Rethinking Multi-domain Generalization with A General Learning Objective

Zhaorui Tan[1,2], Xi Yang*[1], Kaizhu Huang*[3]
[1]Xi'an Jiaotong-Liverpool University, [2]University of Liverpool, [3]Duke Kunshan University
Zhaorui.Tan21@student.xjtlu.edu.cn, Xi.Yang01@xjtlu.edu.cn, kaizhu.huang@dukekunshan.edu.cn

Abstract

Multi-domain generalization (mDG) is universally aimed to minimize the discrepancy between training and testing distributions to enhance marginal-to-label distribution mapping. However, existing mDG literature lacks a general learning objective paradigm and often imposes constraints on static target marginal distributions. In this paper, we propose to leverage a $\mathbf{Y}$-mapping to relax the constraint. We rethink the learning objective for mDG and design a new ***general learning objective*** *to interpret and analyze most existing mDG wisdom. This general objective is bifurcated into two synergistic amis: learning domain-independent conditional features and maximizing a posterior. Explorations also extend to two effective regularization terms that incorporate prior information and suppress invalid causality, alleviating the issues that come with relaxed constraints. We theoretically contribute an upper bound for the domain alignment of domain-independent conditional features, disclosing that many previous mDG endeavors actually* ***optimize partially the objective*** *and thus lead to limited performance. As such, our study distills a general learning objective into four practical components, providing a general, robust, and flexible mechanism to handle complex domain shifts. Extensive empirical results indicate that the proposed objective with $\mathbf{Y}$-mapping leads to substantially better mDG performance in various downstream tasks, including regression, segmentation, and classification. Code is available at https://github.com/zhaorui-tan/GMDG/tree/main.*

1. Introduction

Domain shift, which breaks the independent and identical distributed (*i.i.d.*) assumption amid training and test distributions [51], poses a common yet challenging problem in real-world scenarios. Multi-domain generalization (mDG) [3]) is garnering increasing attention owing to its promising capacity to utilize multiple distinct but related source domains for model optimization, ultimately intending to generalize well to unseen domains. Intrinsically, the primary objective for mDG is the maximization of the joint distribution between observations $\mathbf{X}$ and targets $\mathbf{Y}$ across all domains $\mathcal{D}$:

$$\begin{aligned}\max P(\mathbf{X},\mathbf{Y}\mid\mathcal{D}) =& P(\mathbf{Y}\mid\mathcal{D})P(\mathbf{X}\mid\mathbf{Y},\mathcal{D})\\ =& P(\mathbf{X}\mid\mathcal{D})P(\mathbf{Y}\mid\mathbf{X},\mathcal{D}).\end{aligned} \tag{1}$$

A prevalent approach initiates by maximizing the marginal distribution $P(\mathbf{X}|\mathcal{D})$ before presuming an invariant $P(\mathbf{Y}|\mathbf{X}) = P(\mathbf{Y}|\mathbf{X},\mathcal{D})$ across domains [58], anchored on an assumption that $P(\mathbf{Y}|\mathcal{D})$ remains consistency across $\mathcal{D}$.

Is $P(\mathbf{Y}|\mathcal{D})$ truly static across domains? In other words, does $\mathbf{Y}$ truly lack domain-dependent features? In classification tasks, typically, the influence of $\mathcal{D}$ on $\mathbf{Y}$ is substantially marginal. However, this assumption is not universally applicable, particularly in tasks such as regression or segmentation. Consequently, MDA [16] relaxes the assumption of stable $P(\mathbf{Y}|\mathcal{D})$ by providing an average class discrepancy, allowing both $P(\mathbf{X}|\mathbf{Y},\mathcal{D})$ and $P(\mathbf{Y}|\mathcal{D})$ vary across $\mathcal{D}$. However, MDA has to conduct class-specific sample selection under domains for obtaining $P(\mathbf{X}|\mathbf{Y},\mathcal{D})$,

	Aim1: Learning domain invariance
Others	None
DANN	$\min_\phi H(P(\phi(\mathbf{X})\mid\mathcal{D}))$
CDANN, CIDG, MDA	$\min_\phi H(P(\phi(\mathbf{X}),\mathbf{Y}\mid\mathcal{D}))$
Ours: GAim1	$\min_{\phi,\psi} H(P(\phi(\mathbf{X}),\psi(\mathbf{Y})\mid\mathcal{D}))$
	Reg1: Integrating prior
Others	None
MIRO, SIMPLE	$\min_\phi D_{\mathrm{KL}}(P(\phi(\mathbf{X}),\mathbf{Y})\|\mathcal{O})$
Ours: GReg1	$\min_{\phi,\psi} D_{\mathrm{KL}}(P(\phi(\mathbf{X}),\psi(\mathbf{Y}))\|\mathcal{O})$
	Aim2: Maximizing A Posterior (MAP)
Others	$\min_\phi H(P(\mathbf{Y},\phi(\mathbf{X})))$
Ours: GAim2	$\min_{\phi,\psi} H(P(\mathbf{Y},\phi(\mathbf{X}))) + H(P(\mathbf{Y},\psi(\mathbf{Y})))$
	Reg2: Suppressing invalid causality
Others	None
CORAL	$\min_\phi -H(P(\phi(\mathbf{X})\mid\mathcal{D})) + H(P(\phi(\mathbf{X})))$
MDA,RobutsNet	$\min_\phi -H(P(\phi(\mathbf{X})\mid\mathbf{Y})) + H(P(\phi(\mathbf{X})))$
Ours: GReg2	$\min_{\phi,\psi} -H(P(\phi(\mathbf{X})\mid\psi(\mathbf{Y}))) + H(P(\phi(\mathbf{X})))$

Table 1. A summary of objectives of ERM [14], DANN [13], CORAL [48], CDANN [29], CIDG [28], MDA [16], MIRO [19], SIMPLE [30], RobustNet [10], VA-DepthNet [32] and Ours. All constants are omitted here. 'Others' denotes no other specified methods. For more details, see Supplementary Material 7.

*Corresponding authors

Symbols	Descriptions
$d_n \in \mathcal{D}, n \leq N$; $d' \in \mathcal{D}$:	The n-th observed domains in all domains; Unseen domains in all domains.
$\mathbf{X}, \mathbf{Y}$; $\mathbf{X}_n, \mathbf{Y}_n$; $\mathbf{X}', \mathbf{Y}'$.	All observations and targets; Observations and targets in d_n; Observations and targets in d'.
$P(x)$:	Distributions where x corresponds to the random variables.
ϕ, ψ:	Learnable transformations that codify $\mathbf{X}, \mathbf{Y}$ into the same latent RKHS.
$\phi(\mathbf{X}), \psi(\mathbf{Y})$:	Mapped $\mathbf{X}, \mathbf{Y}$. Within the RKHS realm, $\phi(\mathbf{X}), \psi(\mathbf{Y})$ follow Multivariate Gaussian Distributions.
$\mathcal{O}$; $R(\cdot)$; $\sigma_{\cdot,\cdot}$:	Prior knowledge (oracle model); Empirical risks; Covariance between two variables.
$\mathcal{C} : \phi(\mathbf{X}), \psi(\mathbf{Y}) \to \mathbf{Y}$:	Predictor that predicts $\mathbf{Y}$ from $\phi(\mathbf{X}), \psi(\mathbf{Y})$.
$D_{\mathrm{KL}}(\cdot\|\cdot)$; $H(\cdot)$; $H_e(\cdot,\cdot)$:	KL divergence; Entropy; Cross-entropy.

Table 2. A summary of notations.

which constrains its objective's universality and struggles with tasks beyond basic classification, especially where $\mathbf{Y}$ is not discrete.

To better tackle the $\mathcal{D}$-dependent variations in both $\mathbf{X}$ and $\mathbf{Y}$ for border tasks beside classification, we introduce two learnable mappings, ϕ and ψ, that project $\mathbf{X}$ and $\mathbf{Y}$ into the **same** latent Reproducing Kernel Hilbert Space (RKHS), assumed to extract $\mathcal{D}$-independent features from $\mathbf{X}, \mathbf{Y}$. Incorporating these, Eq. 1 can be changed as

$$\max_{\phi,\psi} P(\phi(\mathbf{X}), \psi(\mathbf{Y})), \quad \text{s.t., } \phi(\mathbf{X}), \psi(\mathbf{Y}) \perp\!\!\!\perp \mathcal{D}. \tag{2}$$

Built upon the optimization of Eq. 2, we further identify two additional issues that warrant consideration. 1). The synergy of *integrating prior information* and domain-invariant feature learning plays a crucial role. pre-trained (oracle) models can be used as priors [19, 30] to regulate feature learning. 2). Issues regarding *invalid causality predicament* within the $P(\mathbf{Y}|\mathcal{D})$ static assumption relaxation during learning the invariance come to light. This is aligned with the causality premise $\phi(\mathbf{X}) \to \psi(\mathbf{Y})$ to maximize $P(\phi(\mathbf{X}), \psi(\mathbf{Y}) \mid \mathcal{D})$. Efforts must be made to suppress invalid causality $\psi(\mathbf{Y}) \to \phi(\mathbf{X})$ during invariant-feature learning (Refer to Eq. 8 for derivation).

Considering these findings, the general objective for mDG, which copes with the above issues and effectively relaxes the static target distribution assumption, is crucial. To be specific, it should consist of **four key parts**: **Aim1**- Learning domain-invariant representations and **Aim2**- Maximizing the posterior; with two regularization **Reg1**- Integrating prior information and the **Reg2**- Suppression of invalid causality. In essence, the objective should certify invariant representations of $\mathbf{X}, \mathbf{Y}$ across domains while preserving the prediction relationship in $\mathbf{X} \to \mathbf{Y}$. As a notable contribution, we redesign the conventional mDG paradigm and uniformly simplify most previous works' empirical objectives, as summarized in Table 1 while Notations are shown in Table 2.

Most current mDG studies only focus on classification. SOTA methods such as MIRO [19] and SIMPLE [30] propose learning similar features by "oracle" models as a substitute for learning domain-invariant representations for mDG. Worth mentioning, we counter MIRO's argument by confirming the persisting necessity of domain-invariant features, even under prior distribution, by theoretically deviating from minimizing the Generalized Jensen-Shannon Divergence (GJSD). MDA [16] pioneered the relaxation of the $P(\mathbf{Y}|\mathcal{D})$ static assumption, yet without explicitly introducing a $\mathbf{Y}$-mapping function, and overlooked the emergence of invalid causality that arises upon the relaxation. Beyond classification, RobustNet [10] and VA-DepthNet [32] explore their methods on mDG settings in segmentation and regression but propose no explicit objective for mDG. Importantly, our theoretical analysis and empirical findings suggest that mere aggregation of all the aforementioned objectives fails to yield a comprehensive general objective for mDG. For instance, term $-H(P(\phi(\mathbf{X}|\mathcal{D})))$, coupled with prior knowledge utilization, could inadvertently precipitate performance degradation.

In this paper, we introduce the General Multi-Domain Generalization Objective (GMDG) to overcome current limitations in current methods, relaxing the static assumption of $P(\mathbf{Y}|\mathcal{D})$ (overall formulation is shown in Section 3). Meanwhile, we propose an actionable solution to the invalidated causality through the minimization of the Conditional Feature Shift (CFS). Our main contributions can be summarized as follows:

- We theoretically prove that domain generalization can be improved through the minimization of Generalized Jensen-Shannon Divergence (GJSD), with the incorporation of prior knowledge, leading to the derivation of an alignment upper bound (PUB) (Section 3).
- We analyze existing approaches, demonstrating their incomplete optimization against the GMDG and identifying unexpected terms they inadvertently introduce (Section 4).
- Our approach is the first try that is designed as compatible with existing mDG frameworks and exhibits performance improvements in a suite of tasks, including regression, segmentation, and classification, as confirmed by our comprehensive experiments (Section 5).

Notably, our results that only used one pre-trained model as prior in classification tasks exceed the SOTA SIMPLE++, which employs 283 pre-trained models as an ensemble oracle, while yielding consistent improvement in regression and segmentation, as shown in Figure 1. This further suggests the superiority of GMDG.

2. Related work

Multi-domain generalization. Most current mDG methods focus only on classification tasks. To learn better D-independent representations for mDG, DANN [13] minimizes feature divergences between the source domains. CDANN [29], CIDG [28], and MDA [16] additionally take conditions into consideration and aim to learn conditionally invariant features across domains. [5, 7, 19, 30] point out

that learning invariant representation to source domains is insufficient for mDG. Thus, MIRO [19] and SIMPLE [19] adopt pre-trained models as an oracle for seeking better general representations across various domains, including unseen target domains. Meanwhile, RobustNet [10] constrains conditional covariance shifts and conducts mDG segmentation, and little exploration focuses on mDG regression, though many other works pay attention to single domain generalization [20, 24, 32, 37]. Our study shows that their objectives optimize partially GMDG, leading to sub-optimal results.

Multi-domain generalization assumptions. In the literature, different assumptions are proposed to simplify the task as described by the original objective in Eq. 1. One assumption is that the $P(\mathbf{Y}|\mathbf{X},\mathcal{D})$ is stable while only marginal $P(\mathbf{X}|\mathcal{D})$ changes across domains [45, 55]. [54] point out that $\mathbf{X}$ is usually caused by $\mathbf{Y}$ thus $P(\mathbf{Y}|\mathcal{D})$ changes while $P(\mathbf{X}|\mathbf{Y},\mathcal{D})$ is sable or $P(\mathbf{X}|\mathbf{Y},\mathcal{D})$ changes but $P(\mathbf{Y}|\mathcal{D})$ stays stable, or a combination of both. Thus, MDA [16] allows both $P(\mathbf{Y}|\mathbf{X},\mathcal{D})$ and $P(\mathbf{X}|\mathcal{D})$ change across domains but needs selecting samples of each class for the calculation. Moreover, it considers no prior. This paper further relaxes these assumptions by extracting domain-invariant features in $\mathbf{X},\mathbf{Y}$.

Using pre-trained models as an oracle. Previous methods such as MIRO [19] have employed pre-trained models as the oracle to regularize ϕ. SIMPLE [30] employs at most 283 pre-trained models as an ensemble and adaptively composes the most suitable oracle model. RobustNet [10] and VA-DepthNet [32], only use pre-trained models as initialization rather than additional supervision.

3. A general multi-Domain generalization objective

General Multi-Domain Generalization Objective (GMDG) essentially comprises a weighted combination of **Four** terms, each term designated by an alias:

$$\begin{aligned}\min_{\phi,\psi}\ & \underbrace{v_{A1}H(P(\phi(\mathbf{X}),\psi(\mathbf{Y})\mid\mathcal{D}))}_{\textbf{GAim1}} \\ & + \underbrace{v_{A2}[H(P(\psi(\mathbf{Y}),\phi(\mathbf{X}))) + H(P(\mathbf{Y},\psi(\mathbf{Y})))]}_{\textbf{GAim2}} \\ & + \underbrace{v_{R1}D_{\mathrm{KL}}(P(\phi(\mathbf{X}),\psi(\mathbf{Y}))\|\mathcal{O})}_{\textbf{GReg1}} \\ & \underbrace{-v_{R2}H(P(\phi(\mathbf{X})\mid\psi(\mathbf{Y}))) + H(P(\phi(\mathbf{X})))}_{\textbf{GReg2}}.\end{aligned} \tag{3}$$

Theoretically, we justify that **GAim1** and **GReg1** can be effectively revised by minimizing the Generalized Jensen-Shannon Divergence (GJSD) with prior knowledge between visible domains for optimization. Meanwhile, we derive an upper bound termed as an alignment Upper Bound with Prior of mDG (PUB). Importantly, we demonstrate using **GReg2** not only cope with the invalid causality brought by $P(\mathbf{Y}|\mathcal{D})$ static assumption relaxation. Regarding **GReg2**, it can be simplified by minimizing the Conditional Feature Shift (CFS), *i.e.,* the shift between unconditional and conditional features, which can be calculated by ψ. More theoretical details are provided as follows.

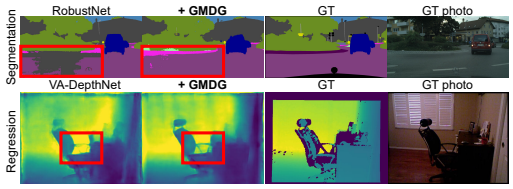

Figure 1. Segmentation and regression results of baselines and +GMDG on samples in unseen domains.

3.1. Theoretical Details

Learning of $\mathcal{D}$-independent conditional features under prior. The generalization alignment upper bound (PUB), a novel GJSD variational upper bound that is tied to domain generalization alignment, is derived based on the generalized Jensen-Shannon divergence (GJSD) [31].

Definition 1 (GJSD). *Given J distributions, $\{P(\mathbf{Z}_j)\}_{j=1}^J$ and a corresponding probability weight vector w, $GJSD_w(\{P(\mathbf{Z}_j)\}_{j=1}^J)$ is defined as:*

$$\begin{aligned}&\sum\nolimits_{j=1}^{J} w_j D_{\mathrm{KL}}(P(\mathbf{Z}_j)\|\sum\nolimits_{j=1}^{J} w_j P(\mathbf{Z}_j)) \\ \equiv & H(\sum\nolimits_{j=1}^{J} w_j P(\mathbf{Z}_j)) - \sum\nolimits_{j=1}^{J} w_j H(P(\mathbf{Z}_j)).\end{aligned} \tag{4}$$

Our method addresses the standard scenario in which the weights are evenly distributed across domains: $w_1 = ... = w_N = 1/N$. To achieve $\phi(\mathbf{X}),\psi(\mathbf{Y}) \perp\!\!\!\perp \mathcal{D}$, minimizing domain gap between $P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))$ can be converted to minimizing GJSD across all domains:

$$\begin{aligned}&\min_{\phi,\psi} GJSD(\{P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))\}_{n=1}^N) \\ \equiv & \min_{\phi,\psi} H(P(\phi(\mathbf{X}),\psi(\mathbf{Y})\mid\mathcal{D})) \\ & - \mathbb{E}[H(P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n)))].\end{aligned} \tag{5}$$

We further involve a prior knowledge distribution $\mathcal{O}$ under the consideration of a variational density model class $\mathcal{Q}$. Drawing upon [9], we have a variational upper bound:

$$\begin{aligned}&GJSD(\{P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))\}_{n=1}^N) \\ \leq & H_c(\mathbb{E}[P(\phi(\mathbf{X}),\psi(\mathbf{Y})]\mid\mathcal{D}),\mathcal{O}) - a,\end{aligned} \tag{6}$$

where $a \triangleq \sum_{n=1}^N H(P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n)))$ is constant *w.r.t* ϕ,ψ, hence ignored during optimization. The novel PUB is derived from Eq. 6, is:

$$\begin{aligned}&\min_{\phi,\psi} PUB(\{P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))\}_{n=1}^N) \\ \triangleq & \min_{\phi,\psi} P(\phi(\mathbf{X}),\psi(\mathbf{Y})\mid\mathcal{D})) \\ & + D_{\mathrm{KL}}(P(\phi(\mathbf{X}),\psi(\mathbf{Y}))\|\mathcal{O})) - a.\end{aligned} \tag{7}$$

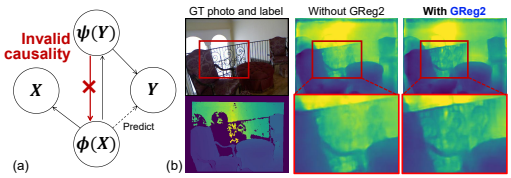

Figure 2. (a) Diagram of causality in the proposed method. (b) Depth predictions on the unseen domain sample between model trained without and with **GReg2**.

Minimizing PUB is the proposed objective for **GAim1** and **GReg1**. This implies that methods like MIRO, solely minimizing GReg1, might result in substantial suboptimality, leaving the domain gap unresolved. We discuss two situations of $\mathcal{O}$ in Section 4.

Suppressing invalid causality. The relaxation of $P(\mathbf{Y}|\mathcal{D})$ static assumption may lead to unexpected causality while learning the invariance. **GAim1** is reformed as:

$$\begin{aligned}\textbf{GAim1} =& H(P(\phi(\mathbf{X})\,|\mathcal{D}))+H(P(\psi(\mathbf{Y})\,|\,\phi(\mathbf{X}),\mathcal{D}))\\ =& H(P(\psi(\mathbf{Y})\,|\mathcal{D}))+H(P(\phi(\mathbf{X})\,|\,\psi(\mathbf{Y}),\mathcal{D})),\end{aligned} \tag{8}$$

where minimizing **GAim1** with relaxation of $P(\mathbf{Y}|\mathcal{D})$ static assumption may lead to $\psi(\mathbf{Y})\to\phi(\mathbf{X})$ since the term $H(P(\psi(\mathbf{Y})|\phi(\mathbf{X}),\mathcal{D}))$ and $\phi(\mathbf{X})\to\psi(\mathbf{Y})$ since the term $H(P(\phi(\mathbf{X})|\psi(\mathbf{Y}),\mathcal{D}))$.

Figure 2 graphically demonstrates the causal diagram under this scenario. Since the prediction relationship from $\phi(\mathbf{X})\to\mathbf{Y}$ and the casual path $\psi(\mathbf{Y})\to\mathbf{Y}$, $\phi(\mathbf{X})\to\psi(\mathbf{Y})$ should be preserved for prediction. However, the casual path $\phi(\mathbf{X})\to\psi(\mathbf{Y})$ may compromise the predication from $\phi(\mathbf{X})\to\mathbf{Y}$ when $\psi(\mathbf{Y})$ is unknown during the inference, leading to generalization degradation. This unveils that the invalid causality from $\psi(\mathbf{X})\to\phi(\mathbf{Y})$ that may happen during the learning invariance needs to be suppressed as $\max_{\phi,\psi} H(P(\phi(\mathbf{X})|\psi(\mathbf{Y})),\mathcal{D})$ while $\min_{\phi,\psi} H(P(\psi(\mathbf{Y})|\phi(\mathbf{X})),\mathcal{D})$, which can be simplified as:

$$\min_{\phi,\psi} H(P(\phi(\mathbf{X}))) - H(P(\phi(\mathbf{X}))|P(\psi(\mathbf{Y}))), \tag{9}$$

where is **GReg2**. See more mathematical details in Supplementary 7. Our experiments also unveil the phenomenon of invalid causality within invariant feature learning, where suppressing it could improve generalizability. The investigation of previous objectives also discloses that, in addressing the varying $P(\mathbf{Y}|\mathcal{D})$, constructs akin to **GReg2** are often implicitly included (see Table 1), though their efficacy was not explicitly stated. Moreover, their efficacy may be compromised due to the lack of ψ and other objective terms.

Then, we assume that $\phi(X),\psi(Y)$ in the RKSH follow Multivariate Gaussian-like Distributions which are denoted as $\mathcal{N}(\phi(\mathbf{X});\mu_{\mathbf{X}},\Sigma_{\mathbf{XX}}),\mathcal{N}(\psi(\mathbf{Y});\mu_{\mathbf{Y}},\Sigma_{\mathbf{YY}})$. $P(\phi(\mathbf{X})|\psi(\mathbf{Y}))$ follows $\mathcal{N}(\phi(\mathbf{X})|\psi(\mathbf{Y});\mu_{\mathbf{X}|\mathbf{Y}},\Sigma_{\mathbf{XX}|\mathbf{Y}})$. **GReg2** can be simplified as:

$$\begin{aligned}&H(\mathcal{N}(\phi(\mathbf{X});\mu_{\mathbf{X}},\Sigma_{\mathbf{XX}}))\\ &-H(\mathcal{N}(\phi(\mathbf{X})\mid\psi(\mathbf{Y});\mu_{\mathbf{X}|\mathbf{Y}},\Sigma_{\mathbf{XX}|\mathbf{Y}}))\\ =&\frac{1}{2}\ln(\frac{|\Sigma_{\mathbf{XX}}|}{|\Sigma_{\mathbf{XX}|\mathbf{Y}}|})\geq 0,\end{aligned} \tag{10}$$

where the inequality stands owing to the *Condition Reducing Entropy*. This implies $H(\mathcal{N}(\phi(\mathbf{X});\mu_{\mathbf{X}},\Sigma_{\mathbf{XX}}))\geq H(\mathcal{N}(\phi(\mathbf{X})\mid\psi(\mathbf{Y});\mu_{\mathbf{X}|\mathbf{Y}},\Sigma_{\mathbf{XX}|\mathbf{Y}}))$, deduced from $|\Sigma_{\mathbf{XX}}|\geq|\Sigma_{\mathbf{XX}|\mathbf{Y}}|\geq 0$, considering they are positive semi-definite. Distinct from the [18, 52] which decompose causal effects through extra networks, our method is based on transfer entropy (TE) by ensuring $TE(\phi(\mathbf{X})\to\psi(Y))\geq TE(\phi(\mathbf{X})\nrightarrow\psi(Y))$, i.e., $H(\phi(\mathbf{X})|\psi(Y))\geq H(\psi(Y)|\phi(\mathbf{X}))$. Thus, minimization of Eq. 10 occurs iff $|\Sigma_{\mathbf{XX}}|=|\Sigma_{\mathbf{XX}|\mathbf{Y}}|$, reformulating the task as $min_{\phi,\psi}|\Sigma_{\mathbf{XX}}|-|\Sigma_{\mathbf{XX}|\mathbf{Y}}|$, where $\Sigma_{\mathbf{XX}|\mathbf{Y}}=\Sigma_{\mathbf{XX}}-\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}$, per [21]. Therefore, **GReg2** is simplified as minimizing Conditional Feature Shift (CFS):

$$\min_{\phi,\psi}|\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}|. \tag{11}$$

3.2. Empirical losses derivations

This section presents the empirical losses used to implement Eq. 3. More detailed derivation can be referred to in Supplementary 7. We introduce the mapping ψ to relax the static target distribution. The implementation of ψ varies across tasks, utilizing MLPs for classification and regression, and ResNet-50 for segmentation. To promote a consistent latent space, the mapped $\psi(\mathbf{Y})$ retains the same dimension as that of $\phi(\mathbf{X})$. $\psi(\mathbf{Y})$ and $\phi(\mathbf{X})$ are separately fed into $\mathcal{C}$ for making predictions and obtaining $\mathcal{L}_{A2}$ for posterior maximization:

$$\mathcal{L}_{A2}(\mathcal{C},\phi,\psi) = H_c(\phi(\mathbf{X}),\mathbf{Y}) + H_c(\psi(\mathbf{Y}),\mathbf{Y}). \tag{12}$$

To mitigate domain shifts and learn domain invariance, we minimize cross-domain conditional feature distribution discrepancies. Specifically, the mean and variance of the joint distribution of $(\phi(\mathbf{X}),\psi(\mathbf{Y}))$ in each domain are estimated using VAE encoders. Consider n-pairs means and variance of n domains, we derive a joint Gaussian distribution expression $P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))\triangleq\mathcal{N}(\boldsymbol{x}_n,\boldsymbol{y}_n;\mu_n,\Sigma_n)$. Accordingly, we establish $\mathbb{E}[P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))]\triangleq\mathcal{N}(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}};\bar{\mu},\bar{\Sigma})$ where $\bar{\mu}=\mathbb{E}[\mu_n],\bar{\Sigma}=\mathbb{E}[\Sigma_n]$. Base on PUB in Eq. 7, we introduce $\mathcal{L}_{A1}$ to minimize the conditional feature gap across domains:

$$\mathcal{L}_{A1}(\phi) = \sum\nolimits_{i=1}^{n}(\log|\Sigma_i| + ||\bar{\mu}-\mu_i||^2_{\Sigma_i^{-1}}). \tag{13}$$

To integrate prior information, similar to MIRO, we utilize VAE encoders to capture the means and variances of

$\mathbf{X}$: $P(\phi(\mathbf{X})) \triangleq \mathcal{N}(\boldsymbol{x}; \mu_x, \Sigma_x)$ and the output features $x_{\mathcal{O}}$ form $\mathcal{O}$. Given that $\mathcal{O}$ preserves the correlation between $\mathbf{X}$ ($\phi(\mathbf{X})$) and $\mathbf{Y}$ ($\psi(\mathbf{Y})$), and is frozen during training, $\mathbf{Y}, \psi(\mathbf{Y})$ is omitted in empirical loss. We propose $\mathcal{L}_{R1}$ to minimize the divergence between features and $\mathcal{O}$:

$$\mathcal{L}_{R1}(\phi) = \log|\Sigma_x| + ||x_{\mathcal{O}} - \mu_x||^2_{\Sigma_x^{-1}}. \tag{14}$$

For suppressing the invalid causality, derived from Eq. 11, the loss is designed to minimize the CFS:

$$\mathcal{L}_{R2}(\phi) = ||\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}||_2, \tag{15}$$

where $\Sigma_{\mathbf{XY}} = \mathbb{E}[(\phi(\mathbf{X}) - \mathbb{E}[\phi(\mathbf{X})])^\top(\phi(\mathbf{Y}) - \mathbb{E}[\phi(\mathbf{Y})])]$, and a similar calculation process is done for $\Sigma_{\mathbf{YY}}$ and $\Sigma_{\mathbf{YX}}$. The final loss is a weighted combination of the above losses[1]:

$$\mathcal{L}(\mathcal{C}, \phi, \psi) = v_{A1}\mathcal{L}_{A1} + v_{A2}\mathcal{L}_{A2} + v_{R1}\mathcal{L}_{R1} + v_{R2}\mathcal{L}_{R2}. \tag{16}$$

4. Connection to previous methods

We validate our objective function's efficiency theoretically and demonstrate its connections with previous objectives, indicating that previous mDG efforts have partly optimized the proposed objective. Refer to Table 1 and Supplementary 8 for a detailed understanding of previous objectives.

Using ψ *v.s.* not using ψ. Previous works rarely employed ψ to map $\mathbf{Y}$, whereas we show its benefits for mDG tasks. Employing *Jensen's inequality*, we obtain $H(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))]) \geq H(\mathbb{E}[P(\phi(\mathbf{X}_n), \mathbf{Y}_n)])$. When other objectives remain the same, we compare the model with parameters θ^{ψ} optimized via the ψ mapping, against another model without ψ using parameters $\theta^{n\psi}$:

$$\sup R(\theta^{n\psi}) \geq \sup R(\theta^{\psi}). \tag{17}$$

The equivalence is valid only if ψ serves as a bijection, a condition prevalent in practical scenarios like classification. Thus, this mapping does not hinder model performance in classification tasks. It also implies that using $\psi(\mathbf{Y})$ can lower generalization risks after optimization, especially when $\mathbf{Y}$ contains features dependent on $\mathcal{D}$. This could potentially yield superior generalization in segmentation and regression tasks. Detailed proof can be seen in Supplementary 7.

Remark 1 (Importance of $\mathbf{Y}$ mapping ψ)**.** *Besides relaxing the static distribution assumption of $\mathbf{Y}$, ψ conveys two other notable benefits: 1). $\mathbf{X}$ and $\mathbf{Y}$ may originate from different sample spaces with distinct shapes. By applying mappings, $\psi(\mathbf{Y})$ can be adapted to the same shape as $\phi(\mathbf{X})$. In practice, concatenating $\phi(\mathbf{X})$ and $\psi(\mathbf{Y})$ is often used as input for VAE encoders to capture $P(\psi(\mathbf{Y}), \phi(\mathbf{X}))$. 2). The derivation of Eq. 11 requires the computation of covariance, which mandates that two variables occupy the same sample space - a condition fulfilled by applying $\psi(\mathbf{Y})$.*

Incorporating conditions leads to lower generalization risk on learning invariant representations. A few past works [13, 48] minimize domain gaps between features without condition consideration. Its objective for **Aim1** is:

$$\begin{aligned} H(P(\phi(\mathbf{X}) \mid \mathcal{D})) \leq & H(P(\phi(\mathbf{X}) \mid \mathcal{D})) + \\ & H(P(\psi(\mathbf{Y}) \mid \phi(\mathbf{X})), \mathcal{D}) = \textbf{GAim1}. \end{aligned} \tag{18}$$

While the other objectives are identical, we consider a model with parameters θ^{nc}, trained with $\min_{\psi} H(P(\phi(\mathbf{X}_n)))$, against another model with θ^{c} parameters, trained with **GAim1**. In this scenario, their empirical risks satisfy:

$$\sup R(\theta^{nc}) \geq \sup R(\theta^{c}). \tag{19}$$

See the mathematical details in Supplementary 7. This reveals that without condition consideration, the minimization of generalization risk is merely partial due to the overlooked risk correlated to $\mathbf{Y}$. Additional evidence supporting the importance of condition consideration is provided by CDANN [29] and CIDG [28]. Our experiments, conducted through a uniform implementation, also lend support to it.

Effect of oracle model $\mathcal{O}$. As stated by MIRO [19] and SIMPLE [30], a generalized $\mathcal{O}$ comprising both seen and unseen domains yields significant improvements. During the derivation of Eq. 7, we find that the disregard **GAim1** term in MIRO [19] and SIMPLE [30] may result in inferior outcomes to our proposed objective.

Remark 2 (Synergy of learning invariance, integrating prior knowledge and suppressing invalid causally)**.** *For readability, we have divided the overall mDG objective into four aspects despite all terms being interconnected. Specifically, as shown by PUB in Eq. 7, **GReg1** collaborating with **GAim1** brings more performance gains than the case when it is solely applied. Moreover, Eq. 8 shows that the side effect of invalid causality in **GAim1** is alleviated by combining with **GReg2**, underscoring the significance of combining learning invariance, integrating prior knowledge, and suppressing invalid causally. It also suggests that all terms are synergistic and contribute together to improved results.*

Validating our assertions via experiments, Section 5.5 ablation study finds that simple cross-domain covariance limitation (**GReg2**) cannot ensure improved results with prior knowledge.

5. Experiments

Four groups of experiments are done to validate the proposed GMDG. A toy example validates the relaxation of

[1] See detailed hyper-parameters settings in Supplementary 10.

	Affine transformations			Squared and cubed transformations		
	ERM	$+\mathcal{L}_{A1}(\phi)$	$+\mathcal{L}_{A1}(\phi,\psi)$	ERM	$+\mathcal{L}_{A1}(\phi)$	$+\mathcal{L}_{A1}(\phi,\psi)$
No DCDS	0.3485	0.3537	**0.3369**	1.5150	0.4652	**0.3370**
With DCDS	0.4144	0.2290	**0.1777**	0.8720	1.5868	**0.8241**

Table 3. Toy experimental results: MSE losses on testing set.$+\mathcal{L}_{A1}(\phi)$ denotes $\mathcal{L}_{A1}$ is used without ψ while $+\mathcal{L}_{A1}(\phi,\psi)$ denotes ψ is used. Best results are highlighted as **bold**. DCDS denotes domain-conditioned distribution shift.

GAim2	$H(P(\psi(\mathbf{Y}) \mid \phi(\mathbf{X}))) + H(P(\mathbf{Y} \mid \psi(\mathbf{Y})))$	**GReg1**	$D_{\mathrm{KL}}(P(\phi(\mathbf{X}), \mathbf{Y} \mid \mathcal{D}) \| P(\mathcal{O}))$
iAim1	$H(P(\phi(\mathbf{X}) \mid \mathcal{D}))$	**GAim1**	$H(P(\phi(\mathbf{X}), \mathbf{Y}) \mid \mathcal{D})$
iReg2	$-H(P(\phi(\mathbf{X}), \mathcal{D}) + H(P(\phi(\mathbf{X})))$	**GReg2**	$-H(P(\phi(\mathbf{X}) \mid \psi(\mathbf{Y}))) + H(P(\phi(\mathbf{X})))$

Table 4. Notations for terms.

$P(\mathbf{Y}|\mathcal{D})$ static assumption brought by ψ of $\mathbf{Y}$. Furthermore, we conduct experiments on regression, segmentation, and classification tasks and use complex benchmark datasets. For a simplification, please refer to Table 4 for the formulations of terms and their alias.

5.1. Toy experiments on synthetic datasets

We perform a regression task on synthetic data to illustrate the impact of using ψ, showcasing its potential for superior results if ψ is not bijective.

Synthetic data. Supplementary Figure 3 illustrates the construction of synthetic data, built on $\mathbf{X}$-$\mathbf{Y}$ pair latent features with a linear relationship, ensuring invariant existence. To better explore this issue, we created four distinct data groups: without and with distribution shift, used affine or squared and cubed transformations as domain-conditioned transformations, and their cross combinations. More description can be seen in Supplementary 10.

Experimental setup. We use two of three constructed domains for training and validation and the last one for testing. Validation and test losses are calculated by MSE. To maintain fairness, all experiments adopt the same network which is selected by the best validation results. Learning aims to find invariant hidden features of X, Y while preserving predictive ability from unseen X to Y.

Results. Toy experiment results are reported in Table 3, which are also visualized in Figure 4. It is observed that across all settings, employing ψ with $\mathcal{L}_{A1}$ yields superior results, outperforming ERM and ERM+$\mathcal{L}_{A1}(\phi)$ without ψ, validating the enhanced generalization effect brought by utilizing ψ whenever Y varies per domain, supporting Eq. 17. Supplementary Figure 4 shows that ψ does learn the invariant representations for Y to relax previous Y-invariant assumption. Specifically, learning the invariance of Y with ψ results in superior invariant representations as the latent representations of X, Y are primarily linear, aligning with X and Y's linear relationship during data construction. The bottom-left figures reveal that though ERM has learned the most invariant $\phi(X)$, it suffers the worst test loss, indicating that a well-learned invariant $\phi(X)$ is not sufficient when Y also has domain-dependent traits. The results also suggest that assuming that Y vary across domains, using $\mathcal{L}_{A1}$ without ψ may not yield superior results.

	SILog↓	Abs Rel↓	RMS↓	Sq Rel↓	RMS log↓	δ_1↑	δ_2↑	TD
Backbone (Swin-L)	11.1473	10.98	56.11	8.86	14.32	87.47	98.05	S
VA-DepthNet (**GAim2**)	10.9357	11.15	56.36	9.02	14.41	87.73	98.02	
GReg1+**GAim2**	10.6548	10.49	52.63	8.04	13.72	89.75	98.12	
GAim1+**GReg1**+**GAim2**	10.1924	10.39	**50.52**	**7.68**	13.39	89.86	**98.17**	
GAim1+**GReg1**+ **GAim2**+**GReg2** (GMDG)	**10.1402**	**10.27**	50.59	7.72	**13.22**	**90.53**	97.98	
Backbone (Swin-L)	14.2078	16.20	81.22	16.30	20.59	72.44	96.35	Co
VA-DepthNet (**GAim2**)	14.7080	16.76	83.17	17.07	21.44	71.46	95.11	
GReg1+**GAim2**	14.1600	16.41	80.78	16.56	21.02	72.06	95.38	
GAim1+**GReg1**+**GAim2**	**13.9978**	15.90	77.97	15.70	20.25	73.37	**95.86**	
GAim1+**GReg1**+ **GAim2**+**GReg2** (GMDG)	14.2803	**15.57**	**77.45**	**15.27**	**19.94**	**74.40**	95.47	
Backbone (Swin-L)	11.6132	12.87	44.51	8.01	15.58	84.57	97.87	O
VA-DepthNet (**GAim2**)	11.5080	12.50	43.98	7.67	15.37	84.78	97.87	
GReg1+**GAim2**	10.4061	11.71	39.02	6.87	13.83	88.27	98.17	
GAim1+**GReg1**+**GAim2**	10.4907	11.53	**38.43**	6.69	13.68	88.46	**98.17**	
GAim1+**GReg1**+ **GAim2**+**GReg2** (GMDG)	**10.4438**	**11.33**	38.95	**6.66**	**13.67**	**88.86**	98.16	
Backbone (Swin-L)	14.7350	18.36	52.31	13.05	20.06	74.74	93.94	H
VA-DepthNet (**GAim2**)	15.0300	17.99	56.54	13.20	20.64	72.40	94.38	
GReg1+**GAim2**	14.7377	17.02	55.39	12.06	19.86	74.05	95.17	
GAim1+**GReg1**+**GAim2**	14.5018	17.14	**52.10**	12.01	19.37	76.13	94.95	
GAim1+**GReg1**+ **GAim2**+**GReg2** (GMDG)	**14.1414**	**15.90**	52.22	**10.72**	**18.95**	**76.27**	**96.10**	
Backbone (Swin-L)	12.9258	14.60	58.54	11.56	17.64	79.81	96.55	Avg.
VA-DepthNet(**GAim2**)	13.0454	14.60	60.01	11.74	17.97	79.09	96.35	
GReg1+**GAim2**	12.4897	13.91	56.96	10.88	17.11	81.03	96.71	
GAim1+**GReg1**+**GAim2**	12.2957	13.74	**54.76**	10.52	16.67	81.96	96.79	
GAim1+**GReg1**+ **GAim2**+**GReg2** (GMDG)	**12.2514**	**13.27**	54.80	**10.09**	**16.45**	**82.52**	**96.93**	

Table 5. Regression results: Comparison of results between proposed and previous methods. Added terms to the baseline are highlighted as **blue**. The best results for each group are highlighted in **bold**. TD: Test Domain.

5.2. Regression on benchmark datasets: Monocular depth estimation

We conduct the Monocular Depth Estimation task as the real-world regression task to further verify GMDG.

Experimental setup. We employ VA-DepthNet [32] with Swin-L [33] backbone as the baseline and follow their hyperparameter settings. Experiments are conducted on NYU Depth V2 [46]. To construct multiple domains, we split the dataset into four categories: 'School' (S), 'Office' (O), 'Home' (H), and 'Commercial' (Co). We conduct the standard leave-one-out cross-validation as an evaluation method. We use the best checkpoint on the seen domains for the evaluation. Note that all models are trained on the newly constructed dataset. Statistical results on popular evaluation metrics such as the square root of the Scale Invariant Logarithmic error (SILog), Relative Squared error (Sq Rel), Relative Absolute Error (Abs Rel), Root Mean Squared error (RMS), and threshold accuracy (δ_1, δ_2) are used as evaluation metrics. See more experimental details in the Supplementary 10.

Results. The Monocular Depth Estimation results are exhibited in Table 5. It can be seen that using terms that are proposed in GMDG leads to better generalization on unseen domains, and using the full GMDG leads to the best results in most metrics. The improvements suggest the feasibility of our GMDG in real-world regression tasks. Specifically, using **GReg2** with other terms significantly improves the

TD	Ci	B	M	Avg.
DeepLabv3+	35.46	25.09	31.94	30.83
IBN-Net	35.55	32.18	38.09	35.27
RobustNet (**GAim2**, **GReg2**)	37.69	34.09	38.49	36.76
GAim1+GAim2+GReg2	38.58	34.72	39.11	37.47
GReg1+GAim2+GReg2	38.13	**35.02**	39.29	37.48
GAim1+**GReg1**+GAim2+GReg2 (GMDG)	**38.62**	34.71	**39.63**	**37.65**

Table 6. Segmentation results: Comparison of mIoU(%) between proposed and previous methods. The models are trained on GTAV and SYNTHIA domains. The added objective terms are highlighted as **blue**. The best results are highlighted in **bold**.

results when Home is the unseen domain, which is the most difficult domain to be generalized for the VA-DepthNet, and barely compromises the performances while using other domains as the unseen. This reveals that suppressing the causality can improve the generalization of the model (refer to Figure 2 (b) for visual results). Note that, except SILog, all metrics results are scaled by 100 for readability; due to the lack of objective targeting on the mDG problem, VA-DepthNet performs worse than its baseline. See more visualizations in Supplementary 11.

5.3. Segmentation on benchmark datasets

Experimental setup. We follow the experimental setup of RobustNet [10] for mDG segmentation experiments, particularly using DeepLabV3+ [8] as the semantic segmentation model architecture, with ResNet-50 backbone and SGD optimizer. As shown in Table 1, RobustNet's objective is equivalent to using **GAim2** and **GReg2**. Consistent with previous methods, mIoU serves as our evaluation metric. Datasets comprise real-world datasets (Cityscapes [11] (Ci), BDD-100K [53] (B), Mapillary [35] (M)) and synthetic datasets (GTAV [41], SYNTHIA [42]). Specifically, we train a model on GTAV and Cityscapes, testing on other datasets. We compare our results to DeepLabv3+[8], IBN-NET[36] and RobustNet [10]. We use Intersection over Union (mIoU) as the evaluation metric. See Supplementary 10 for more experimental details.

Results. Table 6 shows the efficacy of our proposed objective in segmentation tasks upon introducing ψ. Ablation results highlight that using ψ alongside **GAim1** can enhance baseline performance, experimentally substantiating that the introduction of ψ, in relaxing assumptions, boosts performance for better generalization. Using **GReg1** alone also improves average mIoU. Importantly, the most enhancement in average mIoU is observed when **GReg1** and **GAim1** are used together, which finds validation in the PUB derivation in Eq. 7. See more results and visualizations in Supplementary 11.

5.4. Classification on benchmark datasets

Experimental setup. We operate on the DomainBed suite [14] and leverage standard leave-one-out cross-validation as an evaluation method. We experiment on 5 real-world benchmark datasets, including PACS [25], VLCS [12], OfficeHome [50], TerraIncognita [2], and DomainNet [38]. The results are the averages from three trials of each experiment. Following MIRO, two backbones are used for the training (ResNet-50 [15] pre-trained in the ImageNet [15] and RegNetY-16GF backbone with SWAG pre-training [47]). The backbones are trained with our proposed objective barely and further with SWAD [6], respectively. See Supplementary 10 for more experimental details.

TD	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
Non-ensemble methods						
MMD [27]	84.7±0.5	77.5±0.9	66.3±0.1	42.2±1.6	23.4±9.5	58.8
Mixstyle [57]	85.2±0.3	77.9±0.5	60.4±0.3	44.0±0.7	34.0±0.1	60.3
GroupDRO [43]	84.4±0.8	76.7±0.6	66.0±0.7	43.2±1.1	33.3±0.2	60.7
IRM [1]	83.5±0.8	78.5±0.5	64.3±2.2	47.6±0.8	33.9±2.8	61.6
ARM [56]	85.1±0.4	77.6±0.3	64.8±0.3	45.5±0.3	35.5±0.2	61.7
VREx [23]	84.9±0.6	78.3±0.2	66.4±0.6	46.4±0.6	33.6±2.9	61.9
CDANN [29]	82.6±0.9	77.5±0.1	65.8±1.3	45.8±1.6	38.3±0.3	62.0
DANN [13]	83.6±0.4	78.6±0.4	65.9±0.6	46.7±0.5	38.3±0.1	62.6
RSC [17]	85.2±0.9	77.1±0.5	65.5±0.9	46.6±1.0	38.9±0.5	62.7
MTL [4]	84.6±0.5	77.2±0.4	66.4±0.5	45.6±1.2	40.6±0.1	62.9
MLDG [26]	84.9±1.0	77.2±0.4	66.8±0.6	47.7±0.9	41.2±0.1	63.6
Fish [44]	85.5±0.3	77.8±0.3	68.6±0.4	45.1±1.3	42.7±0.2	63.9
ERM [49]	84.2±0.1	77.3±0.1	67.6±0.2	47.8±0.6	44.0±0.1	64.2
SagNet [34]	86.3±0.2	77.8±0.5	68.1±0.1	48.6±1.0	40.3±0.1	64.2
SelfReg [22]	85.6±0.4	77.8±0.9	67.9±0.7	47.0±0.3	42.8±0.0	64.2
CORAL [48]	86.2±0.3	78.8±0.6	68.7±0.3	47.6±1.0	41.5±0.1	64.5
mDSDI [5]	86.2±0.2	79.0±0.3	69.2±0.4	48.1±1.4	42.8±0.1	65.1
Use ResNet-50 [15] as oracle model.						
Style Neophile [20]	**89.11**	-	65.89	-	44.60	-
MIRO [19] (**GReg1**)	85.4±0.4	79.0±0.3	70.5±0.4	50.4±1.1	44.3±0.2	65.9
GMDG	85.6±0.3	**79.2**±0.3	**70.7**±0.2	**51.1**±0.9	**44.6**±0.1	**66.3**
Use RegNetY-16GF [47] as oracle model.						
MIRO	**97.4**±0.2	79.9±0.6	80.4±0.2	58.9±1.3	53.8±0.1	74.1
GMDG	97.3±0.1	**82.4**±0.6	**80.8**±0.6	**60.7**±1.8	**54.6**±0.1	**75.1**
Ensemble methods						
	PACS	**VLCS**	**OfficeHome**	**TerraInc**	**DomainNet**	**Avg.**
Use multiple oracle models.						
SIMPLE [30]	88.6±0.4	79.9±0.5	84.6±0.5	57.6±0.8	49.2±1.1	72.0
SIMPLE++ [30]	**99.0**±0.1	**82.7**±0.4	**87.7**±0.4	**59.0**±0.6	**61.9**±0.5	**78.1**
Use ResNet-50 [15] as oracle model.						
MIRO + SWAD	**88.4**±0.1	79.6±0.2	72.4±0.1	52.9±0.2	47.0±0.0	68.1
GMDG + SWAD	**88.4**±0.1	**79.6**±0.1	**72.5**±0.2	**53.0**±0.7	**47.3**±0.1	**68.2**
Use RegNetY-16GF [47] as oracle model.						
MIRO + SWAD	96.8±0.2	81.7±0.1	83.3±0.1	64.3±0.3	60.7±0.0	77.3
GMDG + SWAD	**97.9**±0.3	**82.2**±0.3	**84.7**±0.2	**65.0**±0.2	**61.3**±0.2	**78.2**

Table 7. Classification results: Comparison of results between the proposed and previous non-ensemble and ensemble mDG methods. The best results for each group are highlighted in **bold**.

Results. Table 7 displays the results of non-ensemble algorithms and ensemble algorithms that employ pre-trained models as oracle models. Specifically, our proposed objectives demonstrate more substantial improvements when a higher-quality pre-trained oracle model ($\mathcal{O}$) is applied. When employing the ResNet-50 model, our approach yields average improvements of approximately 0.3% and 0.1% without and with SWAD, respectively, compared to MIRO. In contrast, when RegNetY-16GF serves as an oracle, GMDG results in significant average improvements of 1.1% and 0.9% without and with SWAD, respectively. Remarkably, our approach outperforms 0.1% more than the SOTA method, SIMPLE++, which relies on an ensemble of 283 pre-trained models as oracle models, whereas ours only engages a single pre-trained model. Overall, these results strongly support GMDG's effectiveness in classifica-

Used objectives	Art	Clipart	Product	Real	Avg.	Imp.
Without $\mathcal{O}$ (**GReg1**)						
GAim2 (ERM)	78.4±0.7	68.3±0.5	85.8±0.4	85.8±0.3	79.6±0.2	0.0
GAim2 + iAim1 (DANN)	79.1±1.0	68.6±0.0	85.6±0.8	86.1±0.5	79.8±0.2	+0.2
GAim2 + GAim1 (CDANN, CIDG)	79.1±0.7	69.1±0.1	85.7±0.5	86.3±0.6	79.9±0.4	+0.3
GAim2 +iReg2 (CORAL+ψ)	79.1±0.1	69.9±0.4	86.0±0.1	86.3±0.4	80.3±0.2	+0.7
GAim2 + GReg2	79.2±0.1	**69.9**±1.4	86.1±0.5	86.1±0.1	80.3±0.3	+0.7
GAim2 + GAim1 + GReg2 (MDA+ψ)	**79.5**±1.1	69.2±1.2	**86.2**±0.2	**86.5**±0.2	**80.3**±0.0	**+0.7**
With $\mathcal{O}$ (**GReg1**)						
GAim2 + GReg1 (MIRO, SIMPLE)	83.2±0.6	72.6±1.1	89.9±0.5	90.2±0.1	84.0±0.2	0.0
GAim2 + GReg1 +iAim1	83.4±0.5	73.1±0.8	89.7±0.4	90.1±0.3	84.1±0.2	+0.1
GAim2 + GReg1 + GAim1	83.7±0.3	74.0±0.6	90.1±0.3	90.3±0.2	84.5±0.2	+0.4
GAim2 + GReg1 + iReg2	82.9±0.5	72.5±0.3	**90.3**±0.3	90.0±0.3	83.9±0.1	-0.1
GAim2 + GReg1 + GReg2	83.4±0.2	72.3±0.2	90.1±0.3	90.1±0.3	84.0±0.2	+0.0
GAim2 + GReg1 + GAim1 + GReg2 (GMDG)	**84.1**±0.2	**74.3**±0.9	89.9±0.4	**90.6**±0.1	**84.7**±0.2	**+0.7**

Table 8. Ablation studies: Results of using different combinations of terms on HomeOffice. Imp. denotes Improvement that gained form **GAim2** and **GAim2 + GReg1**, respectively.

tion tasks. See more results in Supplementary 11.

5.5. Ablation studies

To better compare our objective with previous objectives, we conduct a systematic ablation study on the classification task since most previous objectives are only available for classification due to the lack of ψ. **Experimental setup.** In the ablation studies, we test varied terms (see Table 8) combinations on the HomeOffice dataset using SWAG pre-training [47] and SWAD [6]. Every experiment is repeated in three trials, sharing the same hyper-parameter settings for evaluation. See Supplementary 10 for more details.

Results. Table 8 presents ablation study results. The first column denotes previous methods equivalent to term combinations. The main findings are as follows. See Supplementary 11.1 for more other findings.

1). Previous methods that partially utilize our proposed objectives often yield suboptimal results. Note that **iAim1** is the unconditional version of **GAim1**. By eliminating other factors, it can be seen that employing our proposed full objectives offers the most significant improvements, while previous objectives may lead to inferior results.

2). The effectiveness of using conditions. By conducting uniform implementation and testing, it can be observed that the use of conditions yields superior results compared to the unconditional approach. This observation aligns with Eq. 19, suggesting that minimizing the gap between conditional features across domains leads to improved generalization. The disparity in performance between CDANN and DANN might be attributed to differences in their implementation details.

3). Learning invariance is crucial, regardless of whether integrating prior knowledge. Evidently, learning invariance facilitates improvement whether prior is applied or not, as validated in the PUB derivation in Eq. 7. This contradicts MIRO's argument that achieving similar representations to a prior can replace the need for learning invariance.

4). Impacts of using prior. The significant improvement owes to the use of a pre-trained oracle model ($\mathcal{O}$) preserving correlations between $\mathbf{X}$ and $\mathbf{Y}$ - a concept validated by MIRO and SIMPLE. However, utilizing our full set of objectives can further enhance this improvement by an additional 0.7%. Notably, the invalid causality may not work when using prior knowledge, while the invariance across domains is not permitted. We hypothesize that such invalid causality is inherently eliminated within a 'good' feature space obtained by $\mathcal{O}$, but may be reintroduced when we minimize the domain gap with $\mathcal{O}$. Thus, using the full objective can synergistically produce optimal results.

5). Constraining only the covariance shifts of features across domains (**iReg2**) does not guarantee better results when prior knowledge is available. We find that using the objectives of CORAL performs better than DANN, CDANN, and CIDG. The results suggest that considering the covariance shifts of features does lead to improvements, which we hypothesize are primarily driven by $H(P(\phi(\mathbf{X})))$. However, when a large pre-trained oracle model ($\mathcal{O}$) is provided, the performance actually degrades. This implies that the use of $\mathcal{O}$ implicitly minimizes the covariance shifts of features across domains. Under this scenario, the unexpected effect of $-H(P(\phi(\mathbf{X}|\mathcal{D})))$ hinders improvement, while the benefits brought by $H(P(\phi(\mathbf{X})))$ are diminished by the use of prior knowledge. In contrast, **GReg2** continues to yield improvements. This suggests that GMDG is more versatile and suitable for various situations.

6. Conclusion

In this paper, we propose a general objective, namely GMDG, by relaxing the static distribution assumption of $\mathbf{Y}$ through a learnable mapping ψ. GMDG is applicable to diverse mDG tasks, including regression, segmentation, and classification. Empirically, we design a suite of losses to achieve the overall GMDG, adaptable across various frameworks. Extensive experiments validate the viability of our objective across applications where previous objectives may yield suboptimal results compared to ours. Both theoretical analyses and empirical results demonstrate the synergistic effect of distinct terms in the proposed objective. Simplistically, we assume equal domain weights whilst minimizing GJSD, presenting the future scope for dealing with imbalance situations triggering unequal domain weights.

Acknowledgments

The work was partially supported by the following: National Natural Science Foundation of China under No. 92370119, No. 62376113, and No. 62206225; Jiangsu Science and Technology Program (Natural Science Foundation of Jiangsu Province) under No. BE2020006-4; Natural Science Foundation of the Jiangsu Higher Education Institutions of China under No. 22KJB520039.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 7

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 7, 6

[3] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011. 1

[4] Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100, 2021. 7

[5] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201, 2021. 2, 7

[6] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34: 22405–22418, 2021. 7, 8, 9, 15

[7] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 301–318. Springer, 2020. 2

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 7

[9] Wonwoong Cho, Ziyu Gong, and David I Inouye. Cooperative distribution alignment via jsd upper bound. *Advances in Neural Information Processing Systems*, 35:21101–21112, 2022. 3, 1

[10] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 1, 2, 3, 7, 4

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 7, 5

[12] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013. 7, 6

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 1, 2, 5, 7, 3

[14] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 1, 7, 3, 9, 15

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[16] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pages 292–302. PMLR, 2020. 1, 2, 3, 4, 8

[17] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 124–140. Springer, 2020. 7

[18] Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. *arXiv preprint arXiv:2207.03162*, 2022. 4

[19] Cha Junbum, Lee Kyungjae, Park Sungrae, and Chun Sanghyuk. Domain generalization by mutual-information regularization with pre-trained models. *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 5, 7, 9, 15

[20] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7130–7140, 2022. 3, 7

[21] Steven M Kay. *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993. 4

[22] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021. 7

[23] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 7

[24] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9946, 2022. 3

[25] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 7, 6

[26] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 7

[27] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 7

[28] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1, 2, 5, 3, 4

[29] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018. 1, 2, 5, 7, 4

[30] Ziyue Li, Kan Ren, Xinyang Jiang, Yifei Shen, Haipeng Zhang, and Dongsheng Li. Simple: Specialized model-sample matching for domain generalization. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 3, 5, 7

[31] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991. 3

[32] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. *arXiv preprint arXiv:2302.06556*, 2023. 1, 2, 3, 6, 5

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 6

[34] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 7

[35] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 7

[36] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 7

[37] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2594–2605, 2022. 3

[38] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 7, 6

[39] Samir M Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, and Stefano Rini. Empirical risk minimization with relative entropy regularization: Optimality and sensitivity analysis. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 684–689. IEEE, 2022. 2

[40] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022. 9, 15

[41] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 7, 5

[42] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 7, 5

[43] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 7

[44] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021. 7

[45] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 3

[46] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 6, 4

[47] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022. 7, 8

[48] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016. 1, 5, 7, 3, 4, 6

[49] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998. 7

[50] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for

unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 7, 6

[51] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1

[52] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020. 4

[53] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 7, 5

[54] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR, 2013. 3

[55] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. 3

[56] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021. 7

[57] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 7

[58] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1