains frozen. In our PACS (Li et al., 2017) experiment, Grad-CAM (Selvaraju et al., 2017) maps of these probes emphasize different image regions (Figure 2). The probes also achieve different ID and OOD accuracies (Figures 3(a) and 3(c)). The deeper probes fit more corrupted labels in Figure 3(b), while averaging layer predictions improves OOD accuracy in Figure 3(c). These observations suggest that intermediate layers retain useful information beyond the final classifier. They motivate an ensemble

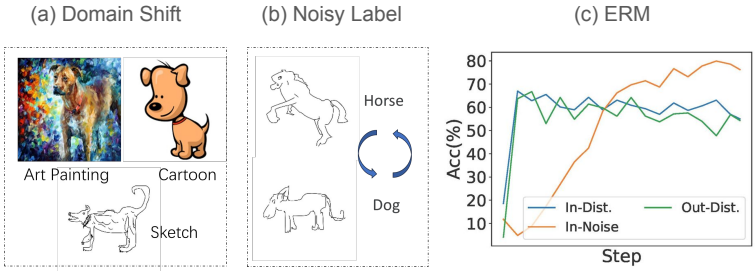


Figure 1: PACS (Li et al., 2017) examples: (a) domain shift; (b) ambiguous labels; (c) with 25% label flipping, ID accuracy stays roughly stable while noise fitting rises and OOD accuracy falls.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

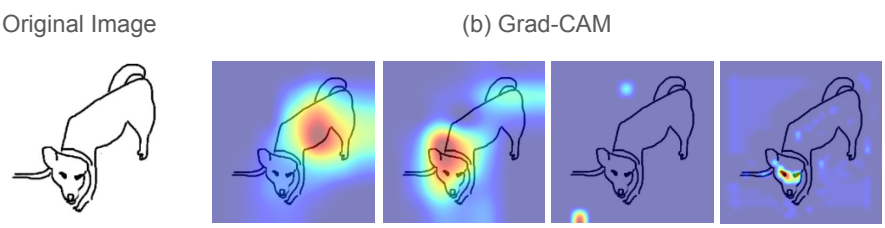


Figure 2: Different layers attend to different regions. From left to right: input, Grad-CAM of the original final-layer model, and added probes at Layers 4, 3, and 2.

across layers, but do not imply that every probe should receive equal weight: naive averaging falls below the backbone in the ablation in Table 3.

We propose **S**elf-**E**nsemble **P**ost **L**earning (SEPL), a post-training framework that freezes an existing backbone and learns classifiers on intermediate features. It consists of feature probing training and prediction ensemble inference. To reduce the effect of noisy labels on probe training, an entropy-based split treats lower-confidence samples as unlabeled and refines the probes with semi-supervised learning. At inference, layer-wise predictions are treated as imperfect annotators and combined using averaging or Dawid–Skene (DS) reliability estimation (Dawid & Skene, 1979).

Our contributions are: **(1) Layer-wise analysis.** We characterize discrimination, spatial attention, and noise fitting across intermediate representations. **(2) Post-training framework.** SEPL combines semi-supervised probe refinement with prediction aggregation on a frozen backbone; either component can be replaced independently. **(3) Empirical evaluation.** Experiments cover three image benchmarks and two medical-image settings, with ablations, architecture and noise stress tests, statistical comparisons, and deployment analyses. The medical experiments assess synthetic corruption on real images, rather than clinical performance.

2 Related Work

Domain generalization. DG methods improve transfer through representation learning and training strategies (Wang et al., 2023; Zhou et al., 2023). Recent examples include stronger ERM training (Teterwak et al., 2025) and domain-specific adapters (Ji et al., 2025). SEPL instead exploits intermediate features after backbone training.

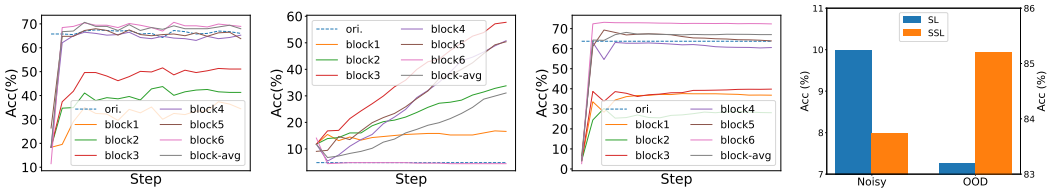
Learning with noisy labels. Robust losses and sample selection reduce the influence of incorrect labels (Song et al., 2023; Wang et al., 2025a; Pan et al., 2026). Such methods can support probe refinement, but label-noise robustness alone does not address domain shift.

Noisy domain generalization. Label noise can undermine DG (Qiao & Low, 2024; Sanyal et al., 2025). Recent methods use language anchors (A^3W), adaptive proxies (D2AP), and cross-domain noise detection (DL4ND) (Dai et al., 2025; Wang et al., 2025b; 2026b). These methods train representations under noise; SEPL refines and aggregates probes on a frozen backbone.

Aggregating intermediate predictions. Deep supervision and multi-exit networks train intermediate classifiers (Lee et al., 2015; Teerapittayanon et al., 2016; Huang et al., 2018), while learning from crowds estimates annotator reliability (Dawid & Skene, 1979; Rodrigues & Pereira, 2018; Guan et al., 2018). SEPL treats layer-wise probes as imperfect annotators and aggregates their predictions. App. A provides detailed related work.

3 Methodology

We consider a C-class classification problem with M source domains, $\mathcal{S} = \{\mathcal{S}^i\}_{i=1}^{M}$, where $\mathcal{S}^i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$. All domains share the input space $\mathcal{X}$ and label space $\mathcal{Y} = \{1, \ldots, C\}$, but their joint distributions differ and the observed source labels may be corrupted. Given a model trained on these sources, our goal is to improve its prediction on an unseen target domain without updating its backbone. Formally, we seek to minimize target risk $R_{\text{test}}(h) = \mathbb{E}_{(\mathbf{x},y)\sim P_{\text{test}}}[\ell(h(\mathbf{x}), y)]$, where



(a) ID accuracy. (b) Noise fitting. (c) OOD accuracy. (d) SL versus SSL.

Figure 3: PACS probes under label noise: ID accuracy and noise fitting increase with depth; aggregation improves OOD accuracy; semi-supervised training reduces noise fitting.

P_{test} is unknown during training. Source validation data guide model selection; target-test labels are never used for training or selection.

3.1 Motivation: what remains useful inside a noisy model?

We analyze ResNet-50 (He et al., 2016) on PACS with 25% source-label flipping. Linear probes are fitted to the input of Layer 1, the outputs of Layers 1–4, and the final pooled features. We select the base checkpoint by source validation with early stopping to limit noise fitting. The following observations motivate the probes, their aggregation, and their refinement.

Complementary spatial evidence. In the example in Figure 2, different probes emphasize the dog's body, head, tail, eyes, and surrounding regions. These maps illustrate differences in attention; they do not by themselves establish which regions cause a prediction. Table 4(a) further quantifies attention changes between models trained with clean and noisy labels. **Intermediate layers offer distinct spatial views**, motivating predictions from more than the final representation.

Discrimination and noise fitting vary with depth. Figures 3(a) and 3(b) show stronger ID discrimination and greater noise fitting in the deeper probes in this PACS experiment. Averaging their predictions improves OOD accuracy in Figure 3(c), but the separate ablation in Table 3 shows that averaging can also hurt. Thus, **layer diversity does not guarantee an ensemble gain**. We use reliability-aware aggregation to account for unequal error patterns; the comparisons below test whether it helps in each setting.

Probe training also needs noise control. Freezing the backbone does not prevent a new classifier from fitting corrupted source labels. In Figure 3(d), replacing supervised probe training with semi-supervised refinement reduces noise-fitting accuracy from 10% to 8% and raises OOD accuracy from 83.2% to 85.2%. **Semi-supervised refinement reduces noise fitting in the probes.** This motivates treating uncertain examples as unlabeled while retaining their inputs for learning.

3.2 Self-ensemble post learning

Frozen features and trainable probes. Let $h = g \circ f$ be the trained model, with feature extractor $f = f_T \circ \cdots \circ f_1$ and classifier g. We attach T probes g_t to selected representations:

$$\mathbf{z}_{i,t} = (f_t \circ \cdots \circ f_1)(\mathbf{x}_i), \qquad \tilde{\mathbf{y}}_{i,t} = g_t(\mathbf{z}_{i,t}), \quad t = 1, \ldots, T. \tag{1}$$

Here g_t includes softmax normalization, and $\tilde{\mathbf{y}}_{i,0} = h(\mathbf{x}_i)$ is the original prediction. The T probes and original model give $K = T + 1$ prediction sources; the default ResNet-50 uses $T = 6$ and $K = 7$. We warm up the probes on observed labels using cross-entropy, keeping f and g fixed.

Prediction aggregation. Given the K probability vectors, the ensemble prediction is

$$\hat{\mathbf{y}}_i = \mathcal{AGG}(\tilde{\mathbf{y}}_{i,0}, \ldots, \tilde{\mathbf{y}}_{i,T}), \qquad \mathcal{AGG}_{\text{AVG}} = \frac{1}{K} \sum_{t=0}^{T} \tilde{\mathbf{y}}_{i,t}. \tag{2}$$

Averaging is simple but ignores differences in probe reliability. We also use Dawid–Skene (DS) aggregation (Dawid & Skene, 1979). For an aggregation set of N examples, reindex the K sources by $j = 1, \ldots, K$ and let $q_{i,j,c'}$ be source j's probability for class c'. Its reliability matrix $\pi_j(c, c')$ describes prediction c' conditional on latent true class c, with each row summing to one. With the class prior fixed, SoftDS uses fractional annotations and HardDS uses one-hot predictions.

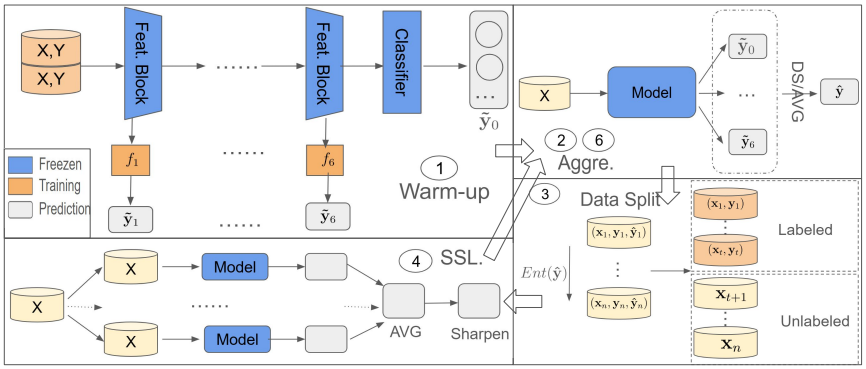


Figure 4: SEPL alternates prediction aggregation, entropy-based data splitting, and semi-supervised probe refinement while keeping the backbone and original classifier fixed.

Under the DS working assumption that annotations are conditionally independent, their class-conditional likelihood and the complete-data objective are

$$L_i(c;\pi) = \prod_{j=1}^{K}\prod_{c'=1}^{C} \pi_j(c,c')^{q_{i,j,c'}}, \qquad \mathcal{L} = \prod_{i=1}^{N}\prod_{c=1}^{C} \left[p_c L_i(c;\pi)\right]^{u_{i,c}}, \tag{3}$$

where $u_{i,c}$ is a latent one-hot class indicator and p_c is the class prior, uniform by default. We estimate reliability by maximizing the marginal log-likelihood

$$\ell(\pi) = \sum_{i=1}^{N} \log\left(\sum_{c=1}^{C} p_c L_i(c;\pi)\right). \tag{4}$$

EM (Expectation-Maximization) (Dawid & Skene, 1979) alternates label and reliability updates. The E-step needs a positive denominator; the M-step weights annotations by the posteriors:

$$\hat{y}_{i,c} = \frac{p_c L_i(c;\pi)}{\sum_{d=1}^{C} p_d L_i(d;\pi)}, \quad \pi_j(c,c') = \frac{\sum_{i=1}^{N} \hat{y}_{i,c} q_{i,j,c'}}{\sum_{i=1}^{N} \hat{y}_{i,c}}. \tag{5}$$

We initialize $\hat{\mathbf{y}}_i$ by averaging predictions and π_j uniformly, then perform the first M-step before updating the posteriors. The updates use at most 100 iterations and stop when the change in π is below 10^{-5} in infinity norm. A row with zero denominator is reset to uniform; no additional smoothing or clipping is used. Algorithm 2 in App. B gives the update order. Table 4 measures the residual dependence between probes; the working independence assumption should not be read as a claim that their errors are independent.

How reliability affects a prediction. For two candidate classes c and d, Equation 5 gives

$$\log \frac{\hat{y}_{i,c}}{\hat{y}_{i,d}} = \log \frac{p_c}{p_d} + \sum_{j=1}^{K}\sum_{c'=1}^{C} q_{i,j,c'} \log \frac{\pi_j(c,c')}{\pi_j(d,c')}. \tag{6}$$

This identity applies where the relevant probabilities are positive. A probe contributes class-specific evidence through its estimated confusion matrix, rather than a single global weight. In HardDS, only the predicted class contributes; SoftDS retains the full probability vector and is our default. The prior term also makes the role of class imbalance explicit. Correlated probes can contribute overlapping evidence, while a poorly estimated prior can favor the wrong class. Tables 4 and 6 examine these two concerns empirically.

Entropy splitting and semi-supervised refinement. The aggregate prediction also supplies a confidence estimate for training. We compute

$$H_i = -\sum_{c=1}^{C} \hat{y}_{i,c} \log \hat{y}_{i,c}, \qquad \mathcal{I}_\ell = \text{Lowest}_{\lfloor \gamma N \rfloor}\{H_i\}_{i=1}^{N}. \tag{7}$$

The operator selects the indices with lowest entropy. Their observed labels are retained in $\mathcal{D}_\ell = \{(\mathbf{x}_i, y_i) : i \in \mathcal{I}_\ell\}$; the remaining inputs form $\mathcal{D}_u = \{\mathbf{x}_i : i \notin \mathcal{I}_\ell\}$, with labels discarded. We fix

Algorithm 1 SEPL: probe training and prediction inference

Require: Noisy source data $\mathcal{D}$; trained backbone f and classifier g; T probes
1: Freeze f and g; initialize and warm up probes $g_1, \ldots, g_T$ on $\mathcal{D}$
2: **for** $r = 1, \ldots, 5$ **do**
3: Compute the backbone and probe probabilities using Equation (1)
4: Aggregate source predictions using AVG or the DS subroutine (Algorithm 2, App. B)
5: Rank examples by entropy and form $\mathcal{D}_\ell, \mathcal{D}_u$ using Equation (7)
6: Update only the probes with MixMatch on $\mathcal{D}_\ell, \mathcal{D}_u$
7: **end for**
8: **if** fixed-source DS inference is required **then**
9: Estimate reliability matrices and the class prior from labeled source data; keep them fixed
10: **end if**
11: Compute predictions for the test inputs with the frozen backbone and trained probes
12: Aggregate with AVG, target-batch DS (Algorithm 2, App. B), or fixed-source Equation (5)
13: **return** Predicted class $\arg\max_c \hat{y}_{i,c}$ for each test input

$\gamma = 0.5$; it is a confidence threshold by rank, not an estimate of the corruption rate, and retained labels need not all be correct.

We then refine the probes using MixMatch (Berthelot et al., 2019; Feng et al., 2023) on $\mathcal{D}_\ell$ and $\mathcal{D}_u$, with sharpening temperature $\tau = 0.5$. Figure 4 shows the sequence: supervised warm-up, aggregation, splitting, and semi-supervised refinement. The last three steps repeat for five rounds by default. Only probe parameters change, and the final probes are combined with the original classifier using Equation 2. SoftMatch (Chen et al., 2023) can also refine the probes.

Inference protocol. The default DS results use *target-batch transductive inference*: reliability matrices are estimated from the current unlabeled target predictions. For independent single-sample inference, we instead estimate the matrices and class prior from labeled source data and fix them before deployment. Neither mode uses target-test labels. We compare both protocols in Table 5 and examine aggregation under imbalance in Table 6. Those results motivate considering AVG, fixed DS, and adaptive DS separately rather than assuming that one rule is always preferable.

4 Experiments

4.1 Protocol and main benchmark results

We use DomainBed (Gulrajani & Lopez-Paz, 2021) with the noisy-label protocol of Qiao & Low (2024). Synthetic label corruption is applied before the $80/20$ source-domain training/validation split, so both subsets contain noisy labels, while target-domain labels remain unchanged. In each task, one domain is held out for testing and all remaining domains supply training data. Base models are selected using source-domain validation. Each task uses 20 base-method hyperparameter configurations, each evaluated in three independent experiments; SEPL is then applied with its reported default configuration. Setting-specific aggregation or probe-count choices use source-training characteristics and validation behavior only. Table 1 provides the configuration summary.

Table 1: Default reproducibility settings for SEPL. Base-model hyperparameters follow the selected corresponding baseline unless otherwise specified.

Component	Default setting
Base-model opt, lr, bs, and epochs	Follow the selected corresponding DomainBed baseline
Probe head	Linear classifier; frozen backbone; Adam optimizer
Dataset split for probing	$\gamma = 0.5$: lower-entropy half labeled, higher-entropy half unlabeled
MixMatch refinement	Temperature $\tau = 0.5$; supervised warm-up; five refinement rounds
DS annotators	Seven predictions: backbone output and six layer-wise probes
DS initialization	Uniform reliability matrices; AVG-initialized label posteriors
DS stopping criterion	Maximum 100 EM iterations; tolerance 10^{-5} for $\boldsymbol{\pi}$

Table 2: Average accuracy (%) with 25% source-label flipping. SEPL uses ERM++. †: results reported by Dai et al. (2025), not rerun here; –: not reported. Bottom rows count wins over ERM++ or all re-implemented baselines. Per-domain means and variation appear in App. C and D.

Method	PACS	VLCS	OfficeHome	HAM10000	Organ
ERM	76.7	77.4	60.7	65.9	69.3
Mixup	77.9	77.2	61.1	64.7	68.3
GroupDRO	77.0	77.3	61.4	64.3	68.7
IRM	77.1	76.4	57.6	64.6	68.8
VREx	77.5	77.6	59.8	65.3	67.3
CORAL	77.0	76.5	63.0	64.2	67.8
ADRMX	80.1	77.1	61.8	64.6	66.6
ERM++	82.4	77.6	66.3	67.2	72.8
$A^3W^\dagger$	82.1	76.1	65.2	–	–
SEPL	**83.7**	**78.7**	**67.5**	**69.5**	**75.5**
Wins vs. ERM++	4/4	3/4	4/4	4/4	3/3
Best vs. rerun baselines	4/4	1/4	4/4	4/4	3/3

Table 3: PACS accuracy (%): (a) ablations with ERM (a single AVG/DS denotes inference after supervised probing; two denote refinement/inference aggregation); (b) ensemble and noisy-label baselines, with SEPL applied to ERM++ and † denoting the reported A^3W result; (c) stress tests with one configuration and three runs. Per-domain results appear in App. C, Tables 11–13.

(a) Probe ablations

Probe / aggregation	Acc.
Base	76.7
Linear+AVG	75.8
Linear+DS	80.1
LSTM+AVG	77.0
LSTM+DS	78.8
AVG+AVG	75.7
AVG+DS	81.2
SEPL	81.0

(b) Method comparison

Method	Acc.
ERM	76.7
ERM Ensemble	78.2
DivideMix	78.2
PASS	80.6
SAGA	78.0
ERM++	82.4
$A^3W^\dagger$	82.1
SEPL	**83.7**

(c) Stress tests

Setting	Base	Before	+SEPL
Asym. noise	ERM	72.4	80.7
	ERM++	80.2	84.3
ViT-B/16	ERM	74.8	80.0
	ERM++	82.2	85.2

We re-implement eight baselines: ERM, Mixup, GroupDRO (Sagawa et al., 2020), IRM (Arjovsky et al., 2019), VREx (Krueger et al., 2021), CORAL, ADRMX, and ERM++ (Teterwak et al., 2025). PACS (Li et al., 2017) contains 9,991 images from seven classes across four style domains; VLCS (Fang et al., 2013) contains 10,729 images from five classes across four photographic domains; OfficeHome (Venkateswara et al., 2017) contains 15,588 images from 65 classes across four domains. The main protocol uses 25% random label flipping. VLCS uses HardDS at inference.

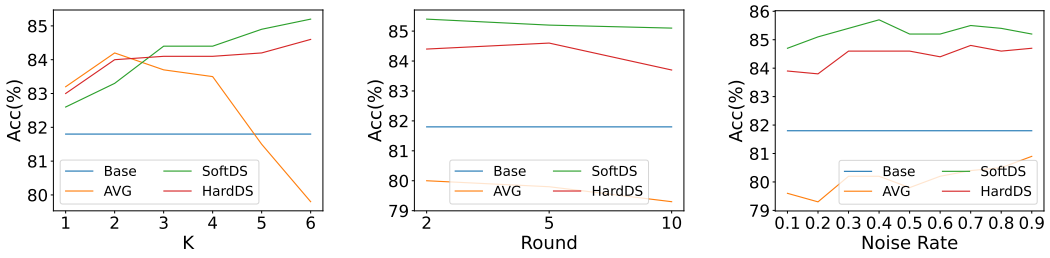
SEPL achieves the highest average accuracy among the compared methods in Table 2. On PACS, VLCS, and OfficeHome, it improves over ERM++ by 1.3, 1.1, and 1.2 percentage points, and over A^3W by 1.6, 2.6, and 2.3 points, respectively. It improves every PACS and OfficeHome domain over ERM++. On VLCS, it improves three of four domains over ERM++ and ranks first among re-implemented baselines on one.

Noise and architecture stress tests. Table 3(c) separates asymmetric noise from the ViT-B/16 architecture test (Dosovitskiy et al., 2021). With asymmetric noise, SEPL raises ERM from 72.4% to 80.7% and ERM++ from 80.2% to 84.3%. With ViT-B/16, the corresponding gains are 74.8% to 80.0% and 82.2% to 85.2%. Each test uses one sampled configuration and three runs. SEPL improves accuracy in both stress tests under this limited search budget.

4.2 Ablations and representation analysis

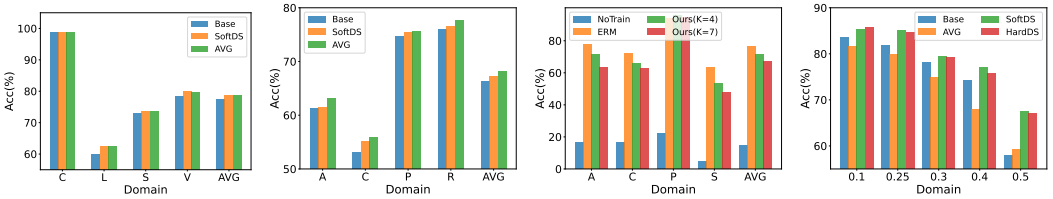
Comparison with noisy-label methods on PACS. SEPL achieves the highest PACS accuracy in Table 3(b), at 83.7%, followed by ERM++ at 82.4%. PASS (Garg et al., 2026) reaches 80.6%, ERM Ensemble and DivideMix (Li et al., 2020) reach 78.2%, and SAGA (Ma et al., 2024) reaches 78.0%.

Probe training and aggregation. The ERM ablation in Table 3(a) differs from the ERM++ comparison in Table 2. Linear+AVG reaches 75.8%, below the 76.7% backbone, while Linear+DS



(a) Number of annotators K. (b) Refinement rounds. (c) Split proportion γ.

Figure 5: Hyperparameter sensitivity. The default uses six probes plus the backbone prediction, five refinement rounds, and $\gamma = 0.5$; more probes or rounds are not necessarily better.



(a) VLCS. (b) OfficeHome. (c) No task fine-tuning. (d) Noise rate.

Figure 6: Changing aggregation, omitting backbone fine-tuning, and increasing label noise. Panels (a)–(b) vary aggregation, (c) starts from ImageNet features, and (d) varies the noise rate.

reaches 80.1%. Default SEPL reaches 81.0% after semi-supervised refinement. LSTM probes do not consistently improve on linear probes, and AVG refinement followed by DS inference reaches 81.2%. The stronger DS results support accounting for probe reliability; the LSTM comparison does not isolate the effect of feature quality. App. C, Table 11, gives results by domain.

Table 4: PACS diagnostics: (a) CAM statistics over 300 OOD comparisons; (b) the three evaluated probe pairs; (c) entropy-split mean $\pm$ std, with per-run values in Table 14 in App. C.

(a) Clean/noisy-label attention comparison Diagnostic	Estimate	95% CI
CAM mass outside clean top-20% region	40.68%	$[39.14, 42.22]\%$
Top-20% CAM-region IoU	0.6016	$[0.5877, 0.6155]$
(b) Layer-wise probe correlations **Probe pair**	**Soft-prediction correlation**	**Error correlation**
Probe 1 vs. Probe 6	0.3173	0.1013
Probe 1 vs. Probe 5	0.3087	0.1177
Probe 5 vs. Probe 6	0.9670	0.6782
(c) Entropy-based split quality **Diagnostic**	**Mean $\pm$ std**	**Reference**
Split precision (%)	40.39 ± 1.43	
Split recall (%)	81.92 ± 3.09	
Split ROC-AUC	0.8270 ± 0.0323	
Retained-subset noise (%)	8.92 ± 1.54	Initial noise: 24.65%

Layer-wise correlation and noise splitting. The corruption-mask evaluation in Table 4 shows that the high-entropy half recovers 81.92% of noisy samples; noise in the retained subset is 8.92%, compared with 24.65% before splitting. Table 4(b) reports both soft-prediction and error correlations for the three selected probe pairs. Probes 5 and 6 are more correlated than either distant pair involving Probe 1. These correlations measure dependence between predictions and errors, rather than between hidden features. In the CAM comparison, the noisy-label model places 40.68% of its mass outside the clean model's top-20% region and has region IoU 0.6016. Together, the probe and CAM measurements document predictive and spatial differences. Establishing their contribution to ensemble gains would require controlled interventions on these properties.

Across DG backbones. In Figure 7, SEPL improves every evaluated DG model on PACS, with the largest gain exceeding four percentage points. Models with higher baseline accuracy generally also reach higher accuracy after post-learning, although this comparison does not isolate representation quality from other differences between models. Using $K = 7$ prediction sources generally outperforms $K = 4$, with VREx an exception. These results span several DG objectives.

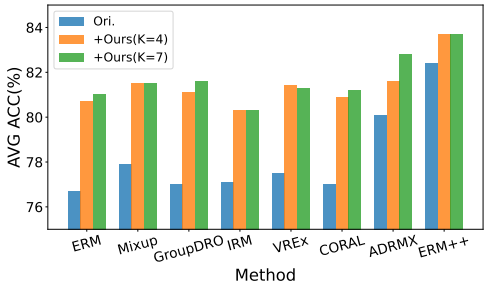


Figure 7: PACS accuracy before and after SEPL for each DG backbone.

Sensitivity and extensibility. Figure 5 examines sensitivity to annotator count, refinement rounds, and the split proportion. More probes or rounds are not uniformly better, and both DS variants outperform the plotted baseline over the tested ranges. We retain $\gamma = 0.5$ rather than tune it to the known synthetic noise rate. Replacing training-time SoftDS with AVG also changes the results on VLCS and OfficeHome (Figures 6(a)–6(b)); on OfficeHome, the gain over ERM++ is about 1.9 percentage points. With an ImageNet-pretrained backbone and no task fine-tuning, accuracy rises from 14.9% to 71.5% (Figure 6(c)), while more probes are not always better. At noise rate 0.5, SEPL exceeds the baseline by nearly ten points (Figure 6(d)). Gains persist throughout the tested noise range.

4.3 Deployment, Statistics, and Computational Cost

Table 5 compares inference protocols on PACS-A. Fixed HardDS reaches 85.2%, versus 84.9% for the backbone, whereas fixed SoftDS reaches 83.2%; adaptive HardDS reaches 86.3%. The results favor target-batch estimation in this setting, but it requires a batch of target predictions. On six independently trained ERM++ checkpoints, all paired gains are positive, with mean 1.27 ± 0.51 points and exact permutation $p = 0.03125$ (Table 5). The same table reports 43.9% lower peak memory during post-training and additional wall-clock cost. DS adds 1.491 ms/image to the median 0.864 ms/image backbone/probe path, for about 2.35 ms/image. Post-learning uses less peak memory than backbone training but adds computation.

Table 5: Inference and computation on PACS-A. Left: mean accuracy $\pm$ std; fixed methods process samples independently, adaptive methods use unlabeled target batches. Right: six-checkpoint paired evaluation and RTX 4090 costs. SEPL time is additional post-training cost.

Inference	Accuracy (%)
Backbone	84.9 ± 1.0
AVG	84.4 ± 1.0
Fixed SoftDS	83.2 ± 1.0
Fixed HardDS	85.2 ± 0.7
Adaptive SoftDS	86.2 ± 0.9
Adaptive HardDS	86.3 ± 1.0

Measurement	ERM++	SEPL
Accuracy (%)	85.20 ± 0.93	86.47 ± 0.86
Peak memory (MiB)	8,762	4,914
Time (s)	503.65	757.24

Paired gain: 1.27 ± 0.51 points; bootstrap 95% CI $[0.90, 1.63]$; exact permutation $p = 0.03125$. Inference: $0.864 + 1.491 \approx 2.35$ ms/image.

5 Illustrative Medical-Image Applications

Skin cancer. HAM10000 (Tschandl et al., 2018) provides 10,015 dermatoscopic images from seven categories. We remove missing-age records, form four age-based domains, merge rare categories into "other", and randomly sample 600 nv images per domain. The resulting task has 5,698 images and four classes: bkl, mel, nv, and others. Group0 remains imbalanced after preprocessing, so its reported accuracy uses AVG aggregation. SEPL improves mean accuracy over ERM++ by 2.3 points and ERM by 3.6 points (Table 2); App. D, Table 15, reports results by age group.

Organ classification. Organ{A,C,S}MNIST (Yang et al., 2021; 2023b) uses axial, coronal, and sagittal CT views as domains. We randomly sample 200 images per class per domain, giving 6,600

Table 6: HAM10000 imbalance analysis. (a) Group0 metrics (mean ± std). (b) Group0 and all-domain accuracy; the "Improved domains" column counts domains with gains over the corresponding uniform-prior variant. Only Group0 improves. App. D, Table 17, gives per-domain results.

(a) Group0: different objectives favor different aggregation rules

Method	Accuracy	Balanced accuracy	Macro F1
ERM++	69.1 ± 0.6	63.1 ± 1.0	61.4 ± 0.4
AVG	$\mathbf{71.2 \pm 0.5}$	59.6 ± 2.3	60.6 ± 1.3
SoftDS	66.0 ± 1.3	$\mathbf{68.7 \pm 0.6}$	62.2 ± 1.2
HardDS	69.6 ± 0.9	65.9 ± 0.9	$\mathbf{62.6 \pm 0.6}$

(b) Prior-aware aggregation: accuracy (%)

Method	Prior	Group0	Mean	Improved domains
SoftDS	Uniform	66.0 ± 1.3	68.1	–
SoftDS	Imbalance-aware	70.0 ± 0.6	68.2	1/4
HardDS	Uniform	69.6 ± 0.9	68.2	–
HardDS	Imbalance-aware	70.1 ± 0.6	68.0	1/4

images across 11 classes at 224×224 resolution. We use $K = 4$ for the sagittal target and $K = 7$ otherwise, based on source-training and validation behavior. Table 2 shows a 2.7-point gain over ERM++ and a 6.2-point gain over ERM; some DG baselines fall below ERM in the same table. Both medical tasks use 25% synthetic label flipping. Medical noisy-label studies also use controlled corruption (Khanal et al., 2023; Karimi et al., 2020; Xue et al., 2019), with different tasks and protocols. These tests assess synthetic corruption rather than clinical annotation errors.

6 Additional Analysis on Medical Data

On the imbalanced HAM10000 Group0, the aggregation rule with the best result depends on the metric: AVG leads in overall accuracy, SoftDS in balanced accuracy, and HardDS in macro F1 (Table 6). We use AVG for the reported Group0 accuracy. Prior-aware SoftDS and HardDS improve Group0 accuracy from 66.0% to 70.0% and from 69.6% to 70.1%, respectively. The improved-domain counts in Table 6(b) show that these gains are confined to Group0. Aggregation should therefore reflect minority-class performance and the reliability of available priors.

7 Limitations

SEPL has four main limitations. First, despite ViT-B/16 and asymmetric-noise stress tests, broader architectures, modalities, realistic noise processes, and naturally occurring clinical disagreement require further evaluation. Second, probes share a backbone, so DS independence is approximate; the prior comparison in Table 6 also shows sensitivity to class imbalance. Correlation modeling, robust prior estimation, regularized reliability weights, and sparse probe selection merit study. Third, choosing between adaptive target-batch inference and fixed single-sample inference involves a trade-off between accuracy and deployment constraints. Table 5 shows additional computation despite lower peak memory; early stopping, cached estimates, and deployment-specific aggregation may help. Finally, SEPL operates after backbone training. Joint representation and reliability learning, as in DL-CL (Rodrigues & Pereira, 2018), offers a possible extension.

8 Conclusion

SEPL improves noisy domain generalization by training intermediate probes on a frozen model and aggregating their predictions. Its entropy-based split supports semi-supervised refinement without assuming the label-noise rate. The benchmark results in Table 2, ablations in Table 3, and ImageNet-only experiment in Figure 6(c) show that intermediate representations can support useful post-training gains. The medical results provide further evidence under controlled label corruption, while the inference and imbalance analyses identify settings where aggregation choices matter. Future work should address probe dependence, prior estimation, and realistic annotation errors.

AI USE STATEMENT

Generative AI assisted with language editing, literature review, and comparisons of reported results. The authors reviewed the AI-assisted content and take responsibility for the manuscript.

ETHICS STATEMENT

The medical experiments use public datasets with synthetic label noise and do not establish clinical validity. Deployment requires validation with realistic annotation errors across patient groups.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify the method; Table 1 lists default settings. Apps. B–D provide implementation details, anonymous code and data links, and per-domain results.

REFERENCES

Paul Albert, Eric Arazo, Noel E O'Connor, and Kevin McGuinness. Embedding contrastive unsupervised features to cluster in- and out-of-distribution noise in corrupted image datasets. In *European Conference on Computer Vision*, pp. 402–419. Springer, 2022a.

Paul Albert, Diego Ortego, Eric Arazo, Noel E O'Connor, and Kevin McGuinness. Addressing out-of-distribution label noise in webly-labelled data. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 392–401, 2022b.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Aristotelis Ballas and Christos Diou. Gradient-guided annealing for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20558–20568, 2025.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. MixMatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

Wanxing Chang, Ye Shi, and Jingya Wang. Csot: Curriculum and structure-aware optimal transport for learning with noisy labels. *Advances in Neural Information Processing Systems*, 36:8528–8541, 2023.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. SoftMatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Zilin Dai, Lehong Wang, Fangzhou Lin, Yidong Wang, Zhigang Li, Kazunori D Yamada, Ziming Zhang, and Wang Lu. A language anchor-guided method for robust noisy domain generalization. *arXiv preprint arXiv:2503.17211*, 2025. URL https://arxiv.org/abs/2503.17211v2.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28 (1):20–28, 1979.

Berker Demirel, Erchan Aptoula, and Huseyin Ozkan. Adrmx: Additive disentanglement of domain features with remix loss. *arXiv preprint arXiv:2308.06624*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

Chuanwen Feng, Yilong Ren, and Xike Xie. Ot-filter: An optimal transport filter for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16164–16174, 2023.

Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, and Gustavo Carneiro. Pass: Peer-agreement based sample selection for training with instance dependent noisy labels. *Image and Vision Computing*, 166:105877, 2026. doi: 10.1016/j.imavis.2025.105877.

Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations*, 2018.

Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B Moeslund. A noisy elephant in the room: Is your out-of-distribution detector robust to label noise? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22626–22636, 2024.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8023–8031, 2023.

Yuyang Ji, Zeyi Huang, Haohan Wang, and Yong Jae Lee. Customizing domain adapters for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 934–944, 2025.

Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2020.

Bidur Khanal, SM Kamrul Hasan, Bishesh Khanal, and Cristian A Linte. Investigating the impact of class-dependent label noise in medical image classification. In *Proceedings of SPIE–the International Society for Optical Engineering*, volume 12464, pp. 1246437, 2023.

Noo-ri Kim, Jin-Seop Lee, and Jee-Hyong Lee. Learning with structural labels for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27610–27620, 2024.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pp. 562–570. Pmlr, 2015.

Gyuseong Lee, Wooseok Jang, Jinhyeon Kim, Jaewoo Jung, and Seungryong Kim. Domain generalization using large pretrained models with mixture-of-adapters. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 8248–8258, 2025.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Jichang Li, Guanbin Li, Hui Cheng, Zicheng Liao, and Yizhou Yu. Feddiv: Collaborative noise filtering for federated learning with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3118–3126, 2024a.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Yijiang Li, Sucheng Ren, Weipeng Deng, Yuzhi Xu, Ying Gao, Edith Ngai, and Haohan Wang. Beyond finite data: Towards data-free out-of-distribution generalization via extrapolation. *arXiv preprint arXiv:2403.05523*, 2024b.

Yi-Peng Liu, Dongxu Zeng, Zhanqing Li, Peng Chen, and Ronghua Liang. Ss-norm: Spectral-spatial normalization for single-domain generalization with application to retinal vessel segmentation. *IET Image Processing*, 17(7):2168–2181, 2023.

Yulin Ma, Jun Yang, and Ruqiang Yan. Sharpness-aware gradient alignment for domain generalization with noisy labels in intelligent fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 73, 2024. doi: 10.1109/TIM.2024.3413128.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Weiran Pan, Wei Wei, and Wenfeng Xie. Debiased sample selection for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32047–32057, 2026.

Kunyu Peng, Di Wen, Kailun Yang, Jia Fu, Yufan Chen, Ruiping Liu, Jiamin Wu, Junwei Zheng, M. Saquib Sarfraz, Luc Van Gool, Danda Pani Paudel, and Rainer Stiefelhagen. EReLiFM: Evidential reliability-aware residual flow meta-learning for open-set domain generalization under noisy labels. *arXiv preprint arXiv:2510.12687*, 2025. URL https://arxiv.org/abs/2510.12687v2.

Kunyu Peng, Di Wen, M. Saquib Sarfraz, Yufan Chen, Junwei Zheng, David Schneider, Kailun Yang, Jiamin Wu, Alina Roitberg, and Rainer Stiefelhagen. Mitigating label noise using prompt-based hyperbolic meta-learning in open-set domain generalization. *International Journal of Computer Vision*, 134:99, 2026a. doi: 10.1007/s11263-025-02643-9.

Yanming Peng, Shijing Wang, Yaping Huang, and Yi Tian. See through the noise: Improving domain generalization in gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31346–31355, 2026b.

Rui Qiao and Bryan Kian Hsiang Low. Understanding domain generalization: A noise robustness perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022a.

Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022b.

Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 28656–28679. PMLR, 2023.

Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Amartya Sanyal, Yaxi Hu, Yaodong Yu, Yian Ma, Yixin Wang, and Bernhard Schölkopf. Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2170–2178. PMLR, 2025. URL https://proceedings.mlr.press/v258/sanyal25a.html.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *European Conference on Computer Vision*, pp. 68–83, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123670069.pdf.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

Yi-Xuan Sun, Ya-Lin Zhang, Bin Han, Longfei Li, and Jun Zhou. Self-cognitive denoising in the presence of multiple noisy label sources. In *Forty-first International Conference on Machine Learning*, pp. 47261–47279, 2024.

Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pp. 2464–2469. IEEE, 2016.

Piotr Teterwak, Kuniaki Saito, Theodoros Tsiligkaridis, Kate Saenko, and Bryan A. Plummer. Erm++: An improved baseline for domain generalization. In *Winter Conference on Applications of Computer Vision*, 2025.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5: 180161, 2018. doi: 10.1038/sdata.2018.161.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Jialiang Wang, Xianming Liu, Xiong Zhou, Gangfeng Hu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Joint asymmetric loss for learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1947–1956, 2025a.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2023. doi: 10.1109/TKDE.2022.3178128.

Kening Wang, Di Wen, Yufan Chen, Ruiping Liu, Junwei Zheng, Jiale Wei, Kailun Yang, Rainer Stiefelhagen, and Kunyu Peng. Towards multi-source domain generalization for sleep staging with noisy labels. *arXiv preprint arXiv:2604.10009*, 2026a. URL https://arxiv.org/abs/2604.10009.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, 2018.

Siqi Wang, Aoming Liu, and Bryan Plummer. Noise-aware generalization: Robustness to in-domain noise and out-of-domain generalization. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6b90ec5be9adbaf8d6316da2211f745d-Abstract-Conference.html.

Yunyun Wang, Xiaodong Liu, and Yi Guo. D2AP: Double debiasing with adaptive proxies for domain generalization in noisy environment. *Knowledge-Based Systems*, 318:113458, 2025b. doi: 10.1016/j.knosys.2025.113458.

Cheng Xue, Qi Dou, Xueying Shi, Hao Chen, and Pheng-Ann Heng. Robust learning at noisy labeled medical images: Applied to skin lesion classification. In *2019 IEEE 16th International symposium on biomedical imaging (ISBI 2019)*, pp. 1280–1283. IEEE, 2019.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. Manydg: Many-domain generalization for healthcare applications. In *The Eleventh International Conference on Learning Representations*, 2023a.

Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023b.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A. Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. In *The Eleventh International Conference on Learning Representations*, 2023.

Xi Yu, Shinjae Yoo, and Yuewei Lin. Clipceil: Domain generalization through clip via channel refinement and image-text alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5372–5382, 2021.

Rui Zhao, Bin Shi, Jianfei Ruan, Tianze Pan, and Bo Dong. Estimating noisy class posterior with part-level labels for noisy label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22809–22819, 2024.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023. doi: 10.1109/TPAMI.2022.3195549.

A Additional Related Work

A.1 Domain generalization

DG differs from domain adaptation in that the target distribution is unavailable during model training (Pan & Yang, 2010; Wang & Deng, 2018; Wang et al., 2023; Liu et al., 2023). Common strategies include data manipulation, representation learning, and optimization. Mixup interpolates examples (Zhang et al., 2018); LISA restricts how examples are mixed (Yao et al., 2022); generative

approaches use large models to produce training variation (Li et al., 2024b). Representation methods include covariance alignment adapted from CORAL (Sun & Saenko, 2016; Gulrajani & Lopez-Paz, 2021), additive disentanglement with ADRMX (Demirel et al., 2023), dependence-based sample reweighting with StableNet (Zhang et al., 2021), patient-domain modeling with ManyDG (Yang et al., 2023a), and multimodal representations with CLIPCEIL (Yu et al., 2024).

Optimization and model-combination approaches target different parts of the training process. Fishr aligns gradient statistics (Rame et al., 2022a), DiWA averages independently trained weights (Rame et al., 2022b), and Model Ratatouille reuses auxiliary models before averaging fine-tuned weights (Ramé et al., 2023). ERM++ strengthens the training recipe (Teterwak et al., 2025); GGA searches for aligned source-domain gradients early in training (Ballas & Diou, 2025). MoA studies parameter-efficient adaptation and ensembles for large pretrained models (Lee et al., 2025). CDA customizes adapters to individual source domains (Ji et al., 2025). SEPL instead trains classifiers on fixed intermediate features after the backbone has been trained.

A.2 Learning with noisy labels

Noise-handling methods include architectural changes, regularization, robust losses, loss correction, and sample selection (Song et al., 2023). In the related setting of source-free domain adaptation, Yi et al. (2023) study noise in target pseudo-labels. SDM uses selective distillation (Sun et al., 2024); LSL augments label information with neighborhood structure (Kim et al., 2024); PLM combines part-specific information with transition modeling (Zhao et al., 2024). OT-Filter and CSOT use transport-based sample selection (Feng et al., 2023; Chang et al., 2023). PASS selects examples through agreement between two classifiers to train a third under instance-dependent noise (Garg et al., 2026). This differs from SEPL's entropy ranking of predictions from layers of one frozen model. JAL combines active and passive asymmetric losses (Wang et al., 2025a); Debiased Sample Selection addresses confirmation bias across classes and individual examples (Pan et al., 2026). These methods are relevant alternatives for probe refinement, but their reported noisy-label benchmarks do not establish performance under our DG protocol. In federated learning, FedDiv combines noise filtering with a consistency-based sampler (Li et al., 2024a).

A.3 Noisy domain generalization and related settings

Qiao & Low (2024) study DG through label-noise robustness, while Sanyal et al. (2025) examine how noise and nuisance features can disrupt the ID/OOD accuracy relationship. Noisy-label OOD detection (Humblot-Renaux et al., 2024) and the DrugOOD benchmark (Ji et al., 2023) study related but distinct problems. SAGA addresses noisy DG in fault diagnosis (Ma et al., 2024). A^3W aligns visual features with language anchors and weights samples by semantic agreement (Dai et al., 2025). D2AP first disentangles samples with class- and domain-level mixture models. Its adaptive proxies address domain imbalance among selected clean samples, while dual classifier heads separate pseudo-label generation from use to reduce confirmation bias (Wang et al., 2025b). DL4ND compares samples with class proxies from other domains, and evaluates both ID robustness and OOD generalization (Wang et al., 2026b). These methods modify training to address noisy DG, whereas SEPL operates on a frozen trained backbone. HyProMeta uses hyperbolic prototypes and prompt-based meta-learning for noisy open-set DG (Peng et al., 2026a). EReLiFM extends this setting with evidential clean/noisy partitioning and domain- and category-conditioned residual flow matching (Peng et al., 2025). Unlike the closed-set experiments here, both require rejecting unseen classes. SeeTN handles noisy continuous gaze annotations using feature–label affinity (Peng et al., 2026b), rather than aggregating categorical probes. FF-TRUST studies noisy multi-source sleep staging with time–frequency regularization on physiological signals (Wang et al., 2026a); its modality and sequential task differ from our image classification experiments. Earlier, DSON evaluated domain-specific normalization under noisy labels on PACS (Seo et al., 2020), providing another connection between DG and noise robustness. Web-label methods also distinguish ID and OOD noise through contrastive representations, clustering, or softened targets (Albert et al., 2022a;b).

A.4 Intermediate classifiers and learning from crowds

Deep supervision and multi-exit architectures use intermediate classifiers for joint learning or efficient prediction (Lee et al., 2015; Teerapittayanon et al., 2016; Huang et al., 2018). Crowd-

sourcing methods estimate labels from imperfect annotators, including DS, DL-CL, and Who Said What (Dawid & Skene, 1979; Rodrigues & Pereira, 2018; Guan et al., 2018). SEPL uses layer outputs as the annotations. Because these outputs share a backbone, the independence assumption is an approximation; Table 4(b) (main text) quantifies their dependence in our PACS setting.

B Implementation Details

B.1 DS aggregation

Algorithm 1 (main text) describes the full learning pipeline. Algorithm 2 operates only on a fixed set of predictions: it estimates reliability matrices and class posteriors without updating the backbone or probes. It is called for source-data aggregation during refinement and for adaptive target-batch inference. Fixed-source inference keeps source-estimated parameters constant and applies the E-step of Equation 5 (main text) once per target sample.

The marginal objective in Equation 4 (main text) is optimized by EM. With posterior weights $\hat{y}_{i,c}$ fixed, the reliability-dependent part of the EM auxiliary function is

$$Q(\pi) = \sum_{j=1}^{K} \sum_{c=1}^{C} \sum_{c'=1}^{C} \left(\sum_{i=1}^{N} \hat{y}_{i,c} q_{i,j,c'} \right) \log \pi_j(c, c'). \tag{8}$$

Maximizing each row subject to $\sum_{c'} \pi_j(c, c') = 1$ gives the M-step in Equation 5 (main text); Bayes normalization gives the E-step in the same equation. For SoftDS, this is a fractional-count likelihood objective, not a calibrated probability density over continuous softmax vectors. It does not imply that probe errors are conditionally independent. HardDS replaces each probability vector with its one-hot prediction; the update rules are otherwise unchanged.

Algorithm 2 DS aggregation subroutine for Algorithm 1 (main text)

Require: Probabilities $q_{i,j,c'}$, prior p_c, $I_{\max} = 100$, $\epsilon = 10^{-5}$
$\hat{\mathbf{y}}_i \leftarrow K^{-1} \sum_{j=1}^{K} \mathbf{q}_{i,j}$; $\pi_j(c, c') \leftarrow 1/C$
for $s = 1, \ldots, I_{\max}$ **do**
  Store $\pi^{\text{old}} \leftarrow \pi$
  Update π using Equation 5 (main text); reset zero-denominator rows to uniform
  Update $\hat{\mathbf{y}}$ using Equation 5 (main text)
  if $\|\pi - \pi^{\text{old}}\|_\infty < \epsilon$ **then**
    break
  end if
end for
return $\hat{\mathbf{y}}$

B.2 Configuration and resources

Table 1 (main text) lists the default settings. Base-model selection uses noisy source validation data; the held-out target labels are not used. The main protocol evaluates 20 base-method configurations in three independent runs. Stress tests use one sampled configuration in three runs. The sensitivity experiments characterize behavior rather than tune each dataset separately. The main runs use the reported default probe locations, refinement rounds, and split proportion, except for the explicitly stated aggregation and probe-count choices. Adaptive DS estimates reliability from unlabeled target batches; fixed DS holds source-estimated reliability constant during inference.

Data availability The datasets analyzed during the current study are available in DomainBed (Gulrajani & Lopez-Paz, 2021) (`https://github.com/facebookresearch/DomainBed`), HAM10000 (Tschandl et al., 2018) (`https://www.kaggle.com/datasets/farjanakabirsamanta/skin-cancer-dataset`), and MedMNIST (Yang et al., 2021; 2023b) (`https://zenodo.org/records/10519652`).

Code availability The implementation and experimental scripts are available at `https://anonymous.4open.science/r/SEPL-6B47/`.

C SUPPLEMENTARY EXPERIMENTAL RESULTS

C.1 COMPARISON WITH RECENT NOISY DG

Table 7 gives the A^3W values from Table II of Dai et al. (2025) (arXiv v2). Both studies use ImageNet-pretrained ResNet-50, leave-one-domain-out evaluation, and 25% symmetric label noise. A^3W uses CLIP text anchors during backbone training, whereas SEPL refines frozen ERM++ features and uses target-batch DS at inference. Table 2 (main text) includes these reported A^3W means; the comparison does not use paired reruns with matched splits, corruption masks, or checkpoints. The numerical differences therefore do not isolate the learning method or establish statistical significance. No A^3W medical results or per-domain win counts are inferred.

DL4ND targets Noise-Aware Generalization, evaluating both in-domain robustness and out-of-domain generalization under noisy supervision (Wang et al., 2026b). Our evaluation focuses on held-out-domain accuracy after post-learning on a frozen backbone. These evaluation goals and noise protocols differ, so we do not directly compare the reported scores. D2AP also targets noisy DG through adaptive proxy debiasing (Wang et al., 2025b); we have not verified results under a matching protocol. DSON evaluates domain-specific normalization with ResNet-18 at $20\%/50\%$ label noise (Seo et al., 2020); our experiments use ResNet-50 at 25% noise.

Other methods address different prediction tasks. HyProMeta and EReLiFM evaluate noisy open-set recognition with rejection of unseen classes (Peng et al., 2026a; 2025); SeeTN studies gaze regression with noisy continuous annotations (Peng et al., 2026b); FF-TRUST studies sleep staging from EEG/EOG signals (Wang et al., 2026a). These tasks differ from closed-set image classification. We did not reimplement DL4ND, D2AP, DSON, or these task-specific methods and do not include their scores as numerical baselines.

Table 7: Average accuracy (%) under 25% symmetric label noise. A^3W values (mean $\pm$ reported variation) are from Dai et al. (2025); SEPL values are from this work.

Method	PACS	VLCS	OfficeHome
A^3W (Dai et al., 2025)	82.1 ± 0.5	76.1 ± 0.3	65.2 ± 0.2
SEPL (this work)	83.7	78.7	67.5

C.2 PER-DOMAIN BENCHMARK RESULTS

Tables 8–10 give the domain-level means and variation underlying Table 2 (main text). PACS domains are Art painting (A), Cartoon (C), Photo (P), and Sketch (S). VLCS columns follow Caltech101 (C), LabelMe (L), SUN09 (S), and VOC2007 (V), consistent with Figure 6(a) (main text); SEPL uses HardDS on VLCS. OfficeHome domains are Art (A), Clipart (C), Product (P), and Real World (R). Each domain is held out in turn. On VLCS, ERM++ has the highest accuracy on Caltech101, while GroupDRO leads on SUN09 (Table 9).

Table 8: Results on PACS.

Domains	A	C	P	S	AVG
ERM	77.5±1.8	72.0±1.4	93.9±0.8	63.3±2.2	76.7
Mixup	81.0±1.1	71.8±0.4	94.9±0.7	64.1±0.7	77.9
GroupDRO	80.8±2.7	71.3±2.2	93.8±1.6	61.9±2.2	77
IRM	78.1±2.1	73.3±2.3	94.4±0.8	62.5±2.7	77.1
VREx	79.9±2.4	74.0±1.8	96.0±0.2	60.2±2.1	77.5
CORAL	79.0±1.5	71.0±1.8	95.8±1.4	62.2±2.0	77
ADRMX	83.4±0.9	75.7±0.9	96.3±0.3	65.1±1.8	80.1
ERM++	84.9±1.0	76.5±0.9	96.0±0.7	72.3±0.9	82.4
Ours	**86.2±0.9**	**78.3±0.1**	**96.7±0.5**	**73.5±1.9**	**83.7**

Table 9: Results on VLCS.

Domains	C	L	S	V	AVG
ERM	98.2±0.3	62.0±1.6	72.3±0.9	77.3±0.8	77.4
Mixup	97.3±0.9	62.3±1.1	71.9±0.4	77.4±0.5	77.2
GroupDRO	97.9±0.3	62.5±0.9	**74.0±1.2**	74.8±0.9	77.3
IRM	98.0±0.7	**63.8±0.4**	68.9±1.1	75.0±2.1	76.4
VREx	98.3±0.4	**63.8±0.9**	73.6±0.9	74.8±0.8	77.6
CORAL	97.4±1.3	63.7±1.0	71.5±1.3	73.5±0.9	76.5
ADRMX	98.5±0.2	**63.8±1.4**	71.5±0.9	74.6±1.4	77.1
ERM++	**99.0±0.2**	60.0±0.4	72.9±0.4	78.4±0.5	77.6
Ours	98.9±0.3	62.4±0.6	73.6±0.8	**79.9±0.8**	**78.7**

Table 10: Results on OfficeHome.

Domains	A	C	P	R	AVG
ERM	55.0±0.2	46.8±1.0	70.0±0.4	70.9±0.3	60.7
Mixup	55.5±1.4	48.1±0.3	69.6±0.3	71.3±0.3	61.1
GroupDRO	56.1±0.8	47.7±0.7	69.8±0.5	72.1±0.1	61.4
IRM	52.6±0.3	42.9±0.3	67.2±0.7	67.5±1.1	57.6
VREx	55.2±0.3	45.0±0.5	68.6±0.1	70.3±0.7	59.8
CORAL	58.9±1.0	48.7±1.1	71.1±0.2	73.2±0.6	63
ADRMX	57.7±0.4	46.8±0.8	70.0±0.1	72.6±0.4	61.8
ERM++	61.3±0.6	53.1±0.1	74.7±0.2	76.0±0.3	66.3
Ours	**62.0±0.2**	**55.2±0.4**	**75.7±0.3**	**77.0±0.3**	**67.5**

C.3 Ablations and stress-test details

Table 11 uses ERM, whereas Table 2 (main text) applies SEPL to ERM++. A single AVG/DS suffix denotes inference after supervised probing; two suffixes denote aggregation during refinement and inference, respectively. Tables 12 and 13 expand the PACS comparisons in Section 4 (main text). SAGA, PASS, and DivideMix are evaluated only on PACS in this study.

Table 11: PACS ablations of probe architecture, semi-supervised refinement, and aggregation.

Domains	A	C	P	S	AVG
Base	77.5±1.8	72.0±1.4	93.9±0.8	63.3±2.2	76.7
Linear+AVG	78.2±0.4	69.9±1.1	91.9±0.5	63.1±1.5	75.8
Linear+DS	79.7±1.5	76.0±1.0	95.0±0.6	69.8±1.1	80.1
LSTM+AVG	79.6±1.4	69.1±1.5	93.8±0.9	65.4±2.4	77
LSTM+DS	79.5±1.8	72.8±1.1	94.9±0.7	68.1±1.3	78.8
AVG+AVG	78.2±1.3	69.9±0.7	90.5±0.6	64.0±3.3	75.7
AVG+DS	82.0±1.5	77.0±0.2	95.1±0.9	70.7±1.1	81.2
Ours	81.8±1.5	76.7±0.1	95.4±0.6	70.4±1.0	81

Table 12: Additional PACS comparisons with ensemble and noisy-label baselines.

Method	A	C	P	S	AVG
ERM	77.5±1.8	72.0±1.4	93.9±0.8	63.3±2.2	76.7
ERM Ensemble	78.5±0.9	73.4±1.1	93.9±0.3	67.3±1.9	78.2
DivideMix	79.0±2.2	73.0±1.3	94.8±0.4	66.1±1.3	78.2
PASS	82.9±0.8	74.2±1.6	93.4±0.7	72.1±2.8	80.6
SAGA	76.2±1.7	73.9±1.8	92.2±0.8	69.7±0.7	78.0
ERM++	84.9±1.0	76.5±0.9	96.0±0.7	72.3±0.9	82.4
SEPL	**86.2±0.9**	**78.3±0.1**	**96.7±0.5**	**73.5±1.9**	**83.7**

Table 13: Stress tests on PACS with asymmetric noise and a ViT-B/16 backbone.

Setting	Method	A	C	P	S	AVG
Asymmetric noise	ERM	72.4±1.5	68.7±2.4	89.4±1.9	59.3±4.9	72.4
	ERM + SEPL	**81.9±0.4**	**76.8±1.2**	**94.6±0.4**	**69.7±3.0**	**80.7**
	ERM++	81.2±1.3	73.6±1.2	95.4±0.3	70.6±1.9	80.2
	ERM++ + SEPL	**87.1±0.5**	**79.2±0.8**	**97.4±0.2**	**73.7±1.4**	**84.3**
ViT-B/16	ERM	80.3±1.5	72.8±2.6	89.7±2.1	56.6±2.8	74.8
	ERM + SEPL	**83.9±1.7**	**77.2±1.9**	**97.3±0.4**	**61.5±0.6**	**80.0**
	ERM++	84.1±1.1	79.9±0.7	93.5±1.4	71.5±0.7	82.2
	ERM++ + SEPL	**88.0±1.3**	**80.1±0.8**	**98.3±0.3**	**74.2±1.0**	**85.2**

C.4 Attention, correlation, and entropy diagnostics

For the CAM analysis, we compare ERM models trained on clean labels and on 25% randomly flipped labels. Grad-CAM uses ResNet-50 Layer 4 and each model's predicted class. The clean model's top-20% CAM region is the reference. Three independently sampled OOD subsets of 100 images yield 300 comparisons. Figure 8 gives examples; Table 4(a) (main text) reports confidence intervals. The reference region is not a human-annotated causal explanation.

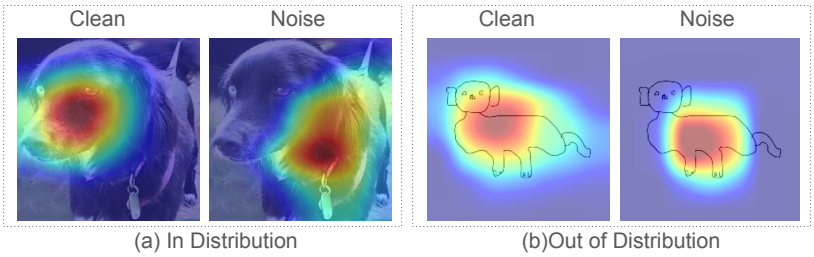


Figure 8: ResNet-50 Layer 4 Grad-CAM under clean and noisy training labels, on ID inputs (left) and OOD inputs (right), supplementing Table 4(a) (main text).

Table 14 reports all three entropy-split runs using the known synthetic corruption mask. Refinement discards labels in the high-entropy half. The corruption mask serves only to evaluate the split.

Table 14: Quality of the entropy-based clean/noisy split on PACS.

Run	Precision	Recall	ROC-AUC	Residual noise
1	41.63	84.70	0.8588	7.52
2	40.72	82.47	0.8280	8.65
3	38.83	78.60	0.7942	10.57
Mean±std	40.39±1.43	81.92±3.09	0.8270±0.0323	8.92±1.54

D Medical Data and Supplementary Results

D.1 HAM10000 preprocessing

The original HAM10000 distribution contains 6,705 nv images and 115 df images (Tschandl et al., 2018). We remove records without age, construct four age-based domains, combine the rare classes into "other", and sample 600 nv images per domain. The resulting bkl/mel/nv/other task has 5,698 images. This reduces the global imbalance but leaves domain-specific imbalance. Group0 uses AVG for the reported accuracy. We flip 25% of training labels synthetically; this is not a measurement of actual diagnostic disagreement. Table 15 provides the four-domain results.

Table 15: Results on the HAM10000 skin-cancer classification task.

Domains	Group0	Group1	Group2	Group3	AVG
ERM	67.8±0.7	73.7±1.4	61.5±1.9	60.5±1.1	65.9
Mixup	67.3±0.8	71.9±0.8	59.7±2.5	60.1±0.2	64.7
GroupDRO	66.8±0.5	73.7±0.1	59.5±0.5	57.0±0.6	64.3
IRM	66.8±1.4	73.8±0.3	63.5±1.9	54.6±2.0	64.6
VREx	68.5±1.4	73.3±0.7	62.4±0.5	57.0±2.0	65.3
CORAL	67.0±0.8	71.6±1.1	61.5±0.7	56.8±2.7	64.2
ADRMX	70.0±0.4	71.0±0.3	60.1±0.1	57.5±1.6	64.6
ERM++	69.1±0.6	75.0±0.7	65.0±0.8	59.8±1.4	67.2
Ours	**71.2±0.5**	**75.8±0.8**	**69.1±0.7**	**61.7±1.1**	**69.5**

D.2 OrganMNIST preprocessing

Organ{A,C,S}MNIST uses CT images from LiTS (Bilic et al., 2023) and is part of MedMNIST (Yang et al., 2021; 2023b). We use the 224×224 MedMNIST+ images, treating axial, coronal, and sagittal views as domains. The source collection has 11 classes and nearly 100,000 images; sampling 200 images per class per domain gives 6,600 images. We apply 25% synthetic label flipping. The reported choice is $K = 4$ for the sagittal target and $K = 7$ for the other targets, based on source-training and validation behavior rather than target-test labels. Table 16 reports accuracy for each of the three held-out views.

Table 16: Results on organ classification.

Dataset	A	C	S	AVG
ERM	75.3±1.5	79.8±0.5	52.8±1.1	69.3
Mixup	74.3±0.4	75.5±0.8	54.9±1.1	68.3
GroupDRO	75.9±1.3	76.6±0.5	53.7±1.1	68.7
IRM	74.8±1.0	77.5±1.7	54.1±1.2	68.8
VREx	72.3±3.4	75.7±1.0	53.9±1.0	67.3
CORAL	74.2±1.2	77.0±0.2	52.1±1.9	67.8
ADRMX	72.3±1.0	74.6±1.0	53.0±0.5	66.6
ERM++	80.5±0.5	79.9±0.4	57.9±0.4	72.8
Ours	**83.8±0.6**	**84.0±0.7**	**58.8±0.9**	**75.5**

D.3 Imbalance-aware metrics and priors

Table 17 reports all four age-domain accuracies underlying the prior comparison in Table 6(b) (main text). Prior-aware SoftDS and HardDS improve accuracy only on Group0.

Table 17: HAM10000 results with uniform and imbalance-aware class priors.

Assumption	Method	Group0	Group1	Group2	Group3	AVG
–	ERM	67.8±0.7	73.7±1.4	61.5±1.9	60.5±1.1	65.9
–	ERM++	69.1±0.6	75.0±0.7	65.0±0.8	59.8±1.4	67.2
Uniform	AVG	71.2±0.5	73.1±0.7	65.3±0.3	60.6±1.6	67.5
Uniform	SoftDS	66.0±1.3	75.8±0.8	69.1±0.7	61.7±1.1	68.1
Uniform	HardDS	69.6±0.9	75.2±0.9	67.1±0.5	60.9±1.4	68.2
Imbalance-aware	SoftDS	70.0±0.6	75.0±0.8	66.6±0.3	61.1±1.4	68.2
Imbalance-aware	HardDS	70.1±0.6	74.8±1.0	66.7±0.6	60.4±1.2	68.0