Improving Pseudo-Labelling and Enhancing Robustness for Semi-Supervised Domain Generalization

Adnan Khan[a], Mai A. Shaaban[b] and Muhammad Haris Khan[c]
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, U.A.E.

Abstract: Beyond attaining domain generalization (DG), visual recognition models should also be data-efficient during learning by leveraging limited labels. We study the problem of Semi-Supervised Domain Generalization (SSDG) which is crucial for real-world applications like automated healthcare. SSDG requires learning a cross-domain generalizable model when the given training data is only partially labelled. Empirical investigations reveal that the DG methods tend to underperform in SSDG settings, likely because they are unable to exploit the unlabelled data. Semi-supervised learning (SSL) shows improved but still inferior results compared to fully-supervised learning. A key challenge, faced by the best performing SSL-based SSDG methods, is selecting accurate pseudo-labels under multiple domain shifts and reducing overfitting to source domains under limited labels. In this work, we propose new SSDG approach, which utilizes a novel uncertainty-guided pseudo-labelling with model averaging (UPLM). Our uncertainty-guided pseudo-labelling (UPL) uses model uncertainty to improve pseudo-labelling selection, addressing poor model calibration under multi-source unlabelled data. The UPL technique, enhanced by our novel model averaging (MA) strategy, mitigates overfitting to source domains with limited labels. Extensive experiments on key representative DG datasets suggest that our method demonstrates effectiveness against existing methods. Our code and chosen labelled data seeds are available on GitHub: https://github.com/Adnan-Khan7/UPLM.

1 INTRODUCTION

Domain shift (Tzeng et al., 2015) (Hoffman et al., 2017) is an important challenge for several computer vision tasks e.g., object recognition (Krizhevsky et al., 2017). Among others, domain generalization (DG) has emerged as a relatively practical paradigm for handling domain shifts and it has received increasing attention in the recent past (Li et al., 2017) (Zhou et al., 2021b) (Khan et al., 2021). The goal is to train a model from the data available from multiple source domains that can generalize well to an unseen target domain. We have seen several DG approaches (Huang et al., 2020) (Wang et al., 2020) that have displayed promising performance across various benchmarks (Li et al., 2017) (Venkateswara et al., 2017). However, the performance of many DG methods is sensitive to the availability of sufficiently annotated quality data from available source domains. As such, this requirement is difficult to meet in several

real-world applications of these models e.g., healthcare, autonomous driving and satellite imagery (Khan et al., 2022b). Besides attaining generalization, it is desirable for the learning algorithms to be efficient in their use of data. This means that the model can be trained using a minimal amount of labelled data to reduce development costs. This concept is closely related to semi-supervised learning (SSL) (Grandvalet and Bengio, 2004) (Tarvainen and Valpola, 2017) which seeks to make use of large amounts of unlabelled data along with a limited amount of labelled data for model training. To this end, this paper studies the relatively unexplored problem of semi-supervised domain generalization (SSDG). It aims to tackle both the challenges of model generalization as well as data-efficiency within a unified framework. Both DG and SSDG share the common goal of training models capable of performing well on unseen target domain using only source domain data for training. However, DG is based on the assumption that all data from source domains is fully labelled, while SSDG operates under the SSL setting, where only few images within each source domain have labels and large num-

[a] https://orcid.org/0000-0002-0583-9863
[b] https://orcid.org/0000-0003-1454-6090
[c] https://orcid.org/0000-0001-9746-276X

ber of images are unlabelled. Figure 1 shows the visual comparison among the settings of three related paradigms.

We note that, the DG methods, which cannot utilize unlabelled data, tend to show degraded performance upon reducing the quantity of labelled data (Zhou et al., 2021a). On the other hand, the SSL methods e.g, FixMatch (Sohn et al., 2020) display relatively better performance than the DG methods under limited labels setting, still their performance is noticeably inferior to the fully labelled setting. SSL methods lose performance in SSDG setting due to differences in data distributions between various source domains and the limited amount of labelled data available, which are unique challenges to SSDG problem.

We propose a systematic approach, namely **U**ncertainty-Guided **P**seudo-**L**abelling with **M**odel Averaging (UPLM) to tackle the challenges in SSDG. First, we develop an uncertainity-guided pseudo-labelling (UPL) technique to overcome the problem of noisy pseudo-labels (PLs), typically produced by confidence-based methods (Sohn et al., 2020) under domain shift. We leverage model's predictive uncertainty to develop a pseudo-label selection criterion that provides accurate PLs by mitigating the impact of miscalibrated predictions, especially for out-domain data. Second, we propose a novel model averaging (MA) technique which overcomes the effect of overfitting to limited labels in source domains to achieve cross-domain generalization at the inference stage. Through empirical results we show the intuition and motivation behind our two components. Our suggested approach demonstrates its effectiveness in addressing the SSDG problem when compared to other SSDG and SSL methods, as evidenced by thorough experimentation on four demanding DG datasets.

2 RELATED WORK

Domain Generalization. (Vapnik, 1999) is recognized as pioneering work in Domain Generalization (DG), introducing Empirical Risk Minimization (ERM) to minimize the sum of squared errors across diverse source domains. It led to various approaches for extracting domain-invariant features, such as (Muandet et al., 2013) employing maximum mean discrepancy (MMD), (Ghifary et al., 2015) introducing a multi-task autoencoder, and (YANG and Gao, 2013) using canonical correlation analysis (CCA). Meta-learning frameworks, like those in (Shu et al., 2021), have also been employed for domain generalization to simulate training domain shifts. For semantic alignment, domain generalization such as

(Kim et al., 2021) and (Dou et al., 2019), leverage self-supervised contrastive formulations (Khan et al., 2022a). The idea of improving diversity in source domains is shown to be effective for DG (Khan et al., 2021). (Volpi et al., 2018) applied a wasserstein constraint in semantic space and (Shankar et al., 2018) introduced Crossgrad training as a DG method to enhance DG. The aforementioned DG methods assume supervised settings with fully labeled source domain data for training. However, there is limited research on enhancing DG performance in scenarios with scarce labeled data. This work addresses the SSDG problem, unifying data efficiency and model generalization, and proposes a principled approach to tackle relevant SSDG challenges.

Uncertainty Estimation in DNNs. Quantifying uncertainty in deep nerual networks (DNNs) has remained an important research direction (Kendall and Cipolla, 2016). These methods proposed to quantify the uncertainty associated with the predictions made by DNNs. For instance, (Gal and Ghahramani, 2016) presented dropout training in DNNs as approximate Bayesian inference to model uncertainty. (Kendall and Gal, 2017) developed a Bayesian DNNs framework that combines input-dependent aleatoric uncertainty with epistemic uncertainty. The work of (Lakshminarayanan et al., 2017) proposed alternative to Bayesian DNNs, which includes ensembles and adversarial training, for estimating predictive uncertainty on out-of-distribution examples. (Smith and Gal, 2018) investigated measures of uncertainty focusing on mutual information and proposed an improvement in uncertainty estimation using probabilistic model ensembles. In this work, we leverage model uncertainty from Monte-Carlo (MC) dropout technique (Gal and Ghahramani, 2016) which is used to develop a pseudo-label selection criterion under multiple domain shifts in the SSDG problem.

Semi-Supervised Domain Generalization. The problem setting in DG assumes fully-supervised settings i.e., the source domains data is completely labelled. In many real-world deployment scenarios, however, this is a strict requirement, as it is costly and some times infeasible to acquire sufficiently labelled data. To address this limitation, a more practical and widely applicable setting is semi-supervised domain generalization (SSDG), which combines model generalization and data efficiency into a single paradigm. For instance, (Lin et al., 2021) introduced a cyclic learning framework to enhance model generalization by promoting a positive feedback between the pseudo-labelling and generalization phases. The authors in (Zhou et al., 2021a) proposed StyleMatch as an effective approach that ex-
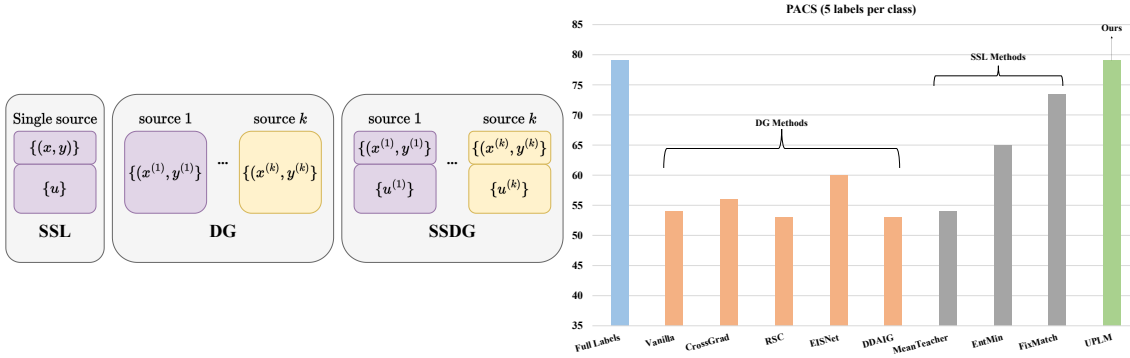
Figure 1: (Left) Visual comparison of SSL, DG, and SSDG setting. (right) Performance comparison of three paradigms.

tends FixMatch with stochastic modeling and multi-view consistency learning to achieve significant improvements in SSDG problem. (Yao et al., 2022) proposed confidence-aware cross pseudo supervision algorithm that utilizes Fourier transformation and image augmentation to enhance the quality of PLs for SSDG medical image segmentation. (Qi et al., 2022) proposed MultiMatch, a SSDG method that extends FixMatch to the multi-task learning framework, utilizing the theory of multi-domain learning to produce high-quality PLs.

3 PROPOSED FRAMEWORK

The problem of semi-supervised domain generalization (SSDG) has two distinct challenges: (1) how to obtain accurate pseudo-labels under multiple domain shifts, and (2) reduce overfitting to source domains under limited labels. To this end, we present a principled approach to SSDG namely uncertainty-guided pseudo-labelling (Section 3.1) with model averaging (Section 3.2) to counter the two challenges in SSDG.

Problem Settings. We first define few notations and then present the formal definition of SSDG. Formally, $\mathcal{X}$ and $\mathcal{Y}$ denote the input and label spaces, respectively. A domain is a combination of the (joint) probability distributions for X and Y, denoted by $P(X,Y)$, over the corresponding spaces $\mathcal{X}$ and $\mathcal{Y}$. We use $P(X)$ and $P(Y)$ to show the marginal distributions of X and Y, respectively. Our focus in this study is on distribution shifts only in $P(X)$, while $P(Y)$ remains constant. This means that all domains share the same label space. Similar to DG, in SSDG, we are provided with K distinct but related source domains $\mathcal{D} = \{\mathcal{D}\}_{k=1}^{K}$, where $\mathcal{D}_k$ denotes the distribution over the input space $\mathcal{X}$ for domain k, and K is the total number of source domains. From each source domain $\mathcal{D}_k$, we are provided with a labelled set comprising of input-label pairs $\mathcal{D}_k^L = \{(x^k, y^k)\}$ and an un-

labelled set $\mathcal{D}_k^U = \{u^k\}$. Note that, $|\mathcal{D}_k^U| \gg |\mathcal{D}_k^L|$. We also assume the existence of a set of target domains $\mathcal{T}$ typically set to 1. The objective in SSDG is to leverage the labelled set $\mathcal{D}_k^L$ from the source domains, along with the unlabelled data $\mathcal{D}_k^U$, to learn a mapping $\mathcal{F}_\theta : \mathcal{D}_k^L \cup \mathcal{D}_k^U \rightarrow \mathcal{Y}$ that can provide accurate predictions on data from an unseen target domain $\mathcal{T}$.

Semi-Supervised DG Pipeline. We instantiate our proposed method in FixMatch, which is an SSL method and performs better than all DG methods in SSDG settings (Figure 1). It combines consistency regularization (Sajjadi et al., 2016) and pseudo-labelling (PL) (Xie et al., 2020) techniques to achieve state-of-the-art results on several SSL benchmarks. The algorithm consists of two standard cross entropy losses, the supervised loss $\mathcal{L}_s$, and an unlabelled loss $\mathcal{L}_u$. The supervised loss is calculated as: $\mathcal{L}_s = -\frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} y_j \log(\hat{y}_j)$, where $\mathcal{S} = \mathcal{D}_{k=1:K}^S$ is the aggregation of labelled set from all K source domains. y_j and $\hat{y}_j$ is the ground truth and the predicted probability for j^{th} labelled example, respectively.

Two augmented versions of the an unlabelled example u are generated i.e., weak and strong augmentations (DeVries and Taylor, 2017) denoted by u' and u'', respectively. Let $q_{u'}$ and $q_{u''}$ be the predicted probability distributions for u' and u'', respectively. For a weakly augmented unlabelled example u', the pseudo-label $\tilde{y}_u'$ is generated if $g_{u'}$ is 1, where $g_{u'}$ is a binary variable and obtained as follows: $g_{u'} = \mathbb{1}[\max(q_{u'}) \geq \tau]$, τ is a scalar hyperparameter denoting the confidence threshold. The cross entropy (CE) loss is used at the model output for a strongly augmented version u'', which introduces the form of consistency regularization to reduce the discrepancy between u' and u''. The unsupervised loss $\mathcal{L}_u$ becomes: $\mathcal{L}_u = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{1}[\max(q_{u'}) \geq \tau] \, \mathrm{CE}(\tilde{y}_u', q_{u''})$, where $\mathcal{U} = \mathcal{D}_{k=1:K}^U$ is the aggregated unlabelled set from all K source domains. The overall loss then becomes: $\mathcal{L}_{final} = \mathcal{L}_s + \lambda \mathcal{L}_u$, where λ is the weight given to the unsupervised loss. The presence of un-

labelled data in different source domains, manifesting different shifts, pose a challenge to confidence-based selection of PLs leading to generation of noisy PLs. Due to various domain shifts, the model is more prone to generating a high confidence for an incorrect prediction, which will then translate into a noisy pseudo-label. To this end, we leverage model's predictive uncertainty to develop pseudo-label selection criterion which allows countering the poor calibration of model, leading to the selection of accurate PLs.

3.1 Uncertainty-Guided Pseudo-Labelling (UPL)

We describe our uncertainty-guided pseudo-labelling (UPL) mechanism to address the challenge of noisy PLs when the unlabelled data could be from different (source) domains. We first quantify the model's predictive uncertainty and then leverage it to construct a uncertainty-guided pseudo-label selection criterion.

Uncertainty Quantification. We choose to use the Monte-Carlo (MC) dropout method to quantify model's predictive output uncertainty $\mathcal{V}_{u'}$ for an unlabelled example u'. It requires the addition of a single dropout layer ($\mathcal{D}$) that is incorporated between the feature extractor network and the classifier. The MC dropout technique requires $\mathcal{N}$ Monte-Carlo forward passes for an unlabelled example u' through the model. This produces a distribution of probability outputs denoted as $\mathbf{c}_{u'} \in \mathbb{R}^{\mathcal{N} \times C}$ where C is the number of classes. Now, we obtain the uncertainty $\mathcal{V}_{u'} \in \mathbb{R}^C$ by computing the variance along the first dimension of $\mathbf{c}_{u'}$. Finally, the $\mathcal{V}_{u'}$ is transformed using tanh function, to obtain a measure of model certainty $\kappa_{u'}$: $\kappa_{u'} = (1 - \tanh(\mathcal{V}_{u'}))$.

Uncertainty Constraint in PL Selection. In SSDG, due to domain shifts, for an unlabelled input, a model can yield high confidence for an incorrect prediction. This happens because the model is typically poorly calibrated for out-domain predictions. So a confidence-based PL selection criterion is prone to generating noisy PLs. To implicitly mitigate the impact of poor calibration of the model under various domain shifts, motivated by (Rizve et al.,), we develop a pseudo-label selection criterion that uses both the predictive confidence and predictive uncertainty of a model. Specifically, for an unlabelled weakly augmented example u', given the confidence of the predicted class label as: $\max(q_{u'})$ and the corresponding certainty as $\kappa_{u'}(\arg\max(q_{u'}))$. The $\max(q_{u'})$ should be greater than the confidence threshold τ and at the same time the $\kappa_{u'}(\arg\max(q_{u'}))$ should be greater than the certainty threshold η as:

$$g_{u'} = \mathbb{1}\left[\max(q_{u'}) \geq \tau\right]\mathbb{1}\left[\kappa_{u'}(\arg\max(q_{u'})) \geq \eta\right] \quad (1)$$

In Figure 2, we plot the relationship between the model output uncertainty and its Expected Calibration Error (ECE) (see also Appendix A). It shows that in all cases when the uncertainty of selected PLs increases, the ECE increases and vice versa. Therefore, choosing PLs that are both certain and confident will likely lead to better PL accuracy via counteracting the negative effects of poor calibration.

3.2 Model Averaging (MA)

In the training stage, the model may overfit to the limited labelled (or pseudo-labelled) data and eventually perform poorly on unseen target domain data. This problem exacerbates when we introduce hard constraints on PLs selection. Consequently the robustness of a model against domains shifts gets affected and could lead to convergence at poor optimum. To address this, we propose a simple yet effective model averaging (MA) technique at the inference stage. Specifically, we take the weighted average of the model parameters obtained from the best performing model on on held-out validation set (θ_{best}), the model checkpoint from last epoch (θ_{last}), and the exponential moving average model (θ_{ema}). The predictions of the three models are averaged using the combined state dictionary, which is created by taking the average of the corresponding weights of the three models denoted by θ_{avg} given as: $\theta_{avg} = \alpha \cdot \theta_{best} + \beta \cdot \theta_{last} + \gamma \cdot \theta_{ema}$ where α, β, and γ are the weights assigned to each model. We set α, β, and γ to 1/3 each, indicating that we give equal importance to each model. θ_{avg} is then used to make predictions on the test data. By using θ_{avg} model, we reduce the reliance on a single model and its parameters, which leads to better generalization at inference stage.

4 EXPERIMENTS

Datasets. We evaluate on four distinct DG datasets: PACS (Li et al., 2017) (9,991 images, 7 classes, four domains), OfficeHome (Venkateswara et al., 2017) (15,588 images, 65 classes, four domains), TerraIncognita (Beery et al., 2018) (24,778 images, 10 classes, four domains), and VLCS (Fang et al., 2013) (10,729 images, 5 classes, four domains).

Training and Implementation Details. We follow the evaluation protocol of (Gulrajani and Lopez-Paz, 2020). For model selection we use the training domain validation protocol. We partition the data from each training domain in 90% training and 10% validation subsets and use only 10 labels per class from each source domain. The model that maximizes the ac-
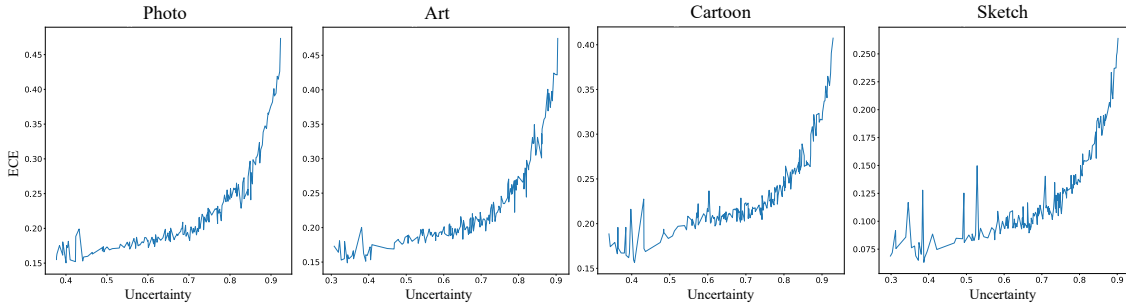
Figure 2: Uncertainty of selected PLs vs Expected Calibration Error (ECE).

curacy on validation set is considered the best model which is then evaluated on the target domain to report classification (top-1) accuracy. All experiments use an NVIDIA Quadro RTX 6000 GPU with 24GB dedicated memory. We use ResNet-50 (He et al., 2016a) model as a backbone with a batch size B of 24 for labelled data and $\mu \times B$ for unlabelled data where $\mu = 5$. We use the SGD (Robbins and Monro, 1951) optimizer and train model for 20 epochs (512 iterations each). The learning rate is set to 0.03 with nesterov momentum (Nesterov, 1983). We do grid search in the range $\{0.2, 0.9\}$ using the validation set for the hyperparameter η in UPL method. The optimal η values are 0.2, 0.5, 0.5 and 0.7 for PACS, TerraIncognita, OfficeHome and VLCS respectively. We report accuracy for target domains and their average, where a model is trained on source domains and evaluated on an (unseen) target domain. Each accuracy on the target domain is an average over three different trials with different labelled examples. Appendix B shows the ablation on different hyperparameters including the number of MC forward passes $\mathcal{N}$, certainty threshold κ and the parameter μ which governs the proportion of unlabelled data within each training batch.

4.1 Results

We investigate the impact of each proposed component for all four datasets. Table 1 presents a comparison of test accuracies achieved by four different methods: FixMatch (baseline), uncertainty-guided PL approach (UPL), model averaging (MA), and our final model (UPLM) across different target domains of four benchmark datasets. The average test accuracy across all target domains is also shown for each dataset. The results demonstrate that the UPLM method achieves the highest test accuracy in three out of four datasets, with an average test accuracy of 78.94% for PACS, 50.61% for OfficeHome, 62.72% for VLCS, and 30.19% for TerraIncognita. On OfficeHome the constraint of uncertainty limits the number of PLs, and hence relatively less improvement is

seen in UPLM as compared to MA. For instance, the class to training examples ratio for VLCS and Office-Home is 1:2146 and 1:238 respectively. Enforcing an uncertainty constraint on this small set of examples reduces their number even further, making the ResNet-50 model more prone to overfitting. Overall, the UPLM method outperforms in the most target domains across all datasets, indicating that the uncertainty-guided PL approach with model averaging leads to improved performance in SSDG.

Table 1: Comparison of FixMatch, UPL, MA, and UPLM.

	Target	FixMatch	UPL	MA	UPLM (Ours)
PACS	Photo	$82.67_{\pm 4.73}$	$89.76_{\pm 3.12}$	$90.40_{\pm 1.24}$	$88.09_{\pm 1.92}$
	Art	$70.79_{\pm 0.88}$	$72.75_{\pm 5.50}$	$76.53_{\pm 1.08}$	$76.84_{\pm 1.02}$
	Cartoon	$70.39_{\pm 3.21}$	$66.87_{\pm 3.46}$	$75.78_{\pm 1.54}$	$74.05_{\pm 5.25}$
	Sketch	$70.19_{\pm 5.00}$	$74.63_{\pm 3.98}$	$71.43_{\pm 3.60}$	$76.79_{\pm 3.38}$
	Average	$73.51_{\pm 2.19}$	$76.35_{\pm 3.41}$	$78.54_{\pm 1.44}$	$\mathbf{78.94_{\pm 1.49}}$
OfficeHome	Art	$38.64_{\pm 3.14}$	$39.37_{\pm 5.09}$	$43.52_{\pm 0.82}$	$42.47_{\pm 0.66}$
	Clipart	$39.28_{\pm 4.05}$	$41.69_{\pm 3.32}$	$41.76_{\pm 0.90}$	$40.58_{\pm 1.94}$
	Product	$58.73_{\pm 1.48}$	$58.10_{\pm 2.36}$	$59.41_{\pm 0.62}$	$58.00_{\pm 1.15}$
	Real World	$56.88_{\pm 2.22}$	$60.87_{\pm 0.65}$	$63.91_{\pm 0.98}$	$61.37_{\pm 1.47}$
	Average	$48.38_{\pm 0.51}$	$50.00_{\pm 0.50}$	$\mathbf{52.15_{\pm 0.59}}$	$50.61_{\pm 1.23}$
VLCS	Caltech101	$43.37_{\pm 32.44}$	$74.08_{\pm 12.60}$	$36.42_{\pm 1.10}$	$85.68_{\pm 3.65}$
	LabelMe	$52.78_{\pm 1.91}$	$59.23_{\pm 6.71}$	$51.49_{\pm 0.72}$	$61.09_{\pm 4.98}$
	SUN09	$49.88_{\pm 1.61}$	$42.96_{\pm 6.19}$	$62.60_{\pm 3.90}$	$50.41_{\pm 5.93}$
	VOC2007	$27.26_{\pm 1.98}$	$41.02_{\pm 12.00}$	$41.87_{\pm 4.70}$	$53.68_{\pm 7.61}$
	Average	$43.32_{\pm 9.13}$	$54.33_{\pm 5.14}$	$48.10_{\pm 1.81}$	$\mathbf{62.72_{\pm 3.66}}$
Terra	Location 38	$15.00_{\pm 13.52}$	$22.14_{\pm 7.50}$	$28.59_{\pm 7.10}$	$32.32_{\pm 18.06}$
	Location 43	$14.07_{\pm 2.46}$	$14.07_{\pm 1.55}$	$17.88_{\pm 7.10}$	$25.82_{\pm 5.94}$
	Location 46	$19.04_{\pm 3.18}$	$21.15_{\pm 4.51}$	$21.77_{\pm 3.18}$	$24.22_{\pm 3.59}$
	Location 100	$22.14_{\pm 16.59}$	$25.23_{\pm 1.99}$	$40.97_{\pm 4.81}$	$38.38_{\pm 9.65}$
	Average	$17.56_{\pm 2.24}$	$20.07_{\pm 2.92}$	$27.30_{\pm 3.38}$	$\mathbf{30.19_{\pm 4.78}}$

Table 2: Comparison of FixMatch, StyleMatch and UPLM on labelled seed examples from (Zhou et al., 2021a).

Target	Baseline (FixMatch)	StyleMatch	UPLM
Photo	$89.18_{\pm 0.30}$	$78.20_{\pm 11.30}$	$\mathbf{91.82_{\pm 1.15}}$
Art	$73.85_{\pm 3.77}$	$78.10_{\pm 1.31}$	$\mathbf{79.05_{\pm 1.74}}$
Cartoon	$74.73_{\pm 3.72}$	$\mathbf{82.02_{\pm 1.11}}$	$78.37_{\pm 2.08}$
Sketch	$74.74_{\pm 5.65}$	$78.60_{\pm 1.87}$	$\mathbf{79.06_{\pm 0.52}}$
Avg.	$78.12_{\pm 1.35}$	$76.60_{\pm 2.77}$	$\mathbf{82.02_{\pm 1.11}}$

Furthermore, we conducted a thorough comparative analysis, using the labelled seed examples of StyleMatch (Zhou et al., 2021a) (Table 2). Considering factors like unavailable source code ((Yuan et al., 2022)) and relatively large batch sizes(StyleMatch) the comparison of SSDG methods becomes difficult.

We optimized our model by adjusting the batch size to 24 and using the ResNet-50 backbone instead of ResNet-18 (He et al., 2016b). These modifications were essential for enhancing both performance and computational efficiency. Notably, in comparison with StyleMatch, our method demonstrated superior performance, particularly in the photo domain, using our randomly chosen seeds (available on our GitHub project page) providing a practical and accessible alternative to the examples employed by StyleMatch.

4.2 Ablation Study and Analysis

t-SNE Plots For Class-Wise Features. Figure 3 plots class-wise feature representations obtained using t-SNE for both the FixMatch and UPLM. Our approach facilitates the learning of more discriminative features, resulting in more tightly clustered features within the same class while maintaining greater distance between features belonging to different classes.
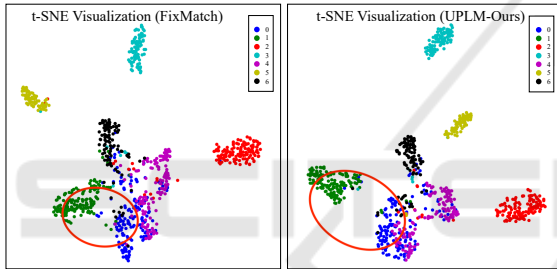

Figure 3: Class-wise feature visualization using t-SNE.

Pseudo-Labelling Accuracy UPL vs FixMatch. We compare the accuracy of PLs on the target domains of the PACS dataset (Table 3). Results indicate that UPL generates more accurate PLs compared to FixMatch.

Table 3: Comparison of Pseudo-Labelling accuracy (%).

	Target	FixMatch	UPL
	Photo	87.09	88.05
	Art	78.80	95.93
PLs Acc.	Cartoon	83.93	89.52
	Sketch	91.55	95.30
	Average	85.34	**92.20**

Performance of Individual Components of MA. Table 4 compares the results of six different variants of the model, with each variant utilizing a different strategy for combining the model's parameters during training. Combining all three models (θ_{avg}), as per our proposal, provides the best performance. **Perfor-**

Table 4: Our proposed θ_{avg} outperforms other variants.

	Target	θ_{last}	θ_{best}	θ_{ema}	$\theta_{(last+ema)}$	$\theta_{(last+best)}$	$\theta_{(best+ema)}$	θ_{avg}
	Photo	87.64	87.72	82.67	89.90	89.30	89.08	**90.40**
	Art	72.98	73.93	70.79	**78.24**	73.11	72.38	76.53
MA (PACS)	Cartoon	73.93	69.16	70.39	75.43	**75.90**	73.08	75.78
	Sketch	68.72	72.23	70.19	65.48	**75.31**	68.90	71.43
	Average	75.82	75.76	73.51	77.26	78.41	75.86	**78.54**

mance Under Various Domain Shifts. We report the performance in various domain shifts in Table 5, e.g., changes in backgrounds, corruptions, textures, and styles. For instance, background shifts only affect the background of an image and not the foreground object's pixel, texture, and structure (Zhang et al., 2022). On the other hand, style shifts involve variations in texture, and object parts across different concepts. To evaluate this, we categorize four DG datasets i.e., PACS, VLCS, OfficeHome, and TerraIncognita based on their exhibited shift(s) into the four categories and report results. UPLM outperforms all other methods in all domain shifts, except for a slight advantage of MA in style. The OfficeHome dataset has a limited number of examples per class, leading to potential overfitting due to uncertainty constraints. However, our MA approach demonstrates strong performance by effectively mitigating mis-calibrated PLs.

Table 5: Accuracy (%) for different types of domain shifts.

Methods	Texture Shifts PACS	Corruption Shifts Terra	Background Shifts VLCS, Terra	Style Shifts OH, PACS
FixMatch	73.51	17.56	30.44	60.94
UPL	76.35	20.07	37.20	63.17
MA	78.54	27.30	37.70	**65.34**
UPLM (Ours)	**78.94**	**30.19**	**46.46**	64.78

5 CONCLUSION

We presented a new SSDG approach (UPLM) featuring uncertainty-guided pseudo-labelling and model averaging mechanisms. The proposed approach leverages the model's predictive uncertainty to develop a pseudo-labelling selection criterion that mitigates the impact of poor model calibration under multi-source unlabelled data. The model averaging technique reduces overfitting to source domains in the presence of limited labels and domain shifts. Results on several challenging DG datasets suggest that our method provides notable gains over the baseline. We believe that our work will encourage the development of more data-efficient visual recognition models that are also generalizable across different domains.

REFERENCES

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473.

DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Dou, Q., Coelho de Castro, D., Kamnitsas, K., and Glocker, B. (2019). Domain generalization via model-agnostic

learning of semantic features. *Advances in Neural Information Processing Systems*, 32.

Fang, C., Xu, Y., and Rockmore, D. N. (2013). 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.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. (2015). Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559.

Grandvalet, Y. and Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.

Gulrajani, I. and Lopez-Paz, D. (2020). In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Hoffman, J., Tzeng, E., Darrell, T., and Saenko, K. (2017). Simultaneous deep transfer across domains and tasks. *Domain Adaptation in Computer Vision Applications*, pages 173–187.

Huang, Z., Wang, H., Xing, E. P., and Huang, D. (2020). Self-challenging improves cross-domain generalization.

Kendall, A. and Cipolla, R. (2016). Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, pages 4762–4769. IEEE.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Khan, A., AlBarri, S., and Manzoor, M. A. (2022a). Contrastive self-supervised learning: a survey on different architectures. In *2022 2nd International Conference on Artificial Intelligence (ICAI)*, pages 1–6. IEEE.

Khan, A., Khattak, M. U., and Dawoud, K. (2022b). Object detection in aerial images : A case study on performance improvement. In *2022 International Conference on Artificial Intelligence of Things (ICAIoT)*, pages 1–9.

Khan, M. H., Zaidi, T., Khan, S., and Khan, F. S. (2021). Mode-guided feature augmentation for domain generalization. In *Proc. Brit. Mach. Vis. Conf.*

Kim, D., Yoo, Y., Park, S., Kim, J., and Lee, J. (2021). Self-reg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 9619–9628.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. (2017). Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550.

Lin, L., Xie, H., Yang, Z., Sun, Z., Liu, W., Yu, Y., Chen, W., Yang, S., and Xie, D. (2021). Semi-supervised domain generalization in real world: New benchmark and strong baseline. *arXiv preprint arXiv:2111.10221*.

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR.

Nesterov, Y. E. (1983). A method of solving a convex programming problem with convergence rate o\bigl(k^2\bigr). In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences.

Qi, L., Yang, H., Shi, Y., and Geng, X. (2022). Multimatch: Multi-task learning for semi-supervised domain generalization. *arXiv preprint arXiv:2208.05853*.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29.

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., and Sarawagi, S. (2018). Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*.

Shu, Y., Cao, Z., Wang, C., Wang, J., and Long, M. (2021). Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633.

Smith, L. and Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

Advances in neural information processing systems, 30.

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pages 4068–4076.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. (2018). Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31.

Wang, S., Yu, L., Li, C., Fu, C.-W., and Heng, P.-A. (2020). Learning from extrinsic and intrinsic supervisions for domain generalization.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698.

YANG, P. Y. and Gao, W. (2013). Multi-view discriminant transfer learning.

Yao, H., Hu, X., and Li, X. (2022). Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3099–3107.

Yuan, J., Ma, X., Chen, D., Kuang, K., Wu, F., and Lin, L. (2022). Label-efficient domain generalization via collaborative exploration and generalization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2361–2370.

Zhang, C., Zhang, M., Zhang, S., Jin, D., Zhou, Q., Cai, Z., Zhao, H., Liu, X., and Liu, Z. (2022). Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7277–7286.

Zhou, K., Loy, C. C., and Liu, Z. (2021a). Semi-supervised domain generalization with stochastic stylematch. *arXiv preprint arXiv:2106.00592*.

Zhou, K., Yang, Y., Qiao, Y., and Xiang, T. (2021b). Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*.

APPENDIX

A Uncertainty κ vs ECE Plot

For each training iteration, we first calculate the mean of uncertainty ($\mathcal{V}$) for all input examples in a batch across class dimension to obtain the overall uncertainty for each example. Also, we compute the corresponding ECE score for this batch. Next, after each epoch, we compute mean of overall uncertainty (for each example) over all examples seen and also compute the mean over ECE. Each epoch yields a pair of mean uncertainty and corresponding mean ECE over all examples. We then sort the mean uncertainty values in ascending order to build x-axis and the corresponding mean (ECE) to plot y-axis.

B Analysis of Hyperparameters

Table 6: Ablation of κ in range [0.2, 0.8].

	Target	0.2	0.3	0.4	0.5	0.6	0.7	0.8
PACS	Photo	87.07	86.11	86.77	78.32	84.79	70.42	65.09
	Art	75.73	71.58	68.70	69.73	68.02	57.13	60.40
	Cartoon	68.09	66.3	65.02	61.56	63.14	59.68	59.90
	Sketch	79.38	72.13	73.12	59.96	71.21	46.17	66.45
	Average	**77.57**	74.03	73.40	67.39	71.79	58.35	62.96
OfficeHome	Art	42.89	37.99	42.69	43.22	38.90	42.89	40.34
	Clipart	40.92	42.15	37.55	42.50	39.31	36.70	41.05
	Product	57.92	58.95	56.25	58.71	56.32	57.02	57.02
	Real World	63.76	59.44	60.80	62.98	61.65	58.30	59.93
	Average	51.37	49.63	49.32	**51.85**	49.05	48.73	49.59
VLCS	Caltech101	29.40	57.03	50.46	85.72	80.71	87.42	80.42
	LabelMe	54.14	53.77	55.76	60.69	54.18	64.72	60.47
	SUN09	65.42	60.02	56.79	52.16	51.22	43.57	49.18
	VOC2007	33.56	48.13	35.55	36.58	38.00	51.18	39.19
	Average	45.63	54.74	49.64	58.79	56.03	**61.72**	57.32
Terra	Location 38	22.09	12.26	4.18	39.54	44.48	11.54	33.57
	Location 43	10.96	22.80	15.19	25.84	16.73	20.10	12.95
	Location 46	23.97	19.46	23.98	20.57	16.66	22.49	22.11
	Location 100	39.72	48.20	32.40	37.63	41.95	30.65	32.38
	Average	24.19	25.68	18.94	**30.90**	29.96	21.20	25.25

B.1 Computational Cost of MC Forward Passes

We use $\mathcal{N} = 10$ Monte Carlo (MC) forward passes in all experiments, with negligible computational overhead. The per-iteration execution times in milliseconds for different values of $\mathcal{N}$ (1, 5, 10, 20, 40, 80, 160) are 134.6, 135.1, 135.6, 137.5, 138.5, 141.7, 146.6, respectively.

B.2 Accuracy with Changing the Amount of Unlabelled Data μ

The average accuracies on PACS for $\mu = [1-6]$ are 65.25, 71.90, 75.47, 73.7, **78.94**, 78.22, respectively. $\mu = 5$ performs best overall, which is used throughout in all our experiments. Note that, μ values beyond 6 are not possible due to computational constraints.

B.3 Effect of Certainty Threshold

We present an ablation study concerning the selection of the certainty threshold $\kappa = 0.2$ (indicating the least certainty) to $\kappa = 0.8$ (indicating the highest certainty), as detailed in Table 6.