Enhancing Unlabeled Data Efficiency for Semi-Supervised Domain Generalization via Reliability-Guided Energy Prior

Panpan Ni, Kunlun Wu, Donghai Zhai*
School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China
*Corresponding author: dhzhai@swjtu.edu.cn

***Abstract*—This paper conducts an in-depth study of domain generalization (DG) to address distribution shifts when applying a model trained on a series of known target domains to unknown target domains. However, existing DG methods heavily rely on labels and high-confidence data, suffering from data efficiency issues and incurring expensive training costs. In this paper, we propose a novel multi-task semi-supervised domain generalization (RPCW) method that relaxes the strict requirement for plentiful labeled data in source domains while emphasizing superior generalization performance in unknown domains. This method comprises a contrastive task guided by reliability with energy prior and a semi-supervised task weighted by domain consistency and coherence, respectively, promoting the model to learn task-related features from unlabeled-confident and unlabeled-unconfident data and thereby jointly optimizing the covariate shift dilemma faced by the feature extractor. Experiments on two standard domain generalization datasets and a heterogeneous geological disaster dataset demonstrate that RPCW achieves better accuracy and generalization performance compared to state-of-the-art methods in scenarios with sparse labels. The code is available at https://github.com/2PNi/RPCW.**

***Index Terms*—domain generalization, energy prior, domain consistency, covariate shift.**

I. Introduction

Recently, computer vision has achieved significant breakthroughs thanks to advances in hardware computational capabilities and the rapid development of deep learning technologies. It has been widely applied in fields such as industrial automation, medical image analysis, and intelligent recognition of geological disasters. Deep convolutional neural networks (DCNNs) play a crucial role in complex visual recognition scenarios. However, various types of engineering production equipment in different settings can collect data that undermines the robustness and generalization capabilities of DCNNs. To mitigate the detrimental impact of domain shifts between source data and out-of-distribution target data on the generalization ability of DCNNs, domain adaptation (DA) methods leverage models trained on source data to match known target domains. Furthermore, domain generalization (DG) enables models to achieve excellent generalization in completely unknown target domains using only source data, thereby avoiding the risk of revealing potential information about the target domain inherent in DA approaches.

This work was supported by the National Natural Science Foundation of China under Grant 62576297.

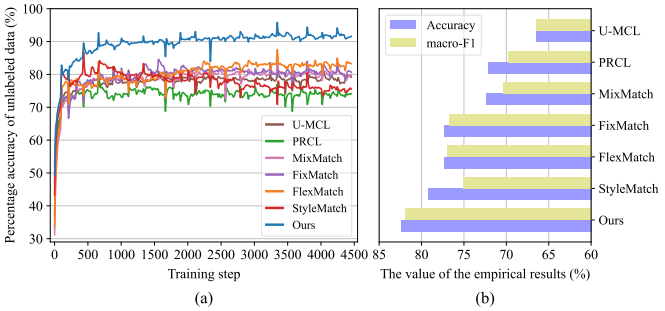


Fig. 1. Each method is implemented with 105 labeled data samples from the source domain under PACS, and the average accuracy results for the unlabeled data are calculated (left), while predictions for the unknown domain are made based on two metrics (right).

Although DG learns robust models capable of handling domain shifts from multiple source domains sharing consistent visual categories, its training process relies heavily on fully labeled data, resulting in high costs and low data efficiency. To address sparsely labeled tasks under domain shift scenarios, semi-supervised domain generalization (SSDG) has been investigated to enhance model generalizability by jointly leveraging a small number of labeled samples and a large amount of unlabeled data across domains.

Existing SSDG methods [1]–[3] tend to focus on improving the quality of pseudo-labels and leveraging reliable unlabeled samples, while disregarding the training value of unlabeled samples that fall below the confidence threshold (i.e., unlabeled-unconfident samples). For example, Yao et al. [1] proposed a Fourier-based model that considers trustworthy pseudo-labels to balance the fluctuation in pseudo-label confidence and extract low-level feature information across domains. FlexMatch [4] dynamically adjusts class thresholds based on the training scenario of the model, alleviating the difficulties associated with learning different classes due to the limitations of fixed thresholds. CWAEE [2] employs Fourier amplitude-based mixed augmentation to generate new samples regularized by consistency, thereby enhancing the generalization capability of SSDG in unknown domain scenarios involving unlabeled-unknown classes. However, these scientific solutions, designed to address the pseudo-label quality task, may constrain the efficiency of unlabeled data by excessively

prioritizing unlabeled-confident samples. Fig. 1(a) illustrates the assignment of predictions for unlabeled-unconfident data when such data are neglected during training in source domains with different distributions. Among semi-supervised algorithms, FlexMatch, which utilizes dynamic class thresholds, demonstrates outstanding performance, particularly excelling in addressing the challenges posed by unconfident data. We observe that the latent-untapped intelligence within unlabeled-unconfident data drives the model away from the vortex of confusion—distribution ambiguity. Notably, the preliminary results shown in Fig. 1(b) indicate that unlabeled-unconfident data facilitate improved generalization of the model in the target domain.

To intuitively understand and address the aforementioned issues, we propose a novel multi-task SSDG method based on reliability priors and consistency weighting (RPCW). To jointly leverage unlabeled-unconfident and trusted samples and enhance unlabeled data efficiency, we introduce two modules based on proxy contrastive and semi-supervised learning, respectively: 1) Reliability-Guided Energy Prior (RGEP) module; 2) Consistency Weighting (CW) module, which is based on bidirectional domain consistency and consistency weighting. We define a series of contrastive proxy tasks based on reliability and Bayesian energy fields to construct positively correlated components, thereby enhancing the capability of DCNN models to obtain robust understanding of the primary objects within unlabeled-unconfident samples. Moreover, we design a semi-supervised task by integrating feature-level weighting and enhanced multi-view consistency regularization, which further enables the model to analyze task-relevant features and improves its ability to accurately predict unlabeled samples. Experimental results demonstrate that the proposed method achieves outstanding robustness and generalization in sparse-label scenarios under domain shifting. In summary, we outline the main contributions of this work as follows:

- A semi-supervised domain generalization framework that integrates multi-tasking and multi-view learning, namely RPCW, is proposed to efficiently leverage unlabeled data and enhance the value of pseudo-label training, thereby enabling the model to achieve generalization across ambiguous distributions.
- We employ innovative techniques guided by reliability-driven energy prior and feature hierarchy weight to extract task-related features and enforce domain consistency from unlabeled-unconfident data.
- Experimental results demonstrate that RPCW outperforms existing classical methods in scenarios with sparse label distribution divergence. Especially, we extend RPCW to the geological disaster dataset xBD with only 7% labeled data, achieving state-of-the-art performance.

II. Methodology

A. Problem Definition of RPCW

For the problem of domain generalization, let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{D}$ represent the feature space, class label space, and domain label space, respectively. A domain of dataset S is characterized as a joint distribution $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{D}$, where X and Y represent the values of random variables $\mathcal{X}$ and $\mathcal{Y}$, respectively. Following existing works, we assume that the marginal distribution P_Y is fixed, while the marginal distribution P_X undergoes distributional shift among domains. We assume to have d source domains and one unknown target domain T, and denote them collectively as $S = \{S^{(1)}, \ldots, S^{(D)}, T\}$. Note that the distribution of T is not similar to any of the source domain $S^{(d)}$, but just shares the same class labels. All the data in the source domains in DG are *labeled*, denoted as $S_L^{(d)} = \{(\mathbf{x}_i, y_i) \mid \mathcal{D} = d\}_{i=1}^{n_d}$ where n_d is the number of labeled samples in domain d and $(\mathbf{x}_i, y_i) \sim P_{X,Y|\mathcal{D}=d}$. Note that the data in the target domain is unlabeled. In semi-supervised domain generalization, we assume the availability of additional unlabeled data in each source domains, denoted as $S_{UL}^{(d)} = \{\mathbf{u}_i \mid \mathcal{D} = d\}_{i=1}^{N_d}$ where N_d is the number of unlabeled samples in domain d and $\mathbf{u}_i \sim P_{X|\mathcal{D}=d}$. The sizes of labeled data is typically much smaller than that of the unlabeled data, i.e., $n_d \ll N_d$.

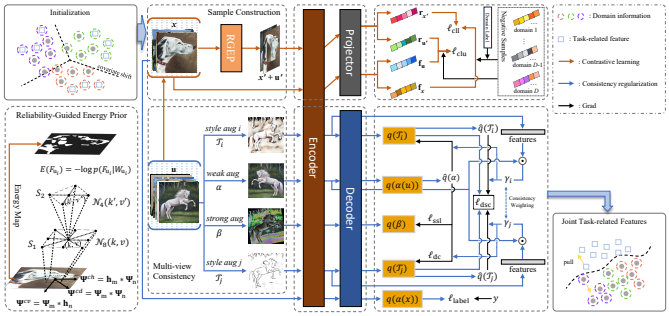


Fig. 2. The proposed multi-task framework, RPCW, includes a reliability-based contrastive task and a domain consistency semi-supervised task to effectively distinguish task-related features in SSDG.

Finally, in the proposed multi-task semi-supervised domain generalization (abbreviated as RPCW) problem, we employ self-supervised contrastive learning to further enhance the efficacy of models to solve the same semi-supervised domain generalization problem. Fig. 2 concretizes the working concept of RPCW in the form of a parallel end-to-end framework. It integrates inductive bias operations for multi-view main-auxiliary task transfer, thereby guiding the model to progressively fit and interpret more complex scenarios.

B. Energy Field on Reliability

Fast wavelet transform employs high-pass decomposition and low-pass orthogonal mirror filter banks, along with subsampling, to compute the detail and approximation components within an image. Compared with Fourier transforms, it additionally provides localization in the time domain, alleviating the limitations imposed by nonstationary signals in spectral analysis. To investigate the relationship between key points of image features and their frequency components, the wavelet coefficients at the two levels of scale are defined as:

$$W_{\mathbf{u}_i}^{(s)}(k,v) = \{W_{\mathbf{h}}^{(s)}\big(\mathbf{u}_i(k,v)\big), W_{\Psi^j}^{(s)}\big(\mathbf{u}_i(k,v)\big)\} \quad (1)$$

where $W_{\mathbf{h}}^{(s)}$ is the low-frequency subband component, while $W_{\Psi^j}^{(s)}$ consists of the high-frequency subband components in three directions ($j \in \{cv, ch, cd\}$). Based on the sparsity of wavelet coefficients, it is assumed that the coefficients of feature points follow a Laplace distribution (i.e., $W_{\mathbf{u}_i}^{(s)} \sim \mathcal{N}(0, b_s)$), while the coefficients of non-feature points follow a Gaussian distribution (i.e., $W_{\mathbf{u}_i}^{(s)} \sim \mathcal{N}(0, \sigma_s^2)$). The likelihood modeling $P\left(W_{\mathbf{u}_i}^{(s)} \mid F_{\mathbf{u}_i}^{(s)}\right)$ describes the probability of a coefficient exhibiting a feature at different scales (i.e., $s \in \{s_1, s_2\}$), which is defined as follows:

$$P\left(W_{\mathbf{u}_i}^{(s)}(k,v) \mid F_{\mathbf{u}_i}^{(s)}(k,v) = 1\right) = \frac{1}{2b_s} \exp\left(-\frac{|W_{\mathbf{u}_i}^{(s)}(k,v)|}{b_s}\right) \quad (2)$$

$$P\left(W_{\mathbf{u}_i}^{(s)}(k,v) \mid F_{\mathbf{u}_i}^{(s)}(k,v) = 0\right) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{W_{\mathbf{u}_i}^{(s)}(k,v)^2}{2\sigma_s^2}\right) \quad (3)$$

where $F_{\mathbf{u}_i}^{(s)}$ is a binary random variable representing the state of pixel feature points. If $F_{\mathbf{u}_i}^{(s)}(k,v) = 1$, it indicates that the pixel (k,v) possesses a feature at scale s; otherwise, it is regarded as noise. $W_{\mathbf{u}_i}^{(s)}$ converts the spatial correlation between natural image pixels into weak correlation among the coefficients. Assuming that the wavelet coefficients are conditionally independent of $F_{\mathbf{u}_i}$, and combining the reliability [5] formula with (2) and (3), we obtain:

$$\begin{aligned} P(W_{\mathbf{u}_i}|F_{\mathbf{u}_i}) &= \prod_{s,k,v} P\left(W_{\mathbf{u}_i}^{(s)}(k,v) \mid F_{\mathbf{u}_i}^{(s)}(k,v)\right) \\ &= \exp\big(-E_d(F_{\mathbf{u}_i})\big) \end{aligned} \quad (4)$$

where E_d represents the energy function of the data term. To align features across different scales and capture spatial correlations, we construct the prior probability $P(F_{\mathbf{u}_i}) = \exp\big(-E_s(F_{\mathbf{u}_i})\big)$ for $F_{\mathbf{u}_i}$ based on reliability. After integrating pixel-level feature information from image samples, the energy formula of the data term is as follows:

$$\begin{aligned} E_d(F_{\mathbf{u}_i}) = &\sum_{k=1,v=1}^{w/2,h/2} \Big(\frac{W_{\mathbf{u}_i}^{(s_1)}(k,v)^2}{2\sigma_{s_1}^2}\big(1 - F_{\mathbf{u}_i}^{(s_1)}(k,v)\big) \\ &+ \frac{|W_{\mathbf{u}_i}^{(s_1)}(k,v)|}{b_{s_1}} F_{\mathbf{u}_i}^{(s_1)}(k,v)\Big) \\ &+ \sum_{k=1,v=1}^{w/4,h/4} \Big(\frac{W_{\mathbf{u}_i}^{(s_2)}(k,v)^2}{2\sigma_{s_2}^2}\big(1 - F_{\mathbf{u}_i}^{(s_2)}(k,v)\big) \\ &+ \frac{|W_{\mathbf{u}_i}^{(s_2)}(k,v)|}{b_{s_2}} F_{\mathbf{u}_i}^{(s_2)}(k,v)\Big) + C_t \end{aligned} \quad (5)$$

where C_t is a constant term related to image dimensions $\mathbf{u}_i \in \mathbb{R}^{w \times h}$, wavelet coefficients, and the parameters of the likelihood distribution. The smoothness term of the energy function achieves feature consistency and mitigates noise interference by performing both inter-scale and spatial smoothing:

$$\begin{aligned} E_s(F_{\mathbf{u}_i}) = &\sum_{k=1,v=1}^{w,h} \sum_{(i,j)}^{\mathcal{N}_8(k,v)} \mathcal{V}\big(F_{\mathbf{u}_i}^{(s_1)}(k,v), F_{\mathbf{u}_i}^{(s_1)}(i,j)\big) \\ &+ \sum_{(i,j)}^{\mathcal{N}_4(k',v')} \mathcal{V}\big(F_{\mathbf{u}_i}^{(s_1)}(k,v), F_{\mathbf{u}_i}^{(s_2)}(k',v')\big) \\ &+ \sum_{k'=1,v'=1}^{w/2,h/2} \sum_{(i,j)}^{\mathcal{N}_8(k',v')} \mathcal{V}\big(F_{\mathbf{u}_i}^{(s_2)}(k',v'), F_{\mathbf{u}_i}^{(s_2)}(i,j)\big) \end{aligned} \quad (6)$$

The potential function $\mathcal{V}(\cdot,\cdot) = 1 - \delta(\cdot,\cdot)$ is used to penalize cases where features are inconsistent within the scale range, and $\mathcal{N}_i$ represents the connectivity number of the spatial domain. $\delta(a_1, a_2)$ is an indicator function; if $a_1 = a_2$, then $\delta = 1$. The evidence term $P(W)$ denotes the probability of observing the coefficient itself, as in (1), independent of the state of the feature. We equate the posterior probability of feature $F_{\mathbf{u}_i}$ to $P(F_{\mathbf{u}_i}|W_{\mathbf{u}_i}) = \exp\big(-\big(E_d(F_{\mathbf{u}_i}) + E_s(F_{\mathbf{u}_i})\big)\big)$ and map it to the energy functions of the data term and the smoothness term. The set of image sample features after performing energy function optimization is defined as:

$$\hat{F}_{\mathbf{u}_i} = \arg\min_F E_d(F_{\mathbf{u}_i}) + E_s(F_{\mathbf{u}_i}) \quad (7)$$

C. Task-specific Contrastive Learning

The energy minimization process adaptively constructs the correlations between image wavelet coefficients and features, reliably distinguishing microscopic signal characteristics and locating key regions. The proposed task-related contrastive module implements a positive sample generation strategy (i.e., region rotation augmentation) to establish a semantic similarity-driven representation generalization learning framework, enabling the model to efficiently utilize unlabeled and low-confidence samples filtered from semi-supervised learning tasks. The domain-supervised contrastive learning loss function reduces the distance between the semantic feature distributions of samples with similar domain prior knowledge labels and is formulated as:

$$\ell_{\text{cll}} = -\frac{1}{n_d}\sum_{i=1}^{n_d} \log \frac{PoS(\mathbf{x}_i)}{PoS(\mathbf{x}_i) + \sum_{\mathbf{k}\in\{\mathbf{Q}_d\}} \exp(\mathbf{f}_i^\top \mathbf{k}/\tau)} \text{ where}$$
$$PoS(\mathbf{x}_i) = \exp(\mathbf{f}_i^\top \mathbf{r}_{\mathbf{x}_i'}/\tau) + \sum_{\{j\neq i \mid \mathcal{D}_i = \mathcal{D}_j \wedge y_i = y_j\}} \exp(\mathbf{f}_i^\top \mathbf{f}_j/\tau) + \exp(\mathbf{f}_i^\top \mathbf{r}_{\mathbf{x}_j'}/\tau) \quad (8)$$

The negative sample features $\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_q$ are derived from the subsets $\mathbf{Q}_1, ..., \mathbf{Q}_D$ of the source domain set d. The temperature coefficient τ is implemented to control the concentration degree of similar distributions. The local feature regions retrieved according to (7) are replaced in the corresponding positions of the augmented samples $\mathbf{r}(\hat{F}_{\mathbf{x}})$ after rotation augmentation. To capture task-related features in unlabeled-unconfident samples and improve model efficiency, the unsupervised-contrastive learning loss of the proposed multi-task learning framework is formulated as:

$$\ell_{\text{clu}} = -\frac{1}{N_d}\sum_{i=1}^{N_d} \log \frac{\exp(\mathbf{f}_i^\top \mathbf{r}_{\mathbf{u}_i'}/\tau)}{\exp(\mathbf{f}_i^\top \mathbf{r}_{\mathbf{u}_i'}/\tau) + \sum_{j=1}^{N_d-1} \exp(\mathbf{f}_i^\top \mathbf{f}_j/\tau)} \quad (9)$$

D. Multi-view Consistency Regularization

Existing semi-supervised learning methods [4], [6], [7] focus on effectively utilizing unlabeled samples while overlooking multiple domains and domain generalization. Hence, we design two additional consistency regularizations based on style transfer and feature consistency weighting and optimize the feature extractor of the multi-task framework in parallel with contrastive loss. The key observation is that style transfer acts as an effective path to remove domain-specific features and serves as a strong augmentation in itself.

Given a sample $\mathbf{u}_i$ from domain d, the style enhancement algorithm [8] is applied to migrate the style information of a random sample from the domain set $z \in \{i, j\}$ and completely replace the initial style of $\mathbf{u}_i$. Then, the model is required to predict that the outcome of the augmented sample still matches the pseudo-label from the weak augmentation α. The loss with respect to such domain consistency is:

$$\ell_{\mathrm{dc}} = \mathbb{E}_d \mathbb{E}_{\mathbf{u} \in S_{UL}^{(d)}} \mathbb{E}_{z \neq d} \left[\mathbb{I}_{\max(q(\alpha)) \geq \psi} \cdot \gamma_z \cdot H\big(q(\mathscr{T}_z) \parallel \hat{q}(\alpha)\big) \right] \text{ where}$$
$$\gamma_z = \frac{\exp(\cos(enc(\alpha), enc(\mathscr{T}_z))/\tau)}{\sum_{z' \in \{i,j\}} \exp(\cos(enc(\alpha), enc(\mathscr{T}_{z'}))/\tau)} \quad (10)$$

Here, $q(\alpha) := p\big(\alpha(\cdot)\big)$ represents the predicted class distribution of the model concerning the augmented operator α, and $\hat{q}(\alpha)$ is the one-hot distribution obtained by solidifying $q(\alpha)$. $\mathscr{T}_z$ denotes the style transfer operator that uses a random sample from a different domain z as the input to equip $\mathbf{u}$ to the style in domain z. $\cos(\cdot, \cdot)$ is the cosine similarity that measure the angular closeness of two feature encodings. $enc(\alpha) := enc\big(\alpha(\mathbf{u})\big)$ is the feature encoding of the sample augmented by the operator α. $H(q \parallel p)$ is the usual cross-entropy loss of the distributions q with respect to the ground truth distribution p. Another novel addition to the multi-view consistency is to use two style transfers as strong augmentations. We refer to the two augmentation as dual branches and let them mutually promote and supervise each other. The mutual style consistency loss is defined as:

$$\ell_{\mathrm{dsc}} = \gamma_i \cdot H\big(q(\mathscr{T}_j) \parallel \hat{q}(\mathscr{T}_i)\big) + \gamma_j \cdot H\big(q(\mathscr{T}_i) \parallel \hat{q}(\mathscr{T}_j)\big) \quad (11)$$

E. Training Objective

RPCW is a multi-task semi-supervised domain generalization model implemented to enhance model efficiency by utilizing unlabeled samples in parallel. It employs joint training on both contrastive and semi-supervised tasks for feature extraction, aiming to learn general task-related features. In the semi-supervised branch, previous approaches included a supervised loss (ℓ_{label}) for labeled samples and an unsupervised loss (ℓ_{ssl}) for unlabeled samples. ℓ_{ssl} leverages highly confident predicted class labels (pseudo-labels) of weakly augmented data to guide the model in aligning the predicted distribution of strongly augmented inputs with the pseudo-labels. Thus, the objective for training the model using the RPCW method is

$$\ell = \ell_{\mathrm{cll}} + \lambda_1 \cdot \ell_{\mathrm{clu}} + \ell_{\mathrm{label}} + \lambda_2 \cdot \ell_{\mathrm{ssl}} + \lambda_3 \cdot \ell_{\mathrm{dc}} + \lambda_4 \cdot \ell_{\mathrm{dsc}} \quad (12)$$

III. Experiments

A. Experimental Setup and Implementation

1) Datasets: **PACS** [9] dataset was constructed to address the issues of minimal domain abstraction differences and non-photorealistic visual diversity, thereby facilitating research on cross-abstract domain instance matching in domain generalization. It consists of 4 domains—C: Cartoon, P: Photo, A: Art painting, and S: Sketch—each containing 7 categories ('giraffe', 'horse', 'elephant', 'house', 'guitar', 'dog', and 'person'), totaling 9,991 images. The visual bias within the **VLCS** [10] dataset encompasses factors such as the preferences of workers during data collection and systematic imaging errors. VLCS is comprised of 4 domains: Caltech101, SUN09, LabelMe, and PASCAL, with each containing 5 common categories, amounting to a total of 10,729 images. **xBD** [11] dataset was developed to support intelligent recognition tasks for high-resolution (1024×1024) satellite remote sensing images of geological disasters, presenting generalization challenges such as imbalanced category distribution, non-unified equipment models, and a lack of expert-level annotation. Four types of geological disasters in xBD—namely earthquake, volcanic eruption, landslide, and tsunami—are selected to assess the robust performance and engineering value of semi-supervised learning methods and SSDG methods in geological remote sensing scenarios, comprising a total of 1,196 images. To enable standardized comparison of the generalization and disaster recognition capabilities between RPCW and baseline algorithms, the training and test sets are split in a 7:3 ratio.

2) Implementation Details: All experiments were conducted under the software environment of Ubuntu 20.04, PyTorch 1.12.1, and CUDA 11.6, as well as the hardware environment of an Intel(R) Xeon(R) Gold 6342 CPU (2.80 GHz) and two NVIDIA RTX 3090 GPUs (24 GB each). We implemented a ResNet34 pretrained on the ImageNet dataset as the backbone for all benchmark models and exploited a normative SGD optimizer with a momentum of 9×10^{-1} and an initial learning rate of 3×10^{-3} to adjust the backbone parameters. The training epochs for PACS, VLCS, and xBD are set to 40, 30, and 40, respectively. The classifier and feature projection head employed an MLP composed of batch normalization and ReLU layers with an initial learning rate of 1×10^{-2}. Moreover, we set the weights of ℓ as $\lambda_1 = 0.6$, $\lambda_2 = 0.4$, $\lambda_3 = 0.3$, and $\lambda_4 = 0.2$. The batch size for labeled and unlabeled samples in the source domain is set to 48. Noted that only source domain samples are considered during model training, and testing is performed exclusively on target domain samples. The temperature coefficient τ and the confidence threshold ψ are set to 0.05 and 0.95. To ensure a fair evaluation of model performance, Top-1 classification accuracy, macro-F1, and mean average precision (mAP) are used as evaluation metrics (higher values indicate better predictive performance).

B. Comparative Analyses

To comprehensively compare the proposed method, RPCW, we considered three different categories: domain generalization methods [12], [13], including DDG, CrossGrad, and

TABLE I

THE RESULTS OF DOMAIN GENERALIZATION PRESENT OUR METHOD AND VARIOUS BENCHMARKS UNDER THE LABEL RATIO STRATEGY OF 210 LABELS ON THE PACS DATASET ACROSS THREE EVALUATION METRICS. **L** AND **U** REPRESENT THE UTILIZATION OF LABELED AND UNLABELED DATA, RESPECTIVELY. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED, RESPECTIVELY.

Methods	L	U	Accuracy (%)					macro-F1 (%)					mAP (%)				
			C	P	A	S	Avg.[a]	C	P	A	S	Avg.	C	P	A	S	Avg.
DDG	✔		55.67	74.13	70.07	43.48	60.84	56.39	71.98	69.50	35.72	58.40	45.75	58.89	54.20	25.43	46.07
CrossGrad	✔		55.25	83.65	71.39	48.98	64.82	64.83	82.01	70.66	40.06	64.39	45.43	72.73	56.09	30.02	51.07
DDAIG	✔		64.80	85.15	78.47	62.83	72.81	64.87	81.83	78.13	64.91	72.44	53.97	73.31	64.94	52.35	61.14
U-MCL	✔	✔	65.83	92.75	70.26	61.15	72.50	67.62	91.44	71.39	56.30	71.69	56.61	85.74	59.67	43.78	61.45
PRCL	✔	✔	75.51	92.63	79.35	57.43	76.23	76.04	91.60	80.10	53.60	75.34	64.30	85.59	68.74	43.76	65.60
MixMatch	✔	✔	68.77	92.57	74.80	66.80	75.74	67.16	91.81	74.81	68.10	75.47	55.16	85.92	61.54	52.97	63.90
FixMatch	✔	✔	69.16	89.64	78.12	78.74	78.92	68.27	88.44	77.37	79.32	78.35	57.10	80.91	64.99	68.65	67.91
FlexMatch	✔	✔	75.85	92.99	83.54	73.47	81.46	77.06	92.01	83.80	75.57	82.11	66.17	86.49	73.55	64.22	72.61
StyleMatch	✔	✔	73.63	89.10	83.54	79.35	81.41	75.11	87.55	82.94	81.37	81.74	63.98	79.68	72.46	71.69	71.95
Ours	✔	✔	**80.20**	**93.41**	**85.84**	**81.16**	**85.15**	**80.76**	**92.07**	**85.29**	**82.50**	**85.16**	**70.30**	**86.65**	**75.47**	**73.04**	**76.37**

[a]Note that '**Avg.**' represents the average operation of the four test domains.

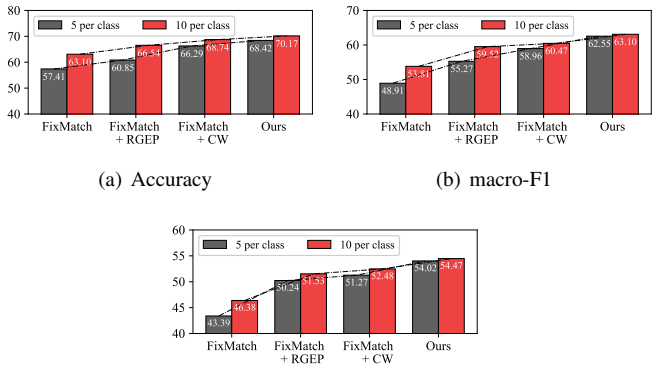
TABLE II

THE RESULTS OF DOMAIN GENERALIZATION PRESENT OUR METHOD AND VARIOUS BENCHMARKS UNDER THE LABEL RATIO STRATEGY OF 105 LABELS ON THE PACS DATASET ACROSS THREE EVALUATION METRICS.

Methods	L	U	Accuracy (%)					macro-F1 (%)					mAP (%)				
			C	P	A	S	Avg.	C	P	A	S	Avg.	C	P	A	S	Avg.
DDG	✔		49.74	65.69	59.52	40.10	53.76	49.32	63.69	59.45	27.21	49.92	38.86	49.84	44.19	21.91	38.70
CrossGrad	✔		52.94	71.92	57.76	41.73	56.09	52.11	70.62	58.10	31.62	53.11	41.63	58.01	43.24	25.49	42.09
DDAIG	✔		55.84	71.80	61.96	42.69	58.07	55.59	69.85	61.75	31.87	54.77	44.96	58.10	46.05	26.13	43.81
U-MCL	✔	✔	65.06	89.76	67.43	43.66	66.48	66.31	87.96	68.08	43.59	66.49	53.60	81.36	56.63	37.44	57.26
PRCL	✔	✔	70.78	89.10	77.64	50.87	72.10	69.18	87.00	77.94	44.77	69.72	58.24	78.77	65.93	36.57	59.88
MixMatch	✔	✔	65.66	90.78	69.73	63.09	72.32	66.13	89.48	69.67	56.35	70.41	52.34	82.37	55.85	44.68	58.81
FixMatch	✔	✔	69.88	86.53	77.00	75.64	77.26	69.82	84.51	75.61	**76.77**	76.68	58.25	75.06	62.40	**66.28**	65.50
FlexMatch	✔	✔	69.62	90.84	80.96	68.00	77.36	71.43	89.39	**81.14**	65.99	76.99	57.98	82.75	69.72	53.43	65.97
StyleMatch	✔	✔	76.24	83.41	78.71	**78.56**	79.23	76.73	76.11	78.66	68.88	75.10	64.60	71.39	66.87	57.52	65.10
Ours	✔	✔	**78.63**	**91.86**	**81.64**	77.47	**82.40**	**80.13**	**90.33**	81.06	76.14	**81.92**	**69.08**	**84.31**	**69.73**	64.92	**72.01**

DDAIG; semi-supervised learning methods [4], [6], [7], including U-MCL, PRCL, MixMatch, FixMatch, and FlexMatch; and a semi-supervised domain generalization method (e.g., StyleMatch [3]). Following existing DG experimental protocols, three domains are randomly selected as source domains (for training), with the remaining domain used as the unseen test domain. In the SSDG setting, 5 or 10 labeled data samples per class are randomly drawn from each source domain, and the rest are used as unlabeled data.

The predictive results of various baseline methods under the 210 and 105 labeled data settings are presented in Table 1 and Table 2, respectively. RPCW achieves state-of-the-art performance with 210 labeled data and attains optimal or sub-optimal results across three metrics with 105 labeled data. In Table I, RPCW achieves the highest average Top-1 Accuracy of 0.8515, macro-F1 of 0.8516, and mAP of 0.7637, surpassing the second-best FlexMatch by 3.69%, 3.05%, and 3.76%, respectively. Compared with the DG algorithm (i.e., DDAIG), RPCW demonstrates significant performance improvements under 'Avg.' across four test cases (target domains), with increases of +12.34% (85.15 vs. 72.81) in Top-1 Accuracy, +12.72% (85.16 vs. 72.44) in macro-F1, and +15.23% (76.37 vs. 61.14) in mAP. Table II further confirms that RPCW consistently drives the model to maintain robust performance under more sparse label generalization scenarios. Although FixMatch achieves the highest macro-F1 (76.77%) and mAP (66.28%) on the 'Sketch' target domain, its overall performance lags behind RPCW and FlexMatch. The highest average mAP in Table II is 0.7201, with RPCW improving upon the second-best result of 0.6597 by 6.04%. In summary, the proposed RPCW method maintains superior generalizability and robustness under varying degrees of sparse label scenarios.



(a) Accuracy (b) macro-F1

(c) mAP

Fig. 3. The results of the three evaluation metrics for VLCS with two core components. Note that all results are the average values obtained after testing with each of the four domains sequentially as the target domain.

C. Ablation Study

The above experimental analyses indicate that our method, RPCW, demonstrates comprehensive performance in sparse label tasks under domain shift. However, how should the importance of the reliability-guided energy prior (RGEP) module and the feature-level consistency weighting (CW) module in RPCW be evaluated? Using FixMatch as the baseline, we sequentially incorporate the two proposed modules to conduct ablation analysis under settings of 5 labels and 10 labels per

TABLE III
OUR METHOD AND VARIOUS SEMI-SUPERVISED (DOMAIN GENERALIZATION) BENCHMARKS ARE EVALUATED USING A STRATEGY WITH 84 LABELS (7%) ON THE XBD DATASET ACROSS THREE METRICS.

Methods	L	U	Accuracy (%)	macro-F1 (%)	mAP (%)
EntMin	✔	✔	83.57	83.88	75.45
MeanTeacher	✔	✔	77.72	77.80	67.76
U-MCL	✔	✔	81.34	82.80	74.94
PRCL	✔	✔	83.57	85.47	79.04
MixMatch	✔	✔	85.24	86.65	80.24
FixMatch	✔	✔	83.57	85.93	79.92
FlexMatch	✔	✔	83.57	85.49	79.37
StyleMatch	✔	✔	81.06	82.57	74.39
Ours	✔	✔	**88.02**	**88.51**	**81.64**

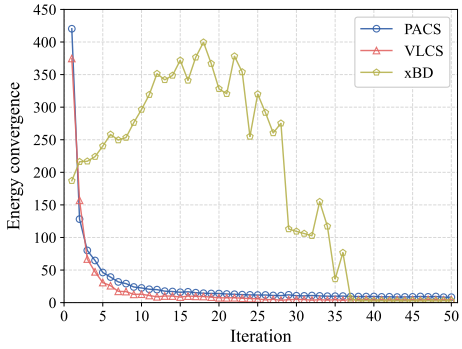


Fig. 4. Iterative optimization of energy fields across three datasets.

category. Fig. 3 reports the average classification performance of the RGEP and CW components across all target domains on the VLCS, covering three metrics. Compared to the baseline in terms of the macro-F1 metric, RPCW shows an average improvement of 11.47% over FixMatch. Fig. 3 demonstrates that each core component has a positive effect on the baseline method, with CW contributing more significantly to the model's generalization performance. Moreover, there is no obvious antagonism between the two proposed modules, suggesting that they complement each other. The joint integration of the modules provides an average improvement of 3.36% and 2.37% in mAP compared to each module individually.

D. Further Analyses

To further explore the practical value of our method in real-world engineering scenarios, we conduct a detailed performance analysis on publicly available geohazard datasets. xBD is a disaster assessment dataset that integrates data collected from the Internet and multispectral satellite sources, characterized by its high resolution and rich label information. However, challenges such as label scarcity and data heterogeneity commonly encountered in real engineering applications can prevent DCNNs from achieving their full potential. Notably, the structural regions in geological remote sensing images often contain complex linear and spectral features, as well as other optical information.

According to the geological disaster identification results recorded in Table III, our method and the MixMatch method achieve the best and second-best performances, respectively, across three evaluation metrics. Additionally, RPCW surpasses the second-best method by 2.78% in Top-1 Accuracy and by 1.86% in macro-F1. Compared to the SSDG method (i.e., StyleMatch), our method outperforms in the geological disaster visual classification task by an average margin of +6.72%. In summary, the proposed method conducts effective analysis of high-dimensional and structurally complex remote sensing images of geological disasters, demonstrating strong generalization performance in scenarios with label scarcity commonly encountered in engineering applications. To visualize the convergence process of the reliability-guided energy prior module, the average energy points for 50 iterations across three datasets are jointly displayed in Fig. 4.

IV. CONCLUSION

In this paper, a novel multi-task semi-supervised domain generalization method called RPCW is proposed to address the negative impacts of label scarcity and distribution shifts, thereby improving the efficiency of unlabeled data and the generalization performance of the model. RPCW consists of a reliability and Bayesian theory-based contrastive learning module and a domain consistency weighting module, which are optimized in parallel. Experimental results demonstrate that RPCW outperforms comparison methods significantly. In future work, we will explore dynamic thresholds to flexibly filter unlabeled data and the issues of model transfer (e.g., the transfer effects at different layers of the model).

REFERENCES

[1] H. Yao, X. Hu, and X. Li, "Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation," in *Proceedings of 36th AAAI*, vol. 36, 2022, pp. 3099–3107.

[2] L. Zhang, J.-F. Li, and W. Wang, "Semi-supervised domain generalization with known and unknown classes," in *Proceedings of 37th NeurIPS*, vol. 36, 2023, pp. 28 735–28 747.

[3] K. Zhou, C. C. Loy, and Z. Liu, "Semi-supervised domain generalization with stochastic stylematch," *International Journal of Computer Vision*, vol. 131, pp. 2377–2387, 2023.

[4] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Proceedings of 35th NeurIPS*, 2021.

[5] A. Bagherinia, B. Minaei-Bidgoli, M. Hosseinzadeh, and H. Parvin, "Reliability-based fuzzy clustering ensemble," *Fuzzy Sets and Systems*, vol. 413, pp. 1–28, 2021.

[6] X. Lu, L. Li, L. Jiao, X. Liu, F. Liu, W. Ma, and S. Yang, "Uncertainty-aware semi-supervised learning segmentation for remote sensing images," *IEEE Transactions on Multimedia*, vol. 27, pp. 5548–5562, 2025.

[7] H. Xie, C. Wang, J. Zhao, Y. Liu, J. Dan, C. Fu, and B. Sun, "Prcl: Probabilistic representation contrastive learning for semi-supervised semantic segmentation," *International Journal of Computer Vision*, vol. 132, pp. 4343–4361, 2024.

[8] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *proceedings of 32nd CVPR*, 2019, pp. 5880–5888.

[9] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of 15th ICCV*, 2017.

[10] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *Proceedings of 14th ICCV*, 2013, pp. 1657–1664.

[11] W. Sun, P. Bocchini, and B. D. Davison, "Applications of artificial intelligence for disaster management," *Natural Hazards*, vol. 103, pp. 2631–2689, 2020.

[12] Z. Sun, Z. Shen, L. Lin, Y. Yu, Z. Yang, S. Yang, and W. Chen, "Dynamic domain generalization," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 1342–1348.

[13] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," in *Proceedings of 6th ICLR*, 2018.