Information Maximization for Long-Tailed Semi-Supervised Domain Generalization

Leo Fillioux[1*], Omprakash Chakraborty[2], Quentin Gopée[1], Pierre Marza[1], Paul-Henry Cournède[1], Stergios Christodoulidis[1], Maria Vakalopoulou[1], Ismail Ben Ayed[2] and Jose Dolz[2]

[1]Université Paris-Saclay, CentraleSupélec, Gustave Roussy, INSERM, CDSU, IHU PRISM, Gif-sur-Yvette
[2]LIVIA, ILLS, ETS Montréal
Corresponding author: leo.fillioux@centralesupelec.fr

Abstract. Semi-supervised domain generalization (SSDG) has recently emerged as an appealing alternative to tackle domain generalization when labeled data is scarce but unlabeled samples across domains are abundant. In this work, we identify an important limitation that hampers the deployment of state-of-the-art methods on more challenging but practical scenarios. In particular, state-of-the-art SSDG severely suffers in the presence of long-tailed class distributions, an arguably common situation in real-world settings. To alleviate this limitation, we propose **IMAX**, a simple yet effective objective based on the well-known InfoMax principle adapted to the SSDG scenario, where the Mutual Information (MI) between the learned features and latent labels is maximized, constrained by the supervision from the labeled samples. Our formulation integrates an α-entropic objective, which mitigates the class-balance bias encoded in the standard marginal entropy term of the MI, thereby better handling arbitrary class distributions. IMAX can be seamlessly plugged into recent state-of-the-art SSDG, consistently enhancing their performance, as demonstrated empirically across two different image modalities.

Keywords: Domain generalization · Semi-supervised learning.

1 Introduction

Deep learning approaches have dominated the computer vision field in the last decade, with unprecedented performances in a plethora of visual recognition problems [19]. However, these models are typically trained under the assumption that training and testing samples follow the same distribution. When distributional drifts violate this *i.i.d.* (independent and identically distributed) assumption, their performance can degrade substantially [12, 26].

To address this problem, Domain Generalization (DG) aims to learn models that remain robust to unseen target domains [20, 24, 36]. In particular, in DG scenarios we assume that labeled data from multiple source domains is available,

[*] Work done during an internship at ILLS.

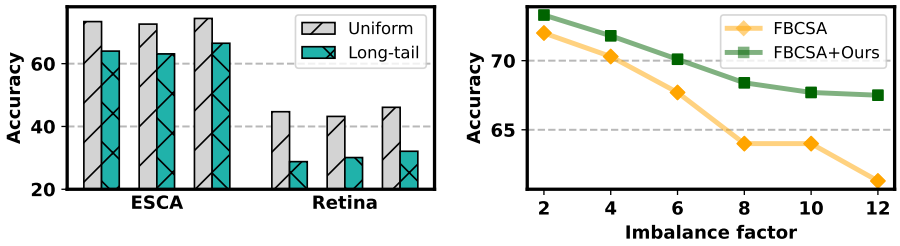

Fig. 1. Performance degradation under class imbalance. Impact on the accuracy of having long-tailed class distribution in the training data using the FBCSA [10] trainer (*left*). Impact of the imbalance factor γ (see Section 4.1) on the performance of FBCSA and our proposed approach (*right*).

and the primary objective is to train a model that can adequately generalize to novel unseen domains. However, in real-world scenarios, such as healthcare, where labeling samples for all the domains may become impractical, the assumption that the data from multiple source domains are fully labeled is unrealistic, and generally unmet. Thus, to reduce the burden of the annotation process, semi-supervised domain generalization (SSDG) has been recently explored [37, 10], which unifies the problems of semi-supervised learning and domain generalization. In this setting, each source domain has access to a few labeled samples and a larger amount of unlabeled data. Thus, beyond striving for cross-domain generalization, the model should achieve this objective efficiently with respect to labeled data requirements.

Very recently, [10] exposed that several existing semi-supervised learning (SSL) approaches outperformed methods tailored to DG in the SSDG scenario. Based on these observations, authors proposed two plug-and-play strategies (i.e., FBCSA [10] and DGWM [9]) that can be integrated into different SSL methods, setting a novel state-of-the-art for this task. Nevertheless, after exploring the proposed frameworks we identified an important limitation that hampers their application in real-life scenarios. In particular, these approaches assume uniform class distributions across the different source domains. However, in real-world applications, many tasks involve inherently imbalanced data, e.g., rare diseases in medical imaging. Indeed, we empirically observed that if we assume *long-tailed* class distributions, the performance of both FBCSA [10] and DGWM [9] is substantially degraded across several tasks (Fig. 1), underscoring the importance of considering this more realistic scenario.

In light of the observation presented above, in this paper we investigate the problem of domain generalization in the context of semi-supervised learning with class-imbalanced datasets, which combines the problems of domain generalization and label-efficient learning in the presence of class imbalance, a more realistic scenario than prior works [10, 9]. Thus, our contributions can be summarized as:

- We introduce a more realistic SSDG setting where, in addition to the standard labeled and unlabeled data across multiple source domains, labeled samples are exposed to imbalanced class distributions.
- We present **IMAX**, an information theoretic approach to tackle this challenging yet realistic scenario. In particular, the proposed solution accommodates the specificities of this challenging setting by deriving a semi-supervised view of the standard Mutual Information (MI).
- To further account for class imbalance, we replace the less flexible marginal entropy term in the modified MI by a learning objective derived from Tsallis divergences [32], which tolerates better variations on class distributions.
- Experiments across several datasets, SSL and SSDG approaches demonstrate that IMAX consistently enhances existing methods in the studied scenario, showcasing its model-agnosticity and *plug-and-play* versatility.

2 Related Work

Semi-supervised Domain Generalization (SSDG) focuses on improving generalization across domains when labeled data is scarce but substantial unlabeled data is accessible. Within the broader computer vision community, state-of-the-art SSDG [10, 9] leverages semi-supervised learning. These methods consist of generating pseudo-labels from weakly augmented samples, and training the model to enforce that predictions from strongly augmented versions align with these pseudo-labels. In particular, FixMatch [28], FreeMatch [34], and StyleMatch [37] are typically employed as SSL frameworks in SSDG [10, 9]. The SSDG setting is particularly well suited to medical imaging, where acquisition conditions can vary widely. In this scenario, different approaches have been developed specifically for medical imaging problems [35, 30, 29]. However, these techniques typically rely on restrictive, modality-dependent assumptions, making them applicable only to a subset of medical imaging modalities. For example, [35] assumes that the distribution shift results in a modification of the hue saturation and contrast, while [29] assumes that structural similarity between classes is preserved across domains. Both conditions do not hold consistently across diverse and real clinical settings, e.g., anatomical variability and pathology-dependent morphology can substantially alter structural patterns.

Mutual information (MI) in deep learning. Our approach is grounded in the well-established *InfoMax* principle [21], which advocates maximizing the MI between a system's inputs and outputs. Several variants of this general principle, which accommodate the specific requirements of different scenarios, have been recently used in machine learning and computer vision tasks, including deep clustering [14, 17, 23], few-shot learning [2, 3, 11], representation learning [1, 13, 16], or generalized category discovery [5]. To the best of our knowledge, however, addressing SSDG from an information-theoretic standpoint remains unexplored.

Class imbalance. While there is active research on learning under class imbalance [22, 6], these methods often fail to achieve expected performance in a long-tailed SSL setting, particularly when labeled and unlabeled distributions mismatch, and therefore cannot be applied to our setting.

3 Methodology

Problem definition and notation. We now formally define the semi-supervised domain generalization (SSDG) problem. Let $\mathcal{X}$ and $\mathcal{Y}$ define the input and label space, respectively. A *domain* is defined by a joint probability distribution P_{XY} over $\mathcal{X} \times \mathcal{Y}$. In this scenario, we only consider distribution shifts in P_X, while assuming P_Y remains the same, i.e., all domains share the same label space. We have access to D distinct but related domain sources $\mathcal{S} = \{S_d\}_{d=1}^{D}$, each associated with a joint distribution $P_{XY}^{(d)}$. In SSDG, only a small portion of the source data has labels, while many instances are unlabeled, for which we can only access the empirical marginal distribution P_X. For each source domain S_d, the labeled subset is defined as $S_L^{(d)} = \{(\mathbf{x}_i^{(d)}, \mathbf{y}_i^{(d)})\}_{i=1}^{m_L^{(d)}}$, where $(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}) \sim P_{XY}^{(d)}$, and the unlabeled subset as $S_U^{(d)} = \{\mathbf{x}_i^{(d)}\}_{i=1}^{m_U^{(d)}}$, with $\mathbf{x}^{(d)} \sim P_X^{(d)}$, which denotes the marginal distribution of $P_{XY}^{(d)}$ over $\mathcal{X}$. The size of the unlabeled dataset is typically much larger than that of the labeled data, i.e., $m_U^{(d)} = |S_U^{(d)}| \gg |S_L^{(d)}| = m_L^{(d)}$.

The goal in SSDG is to learn a domain-generalizable model, parameterized by $\boldsymbol{\theta}$, using both labeled and unlabeled source data, which outputs a softmax prediction vector $\mathbf{p}_i = (p_{ik})_{1 \leq k \leq K}$, with K denoting the number of classes. At test time, the model is directly deployed in an unseen target domain $T = \{\mathbf{x}^*\}$ with $\mathbf{x}^* \sim P_X^*$. The target domain differs from any source domain used for training, i.e., $P_X^* \neq P_X^{(d)}$ for all $d \in \{1, \ldots, D\}$.

Mutual information. The mutual information (MI) between the labels and the input images can be defined as follows:

$$I(Y; X) = \mathcal{H}(Y) - \mathcal{H}(Y|X), \tag{1}$$

where $\mathcal{H}(Y)$ denotes the entropy of the marginal distributions $\pi_k = \mathbb{P}(Y = k; \boldsymbol{\theta})$, and $\mathcal{H}(Y|X)$ refers to the entropy of the conditional probability distribution $\mathbb{P}(Y|X; \boldsymbol{\theta})$, *i.e.*, the remaining uncertainty in Y when X is known. The marginal distribution can be approximated by the soft proportion of samples within each class $k \in \{1, \ldots K\}$ across both the set of all labeled ($\mathcal{D}_L$) and unlabeled ($\mathcal{D}_U$) samples $\mathcal{D} = \{\mathcal{D}_L, \mathcal{D}_U\}$, as $\pi_k \approx \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} p_{ik}$.

Our method. While it is common in the literature to maximize the unsupervised mutual information in Eq. (1) (e.g., often defined over unlabeled samples [3, 17]), we accommodate this term to our current scenario, where a few labeled samples are available. More concretely, we integrate explicit supervision constraints on the conditional probabilities $\mathbf{p}_i$ of the samples within the labeled set. The proposed constrained information maximization can be formally defined as:

$$\max_{\boldsymbol{\theta}} \mathcal{H}(Y) - \mathcal{H}(Y|X) \quad \text{s.t.} \quad \mathbf{y}_i = \mathbf{p}_i \quad \forall \mathbf{x}_i \in \mathcal{D}_L \tag{2}$$

Integrating the equality constraints above into the MI, the objective becomes:

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} \pi_k \log \pi_k - \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \sum_{k=1}^{K} q_{ik} \log p_{ik}, \tag{3}$$

with $q_{ik} = y_{ik}$ if $\mathbf{x}_i \in \mathcal{D}_L$ and $q_{ik} = p_{ik}$ otherwise. Further decomposing the equation above, we can obtain the following objective:

$$\min_{\boldsymbol{\theta}} \quad \sum_{k=1}^{K} \pi_k \log \pi_k - \frac{1}{|\mathcal{D}_L|} \sum_{i \in \mathcal{D}_L} \sum_{k=1}^{K} y_{ik} \log p_{ik} - \frac{1}{|\mathcal{D}_U|} \sum_{i \in \mathcal{D}_U} \sum_{k=1}^{K} p_{ik} \log p_{ik} \quad (4)$$

Semi-Supervised Mutual Information. The last term in Eq. (4) represents the conditional Shannon entropy, and is an unsupervised objective that encourages implicitly confident predictions, as it pushes them toward one vertex of the simplex. Instead of simply optimizing the third objective, i.e., minimizing the entropy of the predictions, we leverage the concepts of *consistency regularization* and *pseudo-labeling*, widely used in the semi-supervised learning literature [28, 34, 37]. More concretely, all unlabeled images undergo an augmentation step, where weak and strong transformations of each image are performed. Then, pseudo-labels derived from the predictions on weak image transformations are used to guide the predictions of their strongly augmented counterparts [28]. Formally, given a *weakly*-augmented version of unlabeled image i, we obtain its softmax prediction vector $\mathbf{p}_i$, from which we can compute its pseudo-label $\hat{y}_i = \arg\max(\mathbf{p}_i)$[1]. Thus, for image i, we can define the pseudo cross-entropy as:

$$-\sum_{k=1}^{K} \mathbb{1}(\max(\mathbf{p}_i) \geq \tau) \hat{y}_{ik} \log p_{ik}^{\mathcal{A}}, \quad (5)$$

where $\mathbf{p}_i^{\mathcal{A}}$ denotes the softmax prediction vector of the strongly augmented image, i.e., $\mathbf{p}_i^{\mathcal{A}} = f_{\boldsymbol{\theta}}(\mathcal{A}(\mathbf{x}_i))$, and τ a scalar hyperparameter denoting the threshold above which a pseudo-label should be considered. Plugging this term into Eq. (4), we have the following SSL view of the MI:

$$\min_{\boldsymbol{\theta}} \quad \underbrace{\sum_{k=1}^{K} \pi_k \log \pi_k}_{-\mathcal{H}(Y)\text{: marginal entropy}} \quad \underbrace{-\frac{1}{|\mathcal{D}_L|} \sum_{i \in \mathcal{D}_L} \sum_{k=1}^{K} y_{ik} \log p_{ik}}_{\mathcal{H}(Y|X_L)\text{: cross-entropy}} \quad \underbrace{-\frac{1}{|\mathcal{D}_U|} \sum_{i \in \mathcal{D}_U} \sum_{k=1}^{K} \mathbb{I}(\max(\mathbf{p}_i) \geq \tau) \hat{y}_{ik} \log p_{ik}^{\mathcal{A}}}_{\mathcal{H}(\hat{Y}|X_U)\text{: pseudo cross-entropy}} \quad (6)$$

Adapting to imbalanced scenarios. While the marginal entropy term $\mathcal{H}(Y)$ in Eq. (6) is of paramount importance to prevent trivial arbitrary solutions [3], it pushes the marginal distribution towards the uniform, thereby encoding a strong bias towards balanced partitions in the labeled set. Nevertheless, as shown in our empirical observations (Fig. 1), and stated in our motivations, assuming uniform

[1] Pseudo-label $\hat{y}_i$ is transformed into a one-hot encoded vector $\hat{\mathbf{y}}_i \in \{0, 1\}^K$ to be used in (5) and (6).

class partitions is unrealistic in real-world problems. Inspired by [33], we propose a generalization of the semi-supervised mutual-information loss in Eq. (6), based on α-divergences, which can better accommodate variations in class distributions. In particular, we focus on Tsallis' [32] formulation of the α-divergences. Indeed, if we write the marginal entropy as an explicit Kullback-Leibler (KL) divergence ($\mathcal{H}(Y) = \log(K) - \mathcal{D}_{KL}(\mathbf{p}||\mathbf{u}_K)$, where $\mathbf{u}_K$ is the probability distribution for the uniform distribution with K classes), it generalizes to Tsallis α-entropy:

$$\mathcal{H}_\alpha(\mathbf{p}) = \log_\alpha(K) - K^{1-\alpha}\mathcal{D}_\alpha(\mathbf{p}\|\mathbf{u}_K) = \frac{1}{\alpha - 1}\left(1 - \sum_k p_k^\alpha\right) \tag{7}$$

A derivation of the above equation is provided in the Appendix of [33].
Final objective. Our final objective is then formally defined as:

$$\min_{\boldsymbol{\theta}} \quad -\mathcal{H}_\alpha(Y) + \mathcal{H}(Y|X_L) + \mathcal{H}(\hat{Y}|X_U), \tag{8}$$

where $\mathcal{H}_\alpha(Y)$ is the α-entropy defined in Eq. (7). The role of each term in the objective presented above is:

- Standard cross-entropy on labeled samples $\mathcal{H}(Y|X_L)$ could be seen as the *penalty* function employed to enforce the constraints $\mathbf{y}_i = \mathbf{p}_i, \forall \mathbf{x}_i \in \mathcal{D}_L$, thereby avoiding the need to explicitly enforce the equality constraints in Eq. (2). Thus, this term basically encourages that, for labeled samples, network predictions align with their corresponding class labels.
- Pseudo cross-entropy for unlabeled samples $\mathcal{H}(\hat{Y}|X_U)$ acts as a pseudo supervisory signal on unlabeled samples, following standard SSL practices [28, 34, 37]. It replaces the conditional entropy $\mathcal{H}(Y|X)$, as its optima may easily lead to degenerate solutions, mapping all samples to a single arbitrary class. Using pseudo-labels, as well as a mechanism to filter only confident ones, prevents these degenerate solutions, while also avoiding the special care required during optimization of $\mathcal{H}(Y|X)$ [3, 8].
- $\mathcal{H}_\alpha(Y)$ acts as a label marginal regularization, offering a more relaxed version of the standard label marginal entropy in MI, which encourages a near-uniform class distribution. By relaxing the strictly uniform constraint, $\mathcal{H}_\alpha(Y)$ tolerates class distributions further from the uniform, which makes the proposed solution more suitable for imbalanced scenarios.

4 Experiments

4.1 Experimental setup

Datasets. We explore two different tasks. **Histology:** ESCA [31], a histopathology patch-level classification dataset with images from 11 classes across four different hospitals, each treated as a separate domain. **Ophthalmology:** we focus on diabetic retinopathy (DR) grading, with five different grades (no DR,

mild DR, moderate DR, severe DR, and proliferative DR). We use four different datasets, each simulating a separate domain: Messidor-2 [7, 18], IDRiD [25], Paraguay [4], and APTOS [15]. Labels are standardized following [27].

Baselines. Since our objective is to design an approach that does not rely on impractical assumptions, and remains broadly applicable, we focus our comparisons on *assumption-free* SSDG methods, such as FBCSA [10] and DGWM [9]. Furthermore, following the recent SSDG literature [10, 9], we leverage three common SSL methods: FixMatch [28], FreeMatch [34], and StyleMatch [37].

Implementation details. In each task, each experiment consists of 20 runs, corresponding to four setups (train on three domains and test on the fourth) and five seeds for each setup, reporting the average across runs. We train for 20 epochs using SGD. Based on the validation set, we empirically set $\alpha = 1.5$ (ESCA), and $\alpha = 2$ (Retina). *Enforcing imbalance:* Given K classes, considering m_L labeled samples per class in each domain, we take $m_L \cdot K$ samples per domain, and enforce an exponentially decaying number of samples per class, shuffling the order of the classes differently for each seed. The hyperparameter γ (set to $\gamma = 10$) tunes the decay, and therefore the imbalance. The imbalance is imposed on the labeled set, while the unlabeled set maintains its original distribution.

4.2 Results

Main results. In this section we evaluate the performance of IMAX when coupled with existing SSDG methods. In addition to the existing FBCSA [10] and DGWM [9], we refer to *"Baseline"* as the direct use of SSL methods, without any SSDG-specific component. Table 1 reports the average accuracy for the ESCA and Retina datasets considering $m_L \in \{5, 10\}$ per class and per domain, across three different frameworks and three SSL methods. We observe that **IMAX consistently improves performance across all but one setting**. Notably, improvements are most pronounced in the low-label regime (up to $+$**7.3%** when $m_L = 5$ *vs.* $+$**5.8%** with $m_L = 10$), highlighting the effectiveness of our approach in scenarios where annotated data is scarce. These results underscore the practical relevance of IMAX, as it enhances generalization of existing SSDG methods under real-world constraints, such as data-scarcity and long-tailed distributions.

Sensitivity to varying distributions (Fig. 1). We now assess robustness under increasingly extreme long-tailed distributions, revealing how quickly existing SSDG methods deteriorate as imbalance grows. Fig. 1 (*right*) depicts the results on the ESCA dataset with the FBCSA strategy, with and without applying the proposed IMAX. The results show that, while both models see their performance decrease as data imbalance increases, **degradation without IMAX is significantly more severe**, particularly for more extreme imbalance.

Impact of the proposed learning objectives (Tab. 2). Integrating the proposed semi-supervised MI objective (Eq. 6) into the baseline already yields substantial gains ($+$**5.0%** and $+$**3.4%** for $m_L \in \{5, 10\}$). We note that, although this formulation improves over the baseline, it remains a rigid case of our α-based objective, since Eq. 6 is a particular case of the more general objective in Eq. 8,

Table 1. Performance in the long-tailed SSDG scenario. We report [standard / +**IMAX**], where +**IMAX** denotes the results obtained by our approach when applied on each SSDG method and across the three SSL strategies.

		ESCA		Retina	
	Method	$m_L = 5$	$m_L = 10$	$m_L = 5$	$m_L = 10$
Baseline	FixMatch/+**IMAX**	61.0 / **68.3**$_{+7.3}$	69.2 / **73.1**$_{+3.9}$	31.4 / **38.4**$_{+7.0}$	34.4 / **40.2**$_{+5.8}$
	FreeMatch/+**IMAX**	59.3 / **63.2**$_{+3.9}$	67.5 / **71.0**$_{+3.5}$	36.0 / **38.7**$_{+2.7}$	34.0 / **37.0**$_{+3.0}$
	StyleMatch/+**IMAX**	60.4 / **67.7**$_{+7.3}$	69.0 / **73.4**$_{+4.4}$	30.9 / **36.0**$_{+5.1}$	34.6 / **38.4**$_{+3.8}$
FBCSA	FixMatch/+**IMAX**	64.0 / **67.7**$_{+3.7}$	72.2 / **73.2**$_{+1.0}$	28.8 / **32.3**$_{+3.5}$	**32.0** / 29.5$_{-2.5}$
	FreeMatch/+**IMAX**	63.1 / **65.5**$_{+2.4}$	71.6 / **72.4**$_{+0.8}$	30.1 / **31.1**$_{+1.0}$	27.9 / **29.8**$_{+1.9}$
	StyleMatch/+**IMAX**	66.5 / **69.9**$_{+3.4}$	73.3 / **75.3**$_{+2.0}$	32.1 / **35.6**$_{+3.5}$	30.0 / **32.0**$_{+2.0}$
DGWM	FixMatch/+**IMAX**	63.2 / **68.8**$_{+5.6}$	71.5 / **74.3**$_{+2.8}$	26.1 / **31.8**$_{+5.7}$	31.4 / **33.2**$_{+1.8}$
	FreeMatch/+**IMAX**	62.2 / **66.1**$_{+3.9}$	71.1 / **72.6**$_{+1.5}$	30.6 / **30.8**$_{+0.2}$	32.3 / **33.5**$_{+1.2}$
	StyleMatch/+**IMAX**	63.9 / **69.3**$_{+5.4}$	71.7 / **73.8**$_{+2.1}$	27.5 / **31.3**$_{+3.8}$	31.8 / **33.7**$_{+1.9}$

when $\alpha = 1$. Relaxing this term with the more flexible $\mathcal{H}_\alpha$ term yields further gains (+**5.0%** and +**3.4%**), demonstrating the effectiveness of IMAX.

Table 2. Ablation on the proposed learning objectives, where SSL baseline can be defined as $\mathcal{H}(Y|X_L) + \mathcal{H}(\hat{Y}|X_U)$. Comparison on the ESCA dataset.

Loss	$m_L = 5$	$m_L = 10$
SSL baseline	61.0	69.2
SSL baseline + $(-\mathcal{H}(Y))$ (Eq. 6)	66.0$_{+5.0}$	72.6$_{+3.4}$
SSL baseline + $(-\mathcal{H}_\alpha(Y))$ (Eq. 8) = **IMAX**	**68.3**$_{+7.3}$	**73.1**$_{+3.9}$

Sensitivity to α. Fig. 2 empirically validates our choices for α. Indeed, while performance varies with α, the validation and test accuracies exhibit highly consistent trends. This alignment indicates that the value selected on the validation set transfers reliably to the test domain, ensuring that α can be tuned in a stable and predictable manner.

5 Conclusion

We introduce the long-tailed SSDG setting, a realistic scenario which has proven to be particularly challenging for current SSDG methods. To accommodate to the challenges of long-tailed SSDG, we propose IMAX, which leverages the InfoMax principle, and introduces a specific learning objective that acounts for class imbalance. Furthermore, IMAX is model-agnostic and plug-and-play, enabling seamless integration with SoTA SSDG frameworks based on SSL, a popular strategy in the literature, showing consistent improvement across settings.

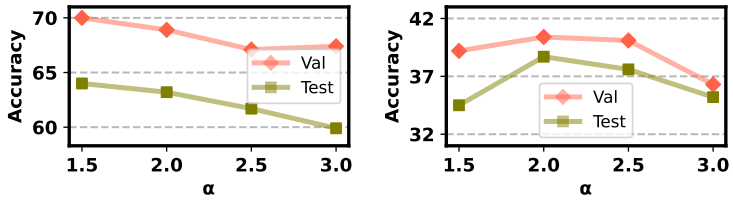

Fig. 2. Ablation on the α parameter. Impact on the α parameter of $\mathcal{H}_\alpha$ on the accuracy for the ESCA (*left*) Retina (*right*) datasets under long-tailed distribution. Results are for FreeMatch on the SSL Baseline setting.

Acknowledgements. This work has benefited from state financial aid, managed by the Agence Nationale de Recherche under the investment program integrated into France 2030, project reference ANR-21-RHUS-0003. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011014802R1 made by GENCI. The author gratefully acknowledges support from ILLS during the internship at ETS Montreal.
Disclosure of Interests. The authors declare no competing interests.

References

1. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. Advances in neural information processing systems **32** (2019)
2. Boudiaf, M., Kervadec, H., Masud, Z.I., et al.: Few-shot segmentation without meta-learning: A good transductive inference is all you need? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
3. Boudiaf, M., Ziko, I., Rony, J., et al.: Information maximization for few-shot learning. Advances in Neural Information Processing Systems **33**, 2445–2457 (2020)
4. Castillo Benítez, V.E., Castro Matto, I., Mello Román, J.C., et al.: Dataset from fundus images for the study of diabetic retinopathy. Data in Brief **36**, 107068 (2021)
5. Chiaroni, F., Dolz, J., Masud, Z.I., et al.: Parametric information maximization for generalized category discovery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1729–1739 (2023)
6. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR (2019)
7. Decencière, E., Zhang, X., Cazuguel, G.: Feedback on a publicly distributed image database: The messidor database. Image Analysis and Stereology **33**, 231–234 (2014)
8. Dhillon, G.S., Chaudhari, P., Ravichandran, A., et al.: A baseline for few-shot image classification. In: International Conference on Learning Representations (2020)
9. Galappaththige, C.J., Izzo, Z., He, X., et al.: Domain-guided weight modulation for semi-supervised domain generalization. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6495–6505 (2025)

10. Galappaththige, C.J., Baliah, S., Gunawardhana, M., et al.: Towards generalizing to unseen domains with few labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23691–23700 (2024)
11. Hajimiri, S., Boudiaf, M., Ayed, I.B., et al.: A strong baseline for generalized few-shot semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023)
12. Hendrycks, D., Zhao, K., Basart, S., et al.: Natural adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
13. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., et al.: Learning deep representations by mutual information estimation and maximization (2019)
14. Jabi, M., Pedersoli, M., Mitiche, A., et al.: Deep clustering: On the link between discriminative models and k-means. IEEE transactions on pattern analysis and machine intelligence **43**(6), 1887–1896 (2019)
15. Karthik, M., Dane, S.: APTOS 2019 Blindness Detection (2019)
16. Kemertas, M., Pishdad, L., Derpanis, K.G., et al.: RankMI: A mutual information maximizing ranking loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14362–14371 (2020)
17. Krause, A., Perona, P., Gomes, R.: Discriminative clustering by regularized information maximization. Advances in neural information processing systems **23** (2010)
18. Krause, J., Gulshan, V., Rahimy, E., et al.: Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology **125**, 1264–1272 (8 2018)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
20. Li, H., Wang, Y., Wan, R., et al.: Domain generalization for medical imaging classification with linear-dependency regularization. Advances in neural information processing systems **33**, 3118–3129 (2020)
21. Linsker, R.: Self-organization in a perceptual network. Computer **21**(3), 105–117 (1988). https://doi.org/10.1109/2.36
22. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
23. Ohl, L., Mattei, P.A., Bouveyron, C.: Generalised mutual information for discriminative clustering. Advances in Neural Information Processing Systems **35** (2022)
24. Ouyang, C., Chen, C., Li, S., et al.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging **42**(4), 1095–1106 (2022)
25. Porwal, P., Pachade, S., Kokare, M., et al.: Idrid: Diabetic retinopathy – segmentation and grading challenge. Medical Image Analysis **59**, 101561 (1 2020)
26. Recht, B., Roelofs, R., Schmidt, L., et al.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400 (2019)
27. Silva-Rodríguez, J., Chakor, H., Kobbi, R., et al.: A foundation language-image model of the retina (flair): encoding expert knowledge in text supervision. Medical Image Analysis **99**, 103357 (2025)
28. Sohn, K., Berthelot, D., Carlini, N., et al.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)

29. Song, J., Chen, H., Qin, J., et al.: Dual-supervised asymmetric co-training for semi-supervised medical domain generalization. IEEE Transactions on Multimedia (2025)
30. Tissir, Z., Chang, Y., Lee, S.W.: Style-aware and uncertainty-guided approach to semi-supervised domain generalization in medical imaging. Mathematics **13**(17), 2763 (2025)
31. Tolkach, Y., Wolgast, L.M., Damanakis, A.I., , et al.: Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. The Lancet Digital Health (2023)
32. Tsallis, C.: Possible Generalization of Boltzmann-Gibbs Statistics. J. Statist. Phys. **52**, 479–487 (1988). https://doi.org/10.1007/BF01016429
33. Veilleux, O., Boudiaf, M., Piantanida, P., et al.: Realistic evaluation of transductive few-shot learning. Advances in Neural Information Processing Systems **34** (2021)
34. Wang, Y., Chen, H., Heng, Q., et al.: Freematch: Self-adaptive thresholding for semi-supervised learning. In: International Conference on Learning Representations (ICLR) (2023)
35. Zhang, R., Xu, Q., Huang, C., et al.: Semi-supervised domain generalization for medical image analysis. In: IEEE 19th International Symposium on Biomedical Imaging (ISBI) (2022)
36. Zhou, K., Liu, Z., Qiao, Y., et al.: Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4396–4415 (2022)
37. Zhou, K., Loy, C.C., Liu, Z., et al.: Semi-supervised domain generalization with stochastic stylematch. International Journal of Computer Vision **131**(9), 2377–2387 (2023)