ENHANCED ADVERSARIAL DOMAIN GENERATION VIA OPTIMIZED BATCH NORMALIZATION AND CLASSIFIER

Anonymous authors
Paper under double-blind review

ABSTRACT

In the endeavor of domain generalization, the objective is to develop a classification model, utilizing multiple source domains, that can subsequently be generalized to unseen target domains. The crux of domain generalization lies in discerning and learning discriminative features that are invariant across domains. Techniques utilizing adversarial domain generalization are paramount in achieving invariant representations. To mitigate the aforementioned impediment, we introduce a novel methodology, termed Auxiliary Classifier in Adversarial Domain Generalization (ACADG). ACADG endeavors to augment the diversity within the source domain through the integration of an auxiliary classifier. By amalgamating standard task-related losses—such as cross-entropy loss for classification and adversarial loss for domain discrimination—the overarching aim is to ensure the acquisition of condition-invariant features across all source domains, while concurrently enhancing the diversity of source domains. We have undertaken comprehensive image classification experiments on benchmark datasets within the realm of domain generalization. Our model demonstrates significant generalization capacity, surpassing contemporaneous state-of-the-art domain generalization methodologies. In the context of mathematical formalization, consider that the aforementioned adversarial loss, auxiliary classifier, and task-related losses are represented with pertinent LaTeX symbols and equations, reflecting the intricate interdependencies and mathematical nuances underpinning the proposed methodology.

1 INTRODUCTION

In contemporary research landscapes, algorithms driven by machine learning have demonstrated unprecedented efficacy across an extensive array of applications. Nonetheless, a core challenge intrinsic to machine learning is the frequent inability of models, resultant of biases in data, to generalize effectively when exposed to data sourced from diverse distributions, owing to discrepancies between training and test sets. The concept of Domain Adaptation Pan et al. (2010); Ben-David et al. (2006) has gained significant scholarly attention as a prospective resolution to this issue, serving to mitigate the disparities between source and target domains. However, the introduction of a new dataset necessitates the reiteration of this methodology, a process which is markedly time-intensive. To cultivate models with enhanced generalizability utilizing data spanning multiple domains, the pursuit of Domain Generalization Li et al. (2017); Muandet et al. (2013) is advocated.

The challenge inherent in domain generalization is significantly magnified due to the absence of congruence between the distributions of the source and target domains, compounded by a lack of insight into the distribution of the target domain throughout the training phase. To enhance the extrapolative capabilities of the models trained, a multitude of strategies have been pursued, each offering distinct vantage points and methodologies. The essence of domain generalization is to cultivate domain-agnostic feature representations across various source domains.

To facilitate the learning of an invariant transformation, Muandet *et al.*Muandet et al. (2013) innovatively implemented a kernel-based strategy, grounded in domain-invariant component analysis. Meanwhile, Zhou *et al.*Zhou et al. (2020a) synthesized three integral components: a label classifier, a domain classifier, and a domain transformation network, in order to map the source training data

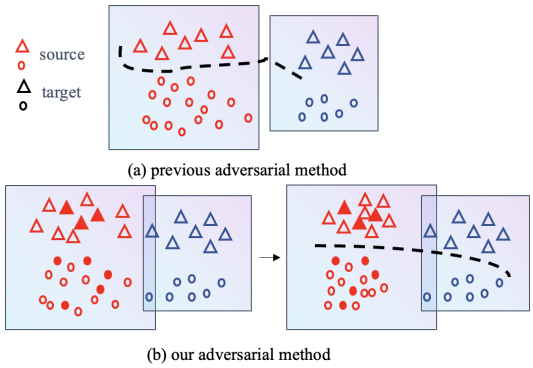


Figure 1: (a) Antecedent methodologies fabricate challenging exemplars for the purpose of instructing a classifier, a process that may not effectively amplify the distribution range of the source data. (b) Conversely, the methodology proposed herein strategically displaces samples from the vicinity of the target data with the intention of broadening the distribution scope of the source data.

into unobserved domains. This ensures that the newly generated data is precisely classified by the label classifier while concurrently deceiving the domain classifier.

Concurrently, a distinct strand of research has delved into exploring a plethora of data augmentation strategies.Yang et al. (2021); Shankar et al. (2018); Volpi et al. (2018); Carlucci et al. (2019). Yang *et al*.Yang et al. (2021) propelled the field of teacher-student learning through an adversarial learning paradigm, which oscillates between representation learning based on knowledge distillation and data augmentation for novel domains. Shankar *et al*.Shankar et al. (2018) proffer a gradient-oriented domain perturbation method, and Zhao *et al*.Zhao et al. (2020) introduce an entropy regularization term founded on adversarial learning to assess the interdependencies between the acquired features and the associated class labels.

The paradigmatic advancements in the aforementioned studies have catalyzed our endeavor to systematically review domain-invariant feature representations. A prevalent postulate within extant methodologies is that the marginal distribution, denoted as $P(X)$, undergoes transformations, while the conditional distribution $P(Y|X)$ sustains its constancy. Consequently, to acquire a domain invariant feature, $P(F(X))$, the integration of an adversarial learning approach becomes imperative. This enables the comprehensive training of the joint classification model across all available source domains, allowing the utilization of domain-invariant features acquired through $F(X)$ to prognosticate on new datasets.

Nonetheless, alterations in $P(X)$ do not inevitably attest to the stability of $P(Y|X)$. For instance, Li *et al*. Li et al. (2018c) advocates for the adversarial training of a class-invariant conditional distribution, represented as $P(F(X)|Y)$. However, the efficacy of this methodology may diminish as the number of classes escalates, presenting potential limitations in its applicability and robustness.

In response to the highlighted challenges, we introduce a paradigm wherein both $P(X)$ and $P(Y|X)$ exhibit variation across domains, focusing on domain generalization within an integrated deep learning architecture. For this objective, we cultivate an invariant conditional neural network designed to reduce the disparity in $P(X|Y)$ over multiple domains. Drawing upon contemporary advances in deep domain generalization, we advocate for an adversarial network architecture, geared towards achieving a domain-agnostic representation by ensuring that the cultivated representations remain perceptually consistent across distinct domains.

To augment the diversity within the source domain and mitigate the detrimental impacts arising from alterations in class priors $P(Y)$ amongst source domains (illustrated in Figure. 1), auxiliary classifiers are introduced in this study. We theoretically substantiate that our auxiliary classifier enhances the model's diversity.

Additionally, we incorporate Adaptive Sharpness Minimization (ASM) Kwon et al. (2021), aimed at realizing smooth minima in task loss, which, in turn, promotes enhanced generalization.

The empirical validity of our methodology is rigorously affirmed through extensive analyses across diverse datasets, including Digits-DG, PACS, Office-Home, and DomainNet, utilizing ResNet backbones. In essence, our research furnishes the following significant contributions:

- We introduce an auxiliary classifier in adversarial generalization (ACADG), designed to bolster source domain diversity by amalgamating two characteristic loss functions: an auxiliary classifier for classification and adversarial loss for domain discrimination.
- To optimize the smoothness of task loss adjacent to optima in adversarial learning, we recommend the employment of a straightforward and innovative approach known as ASM. This approach facilitates stable adversarial domain generalization and yields enhanced generalization within the target domain.
- We execute a comprehensive series of experiments on benchmark datasets to validate our propositions and demonstrate the efficacy of the suggested methodology.

2 Related Works

2.1 Domain Generalization

Domain adaptation (DA) represents a process enabling the transfer of knowledge from designated source domains to targeted domains, extensively discussed in literature Wu et al. (2021); Wang et al. (2019). Predominantly, unsupervised domain adaptation has been a pivotal methodology within DA, addressing domain shift dilemmas through the mitigation of domain discrepancies between labeled source domains and their unlabeled target counterparts Cai et al. (2019); Ma et al. (2021), employing strategies such as domain adversarial learning or domain distance minimization Zhang et al. (2019); Long et al. (2015). However, practical applications often demand preliminary access to domain's information, a requirement that can be resource-intensive or unattainable.

Alternative approaches encompass learning feature or gradient masks for regularization Chattopadhyay et al. (2020); Huang et al. (2020a), and normalization of batches and instances Seo et al. (2020). Our approach resonates with works providing a causal interpretation of domain generalization Mahajan et al. (2020), where we employ auxiliary classifiers for estimating invariant parts of $P(Y|X)$ as conditional distributions vary across domains, thereby enabling the model to generalize features to unseen targets through adversarial training.

2.2 Adaptive Sharpness Minimization

In the realm of neural networks, the efficacy of overparameterized models for enhancing generalization remains a subject of debate Keskar et al. (2017). Smoother minima have been theoretically linked to superior generalization on unseen data Hochreiter & Schmidhuber (1997; 1994); He et al. (2019); Dziugaite & Roy (2017). However, the pursuit of smoothing has often been marred by its computational intensiveness. The advent of Sharpness Aware Minimization (SAM) Foret et al. (2021) marked a paradigmatic shift by offering a pathway to smoother minima and thereby improved generalization. In this work, we leverage SAM with the aspiration to confluence towards smooth minima while simultaneously fostering enhanced generalization to the target domain.

3 Methodology

In a standard domain generalization scenario, we encounter K distinct source domains denoted as $\mathcal{D}^1, \ldots, \mathcal{D}^K$, where each domain $\mathcal{D}^j$ encompasses paired data and labels conforming to a joint distribution $P^j(X, Y)$. The primary objective is to cultivate a model utilizing these source domains and subsequently evaluate it within an unseen target domain. Typically, the distribution within this target domain diverges from those of the source domains.

3.1 Domain Generalization through Adversarial Learning

Initiating our discussion, we elucidate the manner in which the framework of adversarial learning can be instrumental in instilling domain generalization. This methodology encompasses the employment of a feature extractor F, characterized by parameters θ, and a classifier C, articulated

by parameters ϕ. The optimization of θ and ϕ is performed over K source domains through the minimization of cross-entropy loss, aiming to facilitate the refinement of domain generalization capabilities in the developed model.

$$\begin{aligned}\mathcal{L}_{cls} &= -\sum_{i=1}^{K} \mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)} [\log(C^T(Y \mid F(X)))] \\ &= -\sum_{i=1}^{K}\sum_{j=1}^{N_s} \mathbf{y}_j^{(i)} \cdot \log(C(F(\mathbf{x}_j^{(i)})).\end{aligned} \tag{1}$$

In this formulation, $\mathbf{y}_j^{(i)}$ denotes the class label embodied as a one-hot vector $y_j^{(i)}$, with "$\cdot$" signifying the dot product operation. The forecasted label distribution for domain i is depicted as $C^T(Y \mid F(X))$. However, models purely optimized based on classification loss exhibit limitations in learning domain-invariant features. To surmount this obstacle and to amplify the acquisition of domain-invariant features, we advocate the incorporation of adversarial domain generalization as our foundational approach.

3.2 Adversarial Domain Generalization

Adversarial learning emerges as a formidable methodology for the extraction of domain-invariant features. This is realized through a structured confrontation between the feature extractor and the discriminator. The discriminator is designated the responsibility of ascertaining the domain labels of feature gradients present in both source and target domains, whilst the feature extractor endeavours to mislead it. The paramount aim is to attain a balanced state wherein the divergence between feature distributions is condensed to a minimum. This strategy has demonstrated its efficacy across diverse domains grappling with the issue of domain shift, including computer vision and natural language processing fields.

The principal aspiration of acquiring domain-invariant features within the adversarial learning paradigm can be mathematically articulated as:

$$\begin{aligned}\min_F \max_D \mathcal{L}_{adv} &= \sum_{i=1}^{K} \mathop{\mathbb{E}}_{X\sim P_i(X)} [\log D(F(X))] \\ &= \sum_{i=1}^{K}\sum_{j=1}^{N_i} \mathbf{d}_j^{(i)} \cdot \log\left(D\left(F\left(\mathbf{x}_j^{(i)}\right)\right)\right),\end{aligned} \tag{2}$$

where $\mathbf{d}_j^{(i)}$ represents the one-hot vector corresponding to domain label i, satisfying the condition $\sum_{i=1}^{K} D(F(x^{(i)})) = 1$.

The optimization of Eq. 2 is capable of engendering marginal distributions that remain invariant, elucidated as $P_1(F(X)) = P_2(F(X)) = \cdots = P_K(F(X))$. However, Zhao et al.Zhao et al. (2020) have elucidated that this does not ensure the invariance of the conditional distribution $P(Y|F(X))$ across the distinct domains. Such a scenario can culminate in the diminution of the model's generalization capabilities. Although the classifier endeavors to aggregate samples emanating from an identical category within the feature space—contributing to the learning of the invariant conditional distribution—a residual issue persists. To resolve this predicament, Zhao *et al*. Zhao et al. (2020) recommend the implementation of a class-conditional domain classification network, which harmonizes the learning of domain-invariant and discriminative features. The auxiliary classifier intrinsic to adversarial learning is delineated as per Hou et al. (2022) as follows:

$$\begin{aligned}\min_F \max_D \mathcal{L}_{adv} &= \sum_{i=1}^{K}\sum_{j=1}^{N_i} d_j^{(i)} \log(D(F(x_j^{(i)}))) \\ &+ \lambda \sum_{i=1}^{K}\sum_{j=1}^{N_i} E_{X,Y\sim P_i(X,Y)} \log C(y_j^{(i)}|x_j^i)\end{aligned} \tag{3}$$

where $\lambda > 0$ is defined as a coefficient hyperparameter. However, reliance solely on the auxiliary classifier as expressed in Eq. 3 has the potential to culminate in diminished diversity, a phenomenon which will be elucidated in the ensuing discussions.

Proof. We demonstrate that the incorporation of an auxiliary classifier can be equivalently represented by the subsequent formula:

$$\begin{aligned}
&\max_D \mathbb{E}_{x,y\sim P_{X,Y}}[\log C(y \mid x)] \\
&= \mathbb{E}_{x\sim P_X}\mathbb{E}_{y\sim P_{Y|X}}[\log C(y \mid x)] \\
\Rightarrow &\min_D \mathbb{E}_{x\sim P_X}\mathbb{E}_{y\sim P_{Y|X}}[-\log C(y \mid x)] \\
&= \mathbb{E}_{x\sim P_X}[H(p(y \mid x)) + \mathrm{KL}(p(y \mid x)\|C(y \mid x))] \\
\Rightarrow &C^*(y \mid x) \\
&= \arg\min_D \mathrm{KL}(p(y \mid x)\|C(y \mid x)) = p(y \mid x) = \frac{p(x,y)}{p(x)}
\end{aligned} \tag{4}$$

It can be further substantiated that the aforementioned formula is synonymous with the following representation:

$$\begin{aligned}
&\max_F \mathbb{E}_{x,y\sim Q_{X,Y}}\left[\log C^*(y \mid x)\right] = \mathbb{E}_{x,y\sim Q_{X,Y}}\left[\log \frac{p(x,y)}{p(x)}\right] \\
&= \mathbb{E}_{x,y\sim Q_{X,Y}}\left[\log \frac{p(x,y)}{q(x,y)}\frac{q(x)}{p(x)}\frac{q(x,y)}{q(x)}\right] \\
\Rightarrow &\min_F \mathbb{E}_{x,y\sim Q_{X,Y}}\left[\log \frac{q(x,y)}{p(x,y)}\right] \\
&- \mathbb{E}_{x\sim Q_X}\left[\log \frac{q(x)}{p(x)}\right] - \mathbb{E}_{x,y\sim Q_{X,Y}}\left[\log \frac{q(x,y)}{q(x)}\right] \\
\Rightarrow &\min_F \mathrm{KL}\left(Q_{X,Y}\|P_{X,Y}\right) - \mathrm{KL}\left(Q_X\|P_X\right) + H_Q(Y \mid X)
\end{aligned} \tag{5}$$

□

Our research posits that the minimization of the entropy of labels conditioned upon data, derived from the generated distribution ($\min_G H_Q(Y \mid X)$), engenders deterministic labels for the generated data. In essence, it necessitates the data in the source domain, corresponding to each class, to diverge from the classification hyperplane, elucidating the detected diminution in intra-class diversity as documented in Zhao et al. (2020). This phenomenon is conspicuously perceptible when the distributions of disparate classes exhibit substantial overlap, a commonplace occurrence given that neither state-of-the-art classifiers nor humans can achieve absolute classification accuracy on real-world datasets. To augment the diversity inherent to our model, we propose the following refinement to the previously stated formula:

$$\begin{aligned}
&\min_F \max_D \mathcal{L}_{adv}(\theta, \phi) \\
&= \sum_{i=1}^{K}\sum_{j=1}^{N_i} d_j^{(i)} log(D(F(x_j^{(i)}))) \\
&+ \lambda(\sum_{i=1}^{K}\sum_{j=1}^{N_i} E_{X,Y\sim P_i(X,Y)} \log C(y_j^{(i)}|x_j^i) \\
&- \sum_{i=1}^{K}\sum_{s=1}^{K}\sum_{j=1}^{N_i} E_{X,Y\sim P_j(X,Y)} \log C(y_j^s|x_j^s))
\end{aligned} \tag{6}$$

By exploiting Eq. 6, the classifier is endowed with the capability to distinguish between disparate class labels, thereby establishing a discriminative classifier proficient in recognizing varied samples with discrimination. A concise demonstration of this is provided as follows:

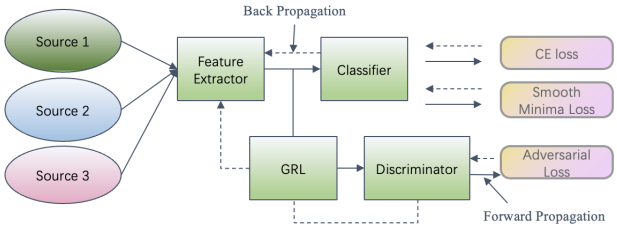


Figure 2: Overview of A Closer Look at Classifier (ACADG). We train an invariant conditional neural network that learns to minimize the discrpancy between $P(X|Y)$ across various domains. we propose to leverage sharpness minimizaton which only focuses on smoothing task loss, leading to stable training which results in effective generalization on target domain.

Proof.

$$\begin{aligned}
&\max_G \mathbb{E}_{x,y\sim Q_{X,Y}} \left[\log C^*(y \mid x)\right] - \mathbb{E}_{x,y\sim Q_{X,Y}} \left[\log C^*_{\mathrm{mi}}(y \mid x)\right] \\
\Rightarrow &\max_G \mathbb{E}_{x,y\sim Q_{X,Y}} \left[\log \frac{p(x,y)}{p(x)}\right] - \mathbb{E}_{x,y\sim Q_{X,Y}} \left[\log \frac{q(x,y)}{q(x)}\right] \\
\Rightarrow &\max_G \mathbb{E}_{x,y\sim Q_{X,Y}} \left[\log \frac{p(x,y)}{q(x,y)}\right] - \mathbb{E}_{x\sim Q_X} \left[\log \frac{p(x)}{q(x)}\right] \\
\Rightarrow &\min_G \mathrm{KL}\left(Q_{X,Y} \| P_{X,Y}\right) - \mathrm{KL}\left(Q_X \| P_X\right)
\end{aligned} \tag{7}$$

□

Pursuing this approach mitigates the issue of deficient intra-class diversity amongst source domain samples. We can articulate our method as an iterative minimax training process. To streamline training, the gradient between the classifier C and the feature extractor F is inverted using a gradient reversal layer. Within this layer, forward- and backward-propagation minimax training can be conducted.

3.3 Sharpness-Aware Minimization

Recent research by Rangwani et al. Rangwani et al. (2022) posits that the convergence of methods towards smooth optima bolsters generalization for supervised learning tasks such as classification. They argue that convergence to a smooth minima concerning task loss stabilizes adversarial training, which subsequently enhances performance in the target domain. To optimize generalization capability, we introduce the concept of adaptive sharpness Kwon et al. (2021) minimization (ASM). ASM's fundamental premise is to locate a smoother minima by applying the ensuing objective:

$$\min_{\theta} \max_{\|\epsilon\| \leq \rho} \mathcal{L}_{obj}(\theta + \epsilon) \tag{8}$$

The model incorporates two hyperparameters: $\rho \geq 0$, which stipulates the maximal norm for ϵ, and $\mathcal{L}$, representing the objective function requiring minimization. To maximize the first-order approximation, ASM seeks the precise solution of the inner maximization:

$$\begin{aligned}
\hat{\epsilon}(\theta) &\approx \arg\max_{\|\epsilon\| \leq \rho} \mathcal{L}_{adv}(\theta) + \epsilon^{\mathrm{T}} \nabla_\theta \mathcal{L}_{adv}(\theta) \\
&= \rho \nabla_\theta \mathcal{L}_{adv}(\theta) / \left\|\nabla_\theta \mathcal{L}_{adv}(\theta)\right\|_2
\end{aligned} \tag{9}$$

The augmentation, represented as $\hat{\epsilon}(\theta)$, is subsequently incorporated into the weight parameters denoted by θ. Following this incorporation, the computation of the gradient update for the parameter θ is represented as $\nabla_\theta \mathcal{L}obj(\theta)|\, \theta + \hat{\epsilon}(\theta)$. The elucidated methodology can be interpreted as a generic formulation aimed at augmenting smoothness, applicable to any objective function $\mathcal{L}_{obj}$. Pursuing this further, we introduce, by analogy, a risk factor cognizant of sharpness from the source, envisioned for the identification of smooth minima:

$$\max_{\|\epsilon\| \leq \rho} R^l_S\left(h_{\theta+\epsilon}\right) = \max_{\|\epsilon\| \leq \rho} \mathbb{E}_{x\sim P_S}\left[l\left(h_{\theta+\epsilon}(x), f(x)\right)\right] \tag{10}$$

3.4 OVERALL FORMULATION

We introduce a methodology termed Smooth Domain Adversarial Training, a focused approach aiming for convergence to smooth minima concerning the task loss, more specifically, the empirical source risk, whilst maintaining the integrity of the original discrepancy term. The optimization objective of the proposed model, referred to as the Auxiliary Classifier in Adversarial Domain Generalization (ACADG), is defined as follows:

$$\min_{\theta} \max_{\Phi} \max_{\|\epsilon\| \leq \rho} \mathbb{E}_{x \sim P_S} \left[l \left(h_{\theta+\epsilon}(x), y(x) \right) \right] + \mathcal{L}_{adv} \tag{11}$$

The primary constituent is the risk attuned to sharpness, and the subsequent constituent represents the discrepancy term, a component not characterized by smoothness within our outlined procedure. A comprehensive depiction of the framework inherent to the proposed methodology is presented in Figure. 2.

4 EXPERIMENTS

Experimental evaluations are conducted on four datasets that are predominantly employed in the realm of Domain Generalization (DG), encompassing Digits-DG Zhou et al. (2020a), PACS Li et al. (2017), Office-Home Venkateswara et al. (2017), and DomainNet Peng et al. (2019). The ensuing sections provide succinct introductions to each of the aforementioned datasets:

Digits-DG benchmark Zhou et al. (2020a) amalgamates four eminent datasets tailored for digit recognition: MNIST, MNIST-M, SVHN, and SYN.

PACS Li et al. (2017) embodies four divergent domains: Photo, Art Painting, Cartoon, and Sketch, each with seven discrete categories. Conforming to previous studiesZhou et al. (2020a), a singular domain is earmarked as the test domain, with the residual serving as source domains.

Office-Home Venkateswara et al. (2017), a venerated benchmark for domain adaptation, has accrued substantial acclaim in recent academic discourse.

DomainNet Peng et al. (2019) serves as a meticulous benchmark, offering an extensive assemblage of 600,000 images across 345 categories.

4.1 RESULTS

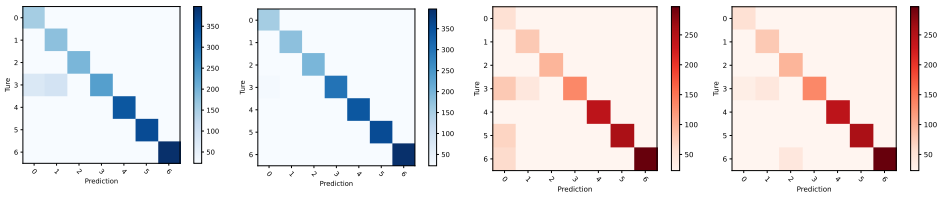
4.2 DETAILS OF IMPLEMENTATION

In adherence to the ubiquitously embraced leave-one-domain-out protocol as illustrated in Li et al. (2017), a singular domain is demarcated as the unseen target domain for evaluative purposes whilst the model is trained utilizing the residual domains. For the PACS dataset, the architecture employs ResNet18 and ResNet50 as backbones, pretrained on ImageNet. For the Office-Home and DomainNet datasets, ResNet50 and ResNet101 backbones are respectively utilized, with pretraining also on ImageNet.

The configuration of our network aligns with the paradigm established by DG_via_ER Zhao et al. (2020). The momentum parameter in SGD is assigned a value of 0.9, and a weight decay of 0.001 is applied. The learning rate is calibrated to $1e-3$ over a span of 30 epochs. The regularization strength, denoted by the ρ value, is adapted to 0.01 for Digits-DG and PACS, 0.02 for Office-Home, and 0.05 for DomainNet experiments.

Comparative Methodologies Our research approach is systematically juxtaposed against an ensemble of contemporary domain generalization methodologies, including but not limited to DeepAll Zhou et al. (2020a), JiGen Carlucci et al. (2019), and CCSA Motiian et al. (2017). All comparative outcomes are presented herein based on the mean accuracy computed from four iterative executions, with detailed references to the respective primary literature sources provided for these results.

Digits-DG Analysis The results pertaining to our methodology's performance on the Digits-DG dataset are exhaustively cataloged in Table 2, as shown in appendix. Our approach distinctly outperforms the baseline methods, surpassing the current state-of-the-art by an average margin of 0.74%,



(a) DG_via_ER (b) CloCls (c) DG_via_ER (d) CloCls

Figure 3: The confusion matrices pertaining to the DG_via_ER and ACADG methodologies, as applied to the PACS dataset, are presented herein. The matrices are delineated for the distinct domains, namely "Photo" and "Art," and are arranged from left to right for comprehensive examination.

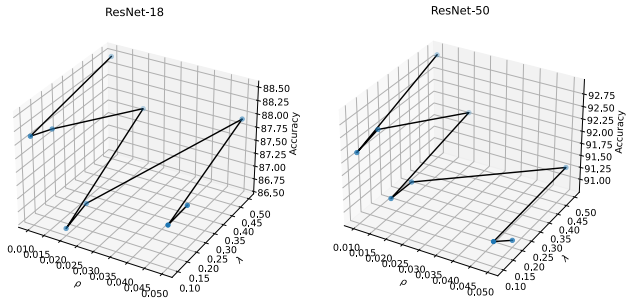


Figure 4: Parameter sensitivity of ACADG to hyper-parameter ρ and λ.

thereby underscoring its intrinsic capacity to synthesize highly generalizable and transferable feature representations for the task of domain generalization.

PACS Evaluation The results, as summarized in Table 3, as shown in appendix, provide a comprehensive insight into the exceptional performance achieved by our methodology. Notably, despite the inherent advantages of SADML Wang et al. (2022) owing to its similarity to the pre-trained ImageNet dataset, our ACADG method exhibits superior performance in domains such as Art and Sketch

Office-Home Benchmark The outcomes, delineated in Table 4, as shown in appendix, serve to highlight the consistent superiority of our proposed approach across a diverse array of tasks, underscoring its effectiveness in the synthesis of diverse datasets through strategic sample augmentation strategies.

DomainNet Dataset Analysis The outcomes stemming from the DomainNet dataset Peng et al. (2019) are succinctly presented in Table 5, as shown in appendix. Our methodology consistently elevates the generalization performance of SADML-based techniques. It is noteworthy that our approach attains the highest mean accuracy of 46.76%, representing a notable improvement of 1.38%.

4.3 Single-Source Domain Generalization

The results of single-source domain generalization, employing ResNet18 on the PACS benchmark, are delineated in Table 1, as shown in appendix. In this particular scenario, a solitary domain is chosen as the source dataset, and the model is subsequently evaluated on a variety of target domains.

5 Empirical Analysis

Ablation Study. Within this section, we undertake an in-depth exploration of the individual components that constitute the holistic framework of ACADG. Specifically, we focus on two pivotal

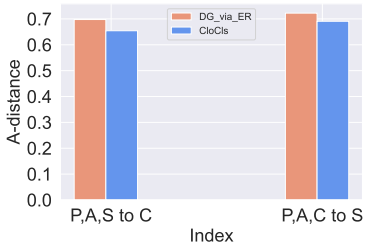


Figure 5: $\mathcal{A}$-distance on learned features for different domain generalization tasks.

elements: the incorporation of an Auxiliary Classifier for adversarial learning and the integration of Sharpness-Aware Minimization (SAM) as an optimizer. Table 6, as shown in appendix, presents a comprehensive analysis of their effects. Notably, upon the removal of both the Auxiliary Classifier and SAM from the overall loss function, the average accuracy of the ACADG method on the PACS dataset experiences a decline of 6.79%. This decline underscores the pivotal role played by ACADG and underscores the profound significance of diversification. Furthermore, it is noteworthy that even in the absence of SAM, ACADG demonstrates superior performance when compared to the majority of existing methodologies, thereby reinforcing the advantages offered by our auxiliary classifier.

Analysis of Confusion Matrices. Figure 3 visually represents the confusion matrices pertaining to the DG_via_ER method, specifically in the context of the PACS dataset. Evidently, the integration of ACADG significantly mitigates instances of erroneous predictions, particularly evident in challenging generalization tasks such as those involving "Cartoon" and "Sketch" domains. This observation underscores the pivotal role played by the auxiliary classifier within the adversarial domain generalization framework.

Parameter Sensitivity Analysis. In Figure 4, we present an analysis of the sensitivity of the ACADG method to hyperparameters, namely ρ and λ. Notably, ρ is allowed to vary within the range of $0.01, 0.02, 0.05$, while λ undergoes changes within the range of $0.1, 0.2, 0.5$. The results exhibit a clear trend, demonstrating that ACADG attains its optimal performance when $\rho = 0.01$ and $\lambda = 0.5$, regardless of whether ResNet-18 or ResNet-50 serves as the backbone architecture. This observation reaffirms the robustness and stability of our proposed method.

Quantification of Domain Divergence. The assessment of domain dissimilarity is conducted through the utilization of the $\mathcal{A}$-distance, a well-established metric for gauging distribution divergence Ben-David et al. (2010). The $\mathcal{A}$-distance serves as a widely-adopted indicator for quantifying the disparity between distributions. Specifically, the $\mathcal{A}$-distance, denoted as $\mathcal{A}_{dis}$, is defined as $\mathcal{A}_{dis} = 2(1 - \epsilon)$, where ϵ represents the test error of a classifier specifically trained to discriminate between the source and target domains. A smaller $\mathcal{A}$-distance value reflects a more effective alignment of distributions. As demonstrated in Figure 5, our proposed methodology exhibits superior capabilities in learning invariant features, thereby minimizing the divergence between the source and target domains in comparison to extant methodologies.

6 Concluding Remarks

In this research endeavor, our primary objective has been the acquisition of domain-invariant conditional distribution, a goal beyond the reach of basic adversarial learning-based solutions. Thoroughly investigating the limitations inherent in prior works, we have introduced an auxiliary classifier as a critical component of our approach. Through the optimization of the proposed regularization term, coupled with the auxiliary classifier and the domain adversarial loss, our framework excels in the generation of domain-invariant features, thereby enhancing its generalization capacity. Empirical findings, garnered from evaluations on both simulated and real-world datasets, unequivocally affirm the efficacy and potency of our proposed methodology.

REFERENCES

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *NeurIPS*, 31, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Guanyu Cai, Yuqin Wang, Lianghua He, and MengChu Zhou. Unsupervised domain adaptation with adversarial residual transform networks. *IEEE transactions on neural networks and learning systems (TNNLS)*, 31(8):3073–3086, 2019.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *arXiv preprint arXiv:2102.08604*, 2021.

Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *ECCV*, pp. 301–318, 2020.

Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *CVPR*, pp. 7119–7129, 2022.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ICLR*, 2021.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: beyond sharp and flat local minima. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 2553–2564, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Liang Hou, Qi Cao, Huawei Shen, Siyuan Pan, Xiaoshuang Li, and Xueqi Cheng. Conditional gans with auxiliary discriminative classifier. In *International Conference on Machine Learning*, pp. 8888–8902. PMLR, 2022.

Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision (ECCV)*, pp. 124–140, 2020a.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, pp. 124–140, 2020b.

Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *ICCV*, pp. 9619–9628, 2021.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

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*, pp. 5542–5550, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018a.

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*, pp. 5400–5409, 2018b.

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)*, pp. 624–639, 2018c.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Ao Ma, Jingjing Li, Ke Lu, Lei Zhu, and Heng Tao Shen. Adversarial entropy optimization for unsupervised domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. *arXiv preprint arXiv:2006.07500*, 2020.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, pp. 5715–5725, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, pp. 8690–8699, 2021.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

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 International Conference on Computer Vision*, pp. 1406–1415, 2019.

Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. A closer look at smoothness in domain adversarial training. In *International Conference on Machine Learning*, pp. 18378–18399. PMLR, 2022.

Seonguk Seo, Yumin Suh, D. Kim, Jongwoo Han, and B. Han. Learning to optimize domain specific normalization for domain generalization. In *European Conference on Computer Vision (ECCV)*, 2020.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

Mengzhu Wang, Jianlong Yuan, Qi Qian, Zhibin Wang, and Hao Li. Semantic data augmentation based distance metric learning for domain generalization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3214–3223, 2022.

Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*, pp. 159–176, 2020.

Zengmao Wang, Bo Du, and Yuhong Guo. Domain adaptation with neural embedding matching. *IEEE transactions on neural networks and learning systems (TNNLS)*, 31(7):2387–2397, 2019.

Xiaofu Wu, Suofei Zhang, Quan Zhou, Zhen Yang, Chunming Zhao, and Longin Jan Latecki. Entropy minimization versus diversity maximization for domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pp. 14383–14392, 2021.

Fu-En Yang, Yuan-Chia Cheng, Zu-Yun Shiau, and Yu-Chiang Frank Wang. Adversarial teacher-student representation learning for domain generalization. *Advances in Neural Information Processing Systems*, pp. 19448–19460, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR*, 2017.

Lei Zhang, Jingru Fu, Shanshan Wang, David Zhang, Zhaoyang Dong, and CL Philip Chen. Guide subspace learning for unsupervised domain adaptation. *IEEE transactions on neural networks and learning systems (TNNLS)*, 31(9):3374–3388, 2019.

Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33: 16096–16107, 2020.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13025–13032, 2020a.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, pp. 561–578, 2020b.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *ICLR*, 2021.

A Appendix

Table 1: Results of single source domain generalization on PACS. Each row and column indicates the source and target domain, respectively. We report the accuracy with the absolute gain from baseline in brackets. Positive gains is colored green.

	Photo	Art	Cartoon	Sketch
Photo	99.64 (+0.00)	64.28 (+1.39)	23.17 (+1.64)	55.63 (+3.97)
Art	97.24 (+1.64)	99.37 (+0.00)	68.43 (+2.01)	54.25 (+1.55)
Cartoon	85.89 (+0.23)	71.23 (+0.08)	99.13 (+0.00)	71.25 (+1.83)
Sketch	42.35 (+1.63)	44.37 (+1.27)	61.29 (+1.85)	99.37 (+0.00)

Table 2: Results following a leave-one-domain-out approach on the Digits-DG are presented. The optimal outcome within this set is emphasized in bold for clear delineation.

Methods	MNIST	MNIST-M	SVHN	SYN	Avg.
DeepAll Zhou et al. (2020a)	95.80	58.80	61.70	78.60	73.70
Jigen Carlucci et al. (2019)	96.50	61.40	63.70	74.0	73.90
CCSA Motiian et al. (2017)	95.20	58.20	65.50	79.10	74.50
MMD-AAE Li et al. (2018b)	96.50	58.40	65.00	78.40	74.60
CrossGrad Shankar et al. (2018)	96.7	61.1	65.3	80.2	75.8
MixStyle Zhou et al. (2021)	96.5	63.5	64.7	81.2	76.5
DDAIG Zhou et al. (2020b)	96.60	64.10	68.60	81.00	77.60
L2A-OT Zhou et al. (2020a)	96.70	63.90	68.60	83.20	78.10
FACT Xu et al. (2021)	97.90	65.60	72.40	90.30	81.50
COMEN Chen et al. (2022)	97.10	67.60	75.10	91.30	82.30
SADML Wang et al. (2022)	98.90	68.20	74.30	92.50	83.50
ACADG	**99.11**	**69.24**	**75.42**	**93.24**	**84.25**

Table 3: Results derived utilizing a leave-one-domain-out methodology on PACS are elucidated below. The superior result within this context is highlighted in bold for unequivocal distinction.

Methods	Art	Cartoon	Photo	Sketch	Avg.
ResNet18					
DeepAll Zhou et al. (2020a)	77.63	76.77	95.85	69.50	79.94
MetaReg Balaji et al. (2018)	83.70	77.20	95.50	70.30	81.70
CrossGrad Shankar et al. (2018)	79.80	76.80	96.00	70.20	80.70
JiGen Carlucci et al. (2019)	79.42	75.25	96.03	71.35	80.51
DDAIG Zhou et al. (2020b)	84.20	78.10	95.30	74.70	83.10
L2A-OT Zhou et al. (2020a)	83.30	78.20	96.20	73.60	82.80
MixStyle Zhou et al. (2021)	84.10	78.80	96.10	75.90	83.70
EISNet Wang et al. (2020)	81.89	76.44	95.93	74.33	82.15
FACT Xu et al. (2021)	85.37	78.38	95.15	79.15	84.51
COMEN Chen et al. (2022)	82.60	81.00	94.60	84.50	85.70
SADML Wang et al. (2022)	87.96	**82.41**	**98.85**	83.21	88.08
ACADG	**88.48**	82.37	98.82	**84.75**	**88.61**
ResNet50					
DeepAll Zhou et al. (2020a)	84.94	76.98	97.64	76.75	84.08
MetaReg Balaji et al. (2018)	87.20	79.20	97.60	70.30	83.60
DDAIG Zhou et al. (2020b)	85.4	78.5	95.7	80.0	84.9
CrossGrad Shankar et al. (2018)	87.5	80.7	97.8	73.9	85.7
EISNet Wang et al. (2020)	86.64	81.53	97.11	78.07	85.84
ATSRL Yang et al. (2021)	90.00	83.50	98.90	80.00	88.10
FACT Xu et al. (2021)	89.63	81.77	96.75	84.46	88.15
SADML Wang et al. (2022)	92.43	84.94	99.13	88.29	91.20
ACADG	**92.90**	**85.13**	**99.21**	**90.38**	**91.91**

Table 4: The results, obtained through the application of a leave-one-domain-out strategy on Office-Home, are disclosed herein. The most exemplary result is accentuated in bold to ensure distinct recognition.

Methods	Art	Clipart	Product	Real	Avg.
DeepAll Zhou et al. (2020a)	57.88	52.72	73.50	74.80	64.72
CCSA Motiian et al. (2017)	59.90	49.90	74.10	75.70	64.90
MMD-AAE Li et al. (2018b)	56.50	47.30	72.10	74.80	62.70
CrossGrad Shankar et al. (2018)	58.40	49.40	73.90	75.80	64.40
MixStyle Zhou et al. (2021)	58.70	53.40	74.20	95.90	65.50
L2A-OT Zhou et al. (2020a)	60.60	50.10	74.80	77.00	65.60
FACT Xu et al. (2021)	60.34	54.85	74.48	76.55	66.56
SWAD Cha et al. (2021)	66.10	57.70	78.40	80.20	70.60
SADML Wang et al. (2022)	68.29	57.63	76.21	80.69	70.71
ACADG	**70.30**	**58.68**	**77.93**	**81.64**	**72.14**

Table 5: The outcomes, following the implementation of a leave-one-domain-out approach on DomainNet, are articulated below. The paramount result within this array is emphasized in bold, serving to facilitate unambiguous identification.

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sktetch	Avg.
C-DANN Li et al. (2018c)	54.60	17.30	43.70	12.10	56.20	45.90	38.30
RSC Huang et al. (2020b)	55.00	18.30	44.40	12.20	55.70	47.80	38.90
Mixup Zhang et al. (2017)	55.70	18.50	44.30	12.50	55.80	48.20	39.20
SagNet Nam et al. (2021)	57.70	19.00	45.30	12.70	58.10	48.80	40.30
MLDG Li et al. (2018a)	59.10	19.10	45.80	13.40	59.60	50.20	41.20
ERM Gulrajani & Lopez-Paz (2021)	58.10	18.80	46.70	12.20	59.60	49.80	40.90
MetaReg Balaji et al. (2018)	59.77	25.58	50.19	11.52	64.56	50.09	43.62
DMG Chattopadhyay et al. (2020)	65.24	22.15	50.03	15.68	59.63	49.02	43.63
SelfReg Kim et al. (2021)	62.40	22.60	51.80	14.30	62.50	53.80	44.60
SADML Wang et al. (2022)	63.72	23.36	51.92	15.93	63.01	54.34	45.38
ACADG	**65.92**	**24.69**	**52.66**	**16.91**	**64.80**	**55.65**	**46.77**

Table 6: We present an ablation study conducted on PACS, utilizing the ResNet-50 architecture, to investigate the contributory impact of the individual components and understand the synergistic effect of the combined elements within the model.

Methods	Auxiliary	SAM	Art	Cartoon	Photo	Sketch	Avg.
ACADG			87.11	78.65	98.22	76.48	85.11
ACADG		✓	87.29	79.30	98.73	77.14	85.62
ACADG	✓		91.35	84.37	99.11	89.72	91.14
ACADG	✓	✓	**92.90**	**85.13**	**99.21**	**90.38**	**91.91**