Towards Robust In-Domain and Out-of-Domain Generalization: Contrastive Learning with Prototype Alignment and Collaborative Attention

Yuan-Jhe Kuo
yuanjhe0817@gapp.nthu.edu.tw
Cheng-Yu Yang
andy855220_grad@gapp.nthu.edu.tw
Chiou-Ting Hsu
cthsu@cs.nthu.edu.tw

Department of Computer Science,
National Tsing Hua University,
Hsinchu, Taiwan

Abstract

Domain generalization focuses on generalizing a model learned from multiple source domains to the unseen target domain. Assuming the target domain has different distribution from the source domains, most methods addressed the out-of-domain generalization issue but slightly concern the in-domain performance on the source domains. Because the target domain is unseen and may distribute similarly with the source domains, we believe both the in-domain and out-of-domain performances are equally important. In addition, the noisy ground truth labels in the source domains also raises serious concerns on model robustness. Therefore, in this paper, we propose a contrastive learning framework with prototype alignment and collaborative attention to address the robust in-domain and out-of-domain generalization issue for image classification. We first design a margin-based contrastive learning to boost the out-of-domain performance by pushing the ambiguous classes apart by at least a margin. Next, we propose using prototype alignment to support the in-domain performance by aligning the latent feature representation of each class to the corresponding class prototype. Finally, we propose a novel collaborative attention method by leveraging the strength from both positive and negative learnings to enhance the model robustness. Experimental results on two benchmarks show that our method achieves competitive in-domain performance and outperforms previous methods in the out-of-domain and noisy label scenario.

1 Introduction

Although deep learning based methods [5, 9, 15] have achieved a great success in many computer vision tasks, these methods usually rely on i.i.d. assumption for data distributions and often have degraded performance when testing on out-of-domain data. This domain shift problem has been extensively studied in *domain generalization* (DG) through, e.g., domain alignment [8, 17, 20], data augmentation [23, 26], and regularization-based methods [2, 12].

Existing DG methods mostly focused on learning a model to well generalize to out-of-distribution data but hardly addressed the in-domain performance on the source domains. Because the target domain is unseen and that its data distribution is totally unpredictable, we believe the in-domain generalization is as important as out-of-domain generalization. Several methods [20, 23, 24] have indeed addressed the in-domain issue. In [23], a data augmentation-based method has been proposed to mix the domain distributions by linearly interpolating the training data and the labels. In [24], the authors proposed to boost in-domain performance by replacing some informative image regions with patches from other images. The methods [23, 24], though achieved good in-domain performance, did not reach good out-of-domain generalization. On the other hand, the method in [20], which addressed the out-of-domain generalization by minimizing the differences of feature distributions between multiple source domains, is not equally effective on in-domain generalization.

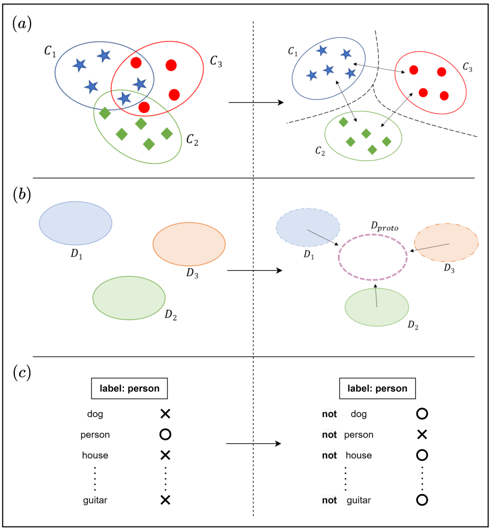


Figure 1: Illustration of the proposed idea: (a) margin-based contrastive learning for out-of-domain generalization, (b) prototype alignment for in-domain generalization, and (c) collaborative learning for improving model robustness.

In addition, noisy labels in source domains also raise a practical concern to domain generalization. In particular, the ground truth labels are usually collected by outsourcing services and are prone to human errors. These noisy "ground-truth" labels inevitably lead to performance drop and require special attention.

Therefore, in this paper, we consider the domain generalization scenario with noisy source labels and aim to simultaneously tackle the in-domain and out-of-domain generalization issues. This scenario is very challenging because there usually exists a trade-off between in-domain and out-of-domain performances. Our proposed method, as illustrated in Fig. 1, includes three major ideas. First, in Fig. 1 (a), we focus on improving out-of-domain performance by identifying the highly overlapped or ambiguous classes in the latent space and then pushing them apart by at least a margin. Second, in Fig. 1 (b), to promote in-domain performance, we explicitly align the source data of the same class but from different domains to the corresponding class prototype D_{proto} so as to learn the domain-agnostic and class-discriminative feature representation. Finally, in Fig. 1 (c), we take advantage of both positive and negative learnings to strengthen the model robustness against label noises. Here, "$\circ$" and "$\times$" indicate whether the corresponding class labels are involved in the model training or not. When including both the ground truth label and the complementary labels, we offer the model with informative supervision and readily improve the model robustness.

Our contributions are summarized as follows:

- We introduce a challenging scenario for achieving robust in-domain and out-of-domain generalization. To the best of our knowledge, this is the first work focusing on addressing these issues.

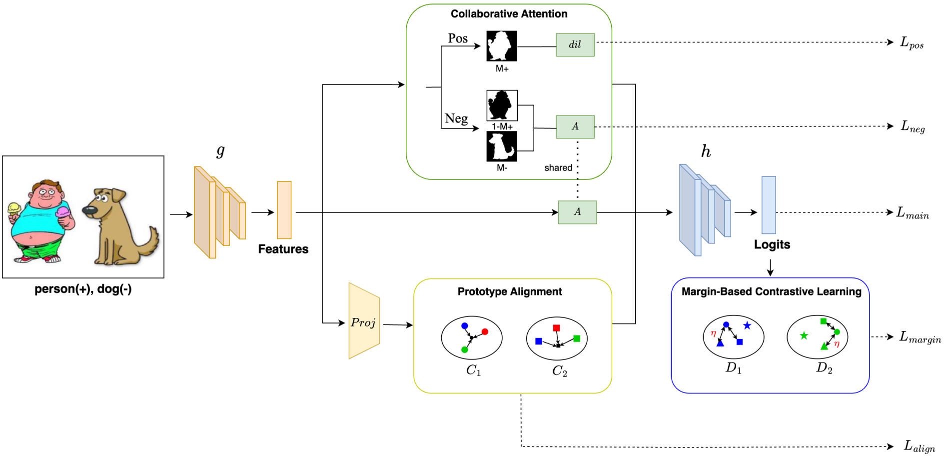


Figure 2: The proposed contrastive learning framework with prototype alignment and collaborative attention.

- The proposed contrastive learning framework together with prototype alignment and collaborative attention cooperatively fulfills the three main goals of this scenario.

- Experimental results show that the proposed method outperforms existing methods under three evaluation protocols on two benchmark datasets.

2 Proposed Method

In this paper, we address the robust domain generalization problem for image classification. Let $\mathcal{S} = \{\mathcal{D}_i\}$ denote the set of source domains, where the i-th domain $\mathcal{D}_i = \left\{(x_j, y_j)\right\}_{j=1}^{N_i}$ contains N_i annotated samples x_j with the ground truth label $y_j \in \{1, 2, ..., C\}$. Our goal is to train an image classification model f which should generalize to any unseen target domain $\mathcal{T}$ and perform well in the source domains $\mathcal{S}$. The model f should also be robust against label noises in the source domains $\mathcal{S}$. As shown in Fig. 2, we decompose the model f into one feature extractor g and one classifier h by $f = h \circ g$ and propose a contrastive learning framework with prototype alignment and collaborative attention to simultaneously achieve the three goals.

2.1 Margin-Based Contrastive Learning

Contrastive learning [18] has been adopted in domain generalization for classification tasks by minimizing the sample distance between intra-class pairs and maximizing the distance between inter-class pairs. However, because contrastive learning is conducted within one mini-batch, the performance and model stability highly depend on the quality of in-batch samples. In particular, when learning from multiple source domains, the intra-class pairs from different source domains may lead to unstable learning directions because of the domain shift. Thus, the inter-class pairs play a more important role in contrastive learning and should be carefully selected during the training stage.

In this paper, we design a margin-based contrastive learning in a per-sample manner and focus on identifying and separating those classes which frequently cause ambiguity to the classification model. To this end, for each input (x,y), we first find a set of ambiguous classes by selecting K classes that yield the highest prediction scores by $f(.)$. Next, we constrain all the samples with the same class label y as x to be distant from these top-K ambiguous classes by at least a margin η. Thus, we define the margin ranking loss L_{margin} as follows:

$$L_{margin} = \sum_{(x,y)} \sum_{k \in \mathcal{K}_y} max(\eta - |\theta(f(x),y) - \theta(f(x),k)|, 0) \quad , \tag{1}$$

where $\theta(f(x),y)$ is the y^{th} component of the C-dimensional prediction vector $f(x)$, and $\mathcal{K}_y$ is the set of top-K ambiguous classes of the class label y.

2.2 Prototype Alignment

Next, we address the in-domain generalization issue by prototype alignment. Here, our goal is to learn feature representation which possess both domain-agnostic and class-discriminative characteristics. To achieve these two objectives, we first define the class prototype m_c as the centroid of the corresponding class c in the latent space. Next, we obtain the latent representation of each class by a projection head $Proj(\cdot)$ and then align the projected class representation to the corresponding class prototype. With this class-wise alignment, we enforce the model to preserve the class-discriminative characteristics while aligning multiple source domains.

During the training stage, we update the class prototypes m_c by moving average with the in-batch class representation by,

$$m_c(t) = \frac{1}{N_c(t)}(N_c(t-1)m_c(t-1) + \sum_{\forall (x,y),\ y=c} Proj(g(x))) \quad , \tag{2}$$

where $Proj(\cdot)$ is the projection head consisting of a two-layer MLP, t is the training step, and N_c is the number of samples in the class c.

Finally, we define the prototype alignment loss L_{align} by,

$$L_{align} = \sum_{c=1}^{C} \sum_{\forall (x,y),\ y=c} \|Proj(g(x)) - m_c(t)\|_2 \quad . \tag{3}$$

2.3 Collaborative Attention

To enhance the model robustness against noisy labels, we propose a novel collaborative attention, in terms of positive learning and negative learning, to supervise the model learning. In classification task, positive learning is popularly used to train the model by minimizing the discrepancy between the prediction and the ground truth label. To resolve the noisy label issue, negative learning [14] has been proposed by minimizing the discrepancy between the complementary prediction and the negative label. Because complementary labels are less sensitive to label noises than the single ground-truth label, collaboration of positive learning and negative learning has been shown [14] to effectively improve the model robustness.

However, in the domain generalization scenario, the domain gap between multiple sources often diminishes the strength of both positive and negative learnings. In addition, by enforcing the model to fit to either the ground truth labels and/or the complementary labels, we risk compromising the domain generalization ability.

To resolve the above-mentioned problem, we propose to include a dilated positive attention and an extended negative attention to collaboratively supervise the model learning. We first identify the positive and negative attention maps of each class by the gradient responses. Given an input (x, y), we assign its ground truth y as the positive label y^+ and randomly select one complementary class as its negative label y^-; here, both y^+ and y^- are presented by one-hot vector. Let $z = g(x)$ be the extracted features of x. We obtain the positive and negative attention maps of x by,

$$M^+(\alpha) = \mathbf{1}_{grad^+ \geq \alpha} \quad , \tag{4}$$

$$M^-(\beta) = \mathbf{1}_{grad^- \geq \beta} \quad , \tag{5}$$

where

$$grad^+ = \frac{\partial h(z) \cdot y^+}{\partial z} \quad , \tag{6}$$

$$grad^- = \frac{\partial h(z) \cdot y^-}{\partial z} \quad . \tag{7}$$

In Equations (4) and (5), $\mathbf{1}$ denotes the indicator function, α and β are the p^{th} percentiles of $grad^+$ and $grad^-$, respectively.

Then, to augment the representation capacity of each class, we propose using dilated positive attention by enlarging the attention map M^+ using a dilation module. The dilation module includes spatial and channel dilations using a fixed dilation kernel size. Then we average the two dilated outputs to derive the dilated attention. In our experiments, we adopt the convolution blocks with 2D-Maxpooling and 1D-Maxpooling with kernel size 3 for spatial and channel dilations, respectively, to construct the dilation module. Then, in terms of the dilated positive attention $dil(M^+)$, we define the positive loss L_{pos} by,

$$L_{pos} = L_{ce}(h(z^+); y^+) \quad , \tag{8}$$

where $L_{ce}(\cdot)$ is the cross-entropy loss, $z^+ = z \odot dil(M^+)$ is the feature map masked by the dilated attention $dil(M^+)$, and $\odot$ is the element-wise product.

As to the negative learning, we propose an extended negative attention by combining both the negative attention M^- and non-positive attention $(1 - M^+)$ as the complementary prediction. The negative attention M^- supports the capability of negative learning for the randomly selected class y^-; and the non-positive attention $(1 - M^+)$ excludes the positive class y^+ from the negative learning and further enhances the capability of negative learning. We define the negative loss L_{neg} by,

$$L_{neg} = L_{ce}(h(\mathcal{A}(z^-)); y^-) \quad , \tag{9}$$

where $\mathcal{A}$ is a self-attention module implemented by CBAM [22], and

$$z^- = z \odot (\frac{1}{2}M^- + \frac{1}{2}(1 - M^+)) \tag{10}$$

is the feature map masked by the negative and non-positive attentions. Then we define the collaborative loss by

$$L_{collab} = L_{pos} + \lambda^{-} L_{neg} \quad , \tag{11}$$

where λ^{-} is a hyper-parameter and is set to be 0.2 in our experiments.

2.4 Total Loss

Finally, we include the image classification loss for the in-domain data (x, y) by,

$$L_{main} = L_{ce}(h(\mathcal{A}(z)); y) \quad . \tag{12}$$

where $L_{ce}(\cdot)$ is the cross-entropy loss, and $\mathcal{A}(\cdot)$ is the self-attention module. Here $\mathcal{A}(\cdot)$ is included to maintain the in-domain performance.

By combining the classification loss L_{main}, the margin ranking loss L_{margin}, the prototype alignment loss L_{align}, and the collaborative loss L_{collab}, we define the total loss L_{all} by,

$$L_{all} = L_{main} + \lambda_1 L_{margin} + \lambda_2 L_{align} + \lambda_3 L_{collab} \quad , \tag{13}$$

where λ_i are the hyper-parameters.

3 Experiments

3.1 Datasets and Evaluation Metrics

We conduct experiments on two image classification benchmarks PACS and VLCS. PACS [16] contains 4 domains, 7 classes, and 9991 examples; and VLCS [6] includes 4 domains, 5 classes, and 10729 examples. We evaluate the model performance in terms of two metrics: in-domain accuracy (ID) and out-of-domain accuracy (OOD) under two noisy-label protocols *symm inc* and *symm exc*. To evaluate ID, we follow [3] to split half of the validation sets as test sets and measure the averaged accuracy on the test sets. To evaluate OOD, we conduct the leave-one-domain-out protocol with model selection by training-domain validation set [10] and report the averaged accuracy on the test domains. To simulate the label noises, we follow [14] and use the symmetric noise protocols *symm inc* and *symm exc* with ratio 0.2 and 0.4 to perturb the labels.

3.2 Implementation Details

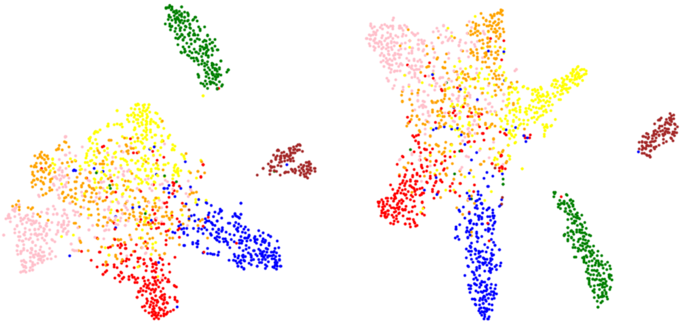
We adopt ResNet-18 [11] and ResNet-50 [11] as backbones in our experiments. The hyper-parameters λ_1, λ_2, and λ_3 in Equation (13) are set as 0.01, 0.05, and 0.05, respectively. To have a fair comparison, we follow the recent domain generalization test-bench [10] and set the batch-size to be 8 for each domain in all of our experiments. Because the recent ensemble-based techniques [1, 3, 13] effectively improve the domain generalization ability, we evaluate our method with SWAD [3] and also report the results without SWAD for comparison. All the experiments are performed with 3 trials of different random seeds and the averaged results are reported.

Components				ResNet-50	
L_{main}	L_{collab}	L_{margin}	L_{align}	ID	OOD
✓				97.89 ± 0.18	86.90 ± 0.30
✓	✓			97.36 ± 0.86	86.97 ± 0.79
✓		✓		97.60 ± 0.17	87.13 ± 0.36
✓			✓	97.57 ± 0.24	86.84 ± 0.14
✓		✓	✓	97.68 ± 0.32	87.45 ± 0.12
✓	✓	✓		97.56 ± 0.33	87.66 ± 0.37
✓	✓	✓	✓	97.80 ± 0.54	87.68 ± 0.48

Table 1: Ablation study of our modules on PACS for in-domain (ID) and out-of-domain (OOD) metrics with SWAD.

L_{collab}	In-Domain			Out-of-Domain		
	0	0.4	drop	0	0.4	drop
	97.68 ± 0.32	95.67 ± 0.70	-2.01	87.45 ± 0.12	83.32 ± 1.44	-4.13
✓	97.80 ± 0.54	96.04 ± 0.59	-1.76	87.68 ± 0.48	84.55 ± 0.45	-3.13

Table 2: Ablation study of collaborative attention on PACS for metrics against *symm exc* label noises with SWAD.



(a) (b)

Figure 3: t-SNE visualization for out-of-domain data on PACS with ResNet-18. (a) ERM [21]; (b) Our contrastive learning framework.

3.3 Ablation Study

In Table 1 and Table 2, we verify the effectiveness of each component in the proposed method on PACS using ResNet-50 backbone with SWAD.

Effectiveness of Collaborative Attention. As shown in Table 1, when including the negative loss in the collaborative attention, we improve the averaged out-of-domain performance by 0.07%. In addition, when further combining with the other two modules, we boost the averaged out-of-domain performance by 0.23%. These results show that our collaborative attention fully utilizes the information from positive and negative classes to improve the model generalizability. In Table 2, when including the collaborative attention module in our framework, we effectively avoid the performance drop on both in-domain and out-of-domain evaluation metrics. These results validate the effectiveness of the proposed collaborative attention.

Effectiveness of Margin-Based Contrastive Learning. In Table 1, when including the margin-based contrastive learning, we improve the averaged performance by 0.23%. When the collaborative attention is included, we improve the average performance from 86.97% to 87.66% and have +0.69% improvement for out-of-domain performance with ResNet-50. These results verify that the proposed margin-based contrastive learning effectively separates the ambiguous classes.

Effectiveness of Prototype Alignment. As shown in Table 1, when the prototype alignment

Method	PACS	VLCS	Avg.
ERM [21]	97.75 ± 0.41	87.21 ± 0.72	92.48
CORAL [20]	97.64 ± 0.33	86.88 ± 0.87	92.26
RSC [12]	97.01 ± 0.58	86.48 ± 0.54	91.75
SagNet [19]	97.53 ± 0.40	86.86 ± 0.83	92.20
Mixup [23]	**97.92 ± 0.54**	86.89 ± 0.93	92.41
Mixstyle [26]	97.31 ± 0.68	86.89 ± 0.92	92.10
ARM [25]	97.86 ± 0.45	87.08 ± 0.95	92.47
SAM [7]	97.84 ± 0.27	86.20 ± 0.55	92.02
MIRO [4]	97.74 ± 0.11	**87.57 ± 1.00**	**92.66**
Ours	97.80 ± 0.54	87.43 ± 0.82	92.62

Table 3: Comparison for in-domain performance with SWAD using ResNet-50.

Method	PACS	VLCS	Avg.
ERM [21]	83.43 ± 0.67	76.52 ± 0.64	79.98
CORAL [20]	83.13 ± 0.74	76.68 ± 0.61	79.90
RSC [12]	81.98 ± 1.05	75.61 ± 0.79	78.79
SagNet [19]	81.40 ± 0.27	76.20 ± 1.02	78.80
Mixup [23]	81.76 ± 1.49	76.83 ± 1.55	79.30
Mixstyle [26]	82.92 ± 0.36	76.04 ± 1.37	79.48
ARM [25]	83.55 ± 1.27	75.18 ± 1.14	79.36
SAM [7]	83.93 ± 1.65	77.00 ± 1.60	80.46
MIRO [4]	**84.14 ± 0.32**	77.96 ± 0.94	81.05
Ours	83.87 ± 0.57	**78.59 ± 0.70**	**81.23**

Table 4: Comparison for out-of-domain performance without SWAD using ResNet-50.

Method	PACS	VLCS	Avg.
ERM [21]	87.56 ± 0.33	78.13 ± 0.16	82.85
CORAL [20]	87.40 ± 0.19	78.20 ± 0.26	82.80
RSC [12]	84.04 ± 0.67	77.99 ± 0.27	81.02
SagNet [19]	86.52 ± 0.63	77.82 ± 0.22	82.17
Mixup [23]	85.90 ± 0.08	78.61 ± 0.14	82.26
Mixstyle [26]	86.15 ± 0.41	78.00 ± 0.39	82.08
ARM [25]	87.31 ± 0.16	78.10 ± 0.25	72.71
SAM [7]	86.28 ± 0.37	78.19 ± 0.26	82.23
EoA [1]	87.55	78.86	83.21
MIRO [4]	87.57 ± 0.21	79.08 ± 0.35	83.33
Ours	**87.68 ± 0.48**	**79.27 ± 0.26**	**83.48**

Table 5: Comparison for out-of-domain performance with SWAD using ResNet-50.

is applied along with other modules, we boost the in-domain performance from 97.56% to 97.80% (+0.24%) and increase the out-of-domain performance by 0.02% with ResNet-50. Although the out-of-domain performance only slightly improves, the improvement on in-domain performance shows that the proposed prototype alignment indeed enables the model to learn domain-agnostic and class-discriminative characteristics by aligning the class representation of different domains.

Visualization. Fig. 3 shows the t-SNE visualization on PACS dataset and compares with ERM [21]. Here, we adopt the leave-one-domain-out protocol and show the results of testing on the target domain *cartoon* by using different colors to indicate different classes. The visualization results show that our method produces well-separated class-wise clusters in the out-of-domain setting and validate the effectiveness of the proposed margin-based contrastive learning framework.

3.4 Comparison

In-Domain Performance. Table 3 shows the in-domain performance on the two benchmarks and compares with other methods which also adopted the same network backbone (ResNet-50) as ours. The results on in-domain testing show that the proposed method is competitive with the state-of-the-art method [4] and verify the effectiveness of our method on maintaining good in-domain performance.

Out-of-Domain Performance. In Table 4, we show the out-of-domain performance and compare with other methods. The proposed method improves ERM [21] by +0.44% in

Method	*symm inc*						*symm exc*					
	In-Domain			Out-of-Domain			In-Domain			Out-of-Domain		
	0	0.2	0.4	0	0.2	0.4	0	0.2	0.4	0	0.2	0.4
ERM []	97.75 ± 0.41	97.35 ± 0.50	96.15 ± 0.51	87.56 ± 0.33	$\mathbf{86.34 \pm 0.19}$	84.72 ± 0.81	97.75 ± 0.41	97.16 ± 0.44	95.76 ± 0.64	87.56 ± 0.33	86.20 ± 0.38	83.94 ± 0.81
RSC []	97.01 ± 0.58	96.40 ± 0.82	95.11 ± 0.76	84.04 ± 0.67	82.91 ± 0.92	78.62 ± 1.04	97.01 ± 0.58	96.34 ± 0.54	94.32 ± 1.16	84.04 ± 0.67	82.29 ± 0.93	76.80 ± 2.88
Mixup []	97.92 ± 0.54	97.23 ± 0.46	96.37 ± 0.44	85.90 ± 0.08	85.36 ± 0.43	84.13 ± 0.52	97.92 ± 0.54	96.87 ± 0.56	95.75 ± 0.40	85.90 ± 0.08	84.96 ± 0.25	83.24 ± 0.46
SagNet []	97.53 ± 0.40	97.05 ± 0.66	96.40 ± 0.82	86.52 ± 0.63	85.50 ± 0.31	83.60 ± 0.51	97.53 ± 0.40	96.84 ± 0.64	95.86 ± 0.86	86.52 ± 0.63	85.94 ± 0.20	82.85 ± 0.49
CutMix []	97.77 ± 0.16	97.32 ± 0.20	96.15 ± 0.59	85.31 ± 0.26	84.58 ± 0.63	82.50 ± 0.53	97.77 ± 0.16	97.08 ± 0.38	95.78 ± 0.62	85.31 ± 0.26	84.29 ± 0.61	81.75 ± 0.40
SAM []	97.84 ± 0.27	97.17 ± 0.38	96.16 ± 0.55	86.28 ± 0.37	85.65 ± 0.40	83.66 ± 0.20	97.84 ± 0.27	97.28 ± 0.48	95.12 ± 0.60	86.28 ± 0.37	85.66 ± 0.20	82.12 ± 1.05
Ours	97.80 ± 0.54	$\mathbf{97.35 \pm 0.38}$	$\mathbf{96.58 \pm 0.48}$	$\mathbf{87.68 \pm 0.48}$	86.22 ± 0.57	$\mathbf{84.92 \pm 0.70}$	97.80 ± 0.54	$\mathbf{97.33 \pm 0.74}$	$\mathbf{96.04 \pm 0.59}$	$\mathbf{87.68 \pm 0.48}$	$\mathbf{86.20 \pm 0.16}$	$\mathbf{84.55 \pm 0.45}$

Table 6: Comparison on PACS for in-domain and out-of-domain metrics against *symm inc* and *symm exc* label noises with SWAD using ResNet-50.

PACS and +2.07% in VLCS, and achieves +0.18% improvement of averaged PACS and VLCS over the state-of-the-art method [4]. In addition, because SWAD [3] has been shown to be a state-of-the-art flatness-aware optimizer, we also evaluate the proposed method using SWAD and show the comparisons with other methods in Table 5. The results show that when including SWAD, the proposed method outperforms all the other methods and achieves +0.15% averaged improvement over MIRO [4] on two benchmarks.

Model Robustness. In Table 6, we report the comparison results under the two protocols *symm inc* and *symm exc*. The results show that the proposed method outperforms other competitors on both protocols and verify the robustness of our model against severe label noises.

4 Conclusion

In this paper, we propose a novel contrastive learning framework with prototype alignment and collaborative attention for robust in-domain and out-of-domain generalization. The proposed margin-based contrastive learning resolves the inter-class ambiguity and improves the out-of-domain generalizability. In addition, the proposed prototype alignment reduces the in-domain discrepancy by matching the latent feature representation of each class to the corresponding class prototype. Finally, the proposed collaborative attention method, by combining the dilated positive attention and the extended negative attention, effectively strengthens the model robustness. Experimental results on two benchmarks show that the proposed framework not only improves the baseline in terms of in-domain and out-of-domain evaluation metrics but also provides improved robustness against noisy labels.

References

[1] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *arXiv preprint arXiv:2110.10832*, 2021.

[2] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.

[3] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

[4] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. *arXiv preprint arXiv:2203.10789*, 2022.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[6] 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*, pages 1657–1664, 2013.

[7] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[10] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

[11] 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*, pages 770–778, 2016.

[12] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020.

[13] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[14] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 101–110, 2019.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[16] 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*, pages 5542–5550, 2017.

[17] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[18] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[19] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.

[20] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[21] Vladimir Vapnik. Statistical learning theory new york. *NY: Wiley*, 1(2):3, 1998.

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[23] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020.

[24] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019.

[25] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.

[26] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.