Augmentation-Invariant Learning Strategy via Data Augmentation for Improving Model Generalization

Yu Miao[1], Juanjuan Zhao[1,2*], Sijie Song[1], Ran Gong[3], Yuanqian Zhu[2], Lusha Qi[2], Yan Qiang[4]

[1]College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China
[2]School of Software, Taiyuan University of Technology, Taiyuan, China
[3]Huawei Technologies Co., Ltd, Nanjing, China
[4]School of Software, North University of China, Taiyuan, China
2023310112@link.tyut.edu.cn, zhaojuanjuan@tyut.edu.cn

Abstract

Data augmentation is an effective technique for regularizing deep networks, which helps to enhance the generalizability and robustness of the model. However, in the field of medical imaging, traditional data augmentation techniques such as cropping, rotation, and degradation may inadvertently alter the critical characteristics of pathological lesions. Conventional semantic augmentation methods, such as altering the color and contrast of the object background, may also affect the structural features of medical images in uncontrolled semantic directions. Such operational conditions compromise the model's diagnostic reliability in medical contexts. To address this issue, we propose a surprisingly efficient implicit augmentation-invariant learning strategy (AILS) via variational Bayesian inference on differentially constrained feature manifolds. Parameterizing probability measures over tangent space through deep networks enables precise estimation of semantic direction distributions. Subsequently, geodesic-aware semantic features are sampled from the reparameterized variational posterior, achieving semantic-consistent feature augmentation. Simultaneously, to mine augmentation distribution invariance, we design the AiHLoss, which constrains the augmentation distribution to facilitate the network to learn augmentation invariance. Extensive experiments demonstrate that AILS exhibits high performance on public medical image datasets, outperforming existing augmentation methods.

Code — https://github.com/miao-zi/AILS

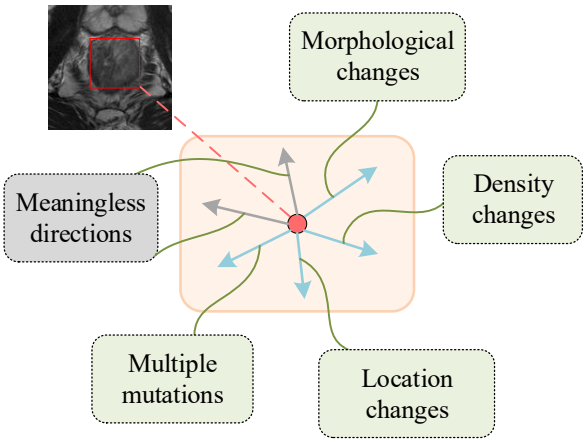

Figure 1: The schematic representation of semantic directions in deep feature space. The central red dot marker denotes the lesion region in the original image, with arrows oriented at distinct angular positions indicating different semantic transformation trajectories.

Introduction

The inadequacy of model generalization due to data scarcity (Upadhyay and Bhandari 2024) is a ubiquitous issue in the scientific community, particularly in the realm of medical imaging, where researchers frequently confront the challenge of "big data with small samples" (Kim, Geenjaar, and Calhoun 2024; Liu et al. 2024). Despite the rapid advancement of modern medical imaging technology, which has generated vast amounts of image data, the number of annotated samples available for practical model training remains relatively limited due to factors such as high annotation costs, domain uniqueness, and strong specialization. This significantly hampers the generalization performance of deep learning models in medical image analysis. Data augmentation is a highly effective technique for in-domain generalization, capable of substantially mitigating overfitting issues during the training of deep networks (El Jiani, El Filali et al. 2022). In the field of natural images, data augmentation techniques have been widely applied and have achieved remarkable results. However, when the focus shifts to medical images, the situation becomes more complex (Goceri 2023). Many explicit operational methods (Harris et al. 2020; Li et al. 2024; Zhang et al. 2018; Yun et al. 2019; DeVries and Taylor 2017; Zhong et al. 2020) that are effective in traditional natural image processing are no longer applicable in this domain.

In recent years, research efforts have explored implicit data augmentation to enhance network generalization with-

*Corresponding author

out generating new data. For instance, Wang et al. (Wang et al. 2021) proposed a semantic augmentation algorithm, where they introduced semantic linearization – a technique that enhances the semantic richness of training data by adjusting the feature representation along certain semantic directions. As illustrated in Figure 1, distinct semantic directions exist within the deep feature space. Applying effective sampling shifts to the research subject can generate diverse feature representations. In essence, the process effectively alters the object's background or perspective by exploring specific directions within the deep feature space that carry particular semantic meaning. However, while this algorithm pioneers the augmentation of semantic information, transformations along certain semantic directions, such as altering object background colors and contrast, may adversely affect the structural features of medical images in uncontrolled semantic directions. In the medical field, recent research focuses on semantic-preserving transformations, such as robust gradient learning (Chen et al. 2023), which constructs augmentation distributions using properties of reproducing kernel Hilbert spaces. However, its linear assumptions restrict its application in complex scenarios. Zhu et al. (Zhu et al. 2024) pioneered the foundational concept of Bayesian semantic data augmentation (BSDA), yet its constrained adaptability to Bayesian framework presents significant opportunities for enhancement. This latent potential not only reveals fertile ground for methodological refinement but catalyzes our development of an evolved paradigm addressing these gaps. Consequently, there is an urgent need for a feasible, efficient, and highly generalizable framework paradigm tailored to the data-scarce and challenging field of medical imaging.

Based on the established theoretical assumption (Upchurch et al. 2017) that deep features of networks often exhibit linear properties, translating along one of certain specific directions in the feature space will yield feature representations corresponding to another sample with the same category label but different semantics. Furthermore, inspired by implicit augmentation approaches, we can effectively enhance the feature diversity of the original training set samples by exploring numerous such semantic directions. In this paper, we propose an efficient implicit augmentation-invariant learning strategy (AILS) through manifold-constrained variational Bayesian augmentation. Without auxiliary generative networks or explicit sample generation, AILS introduces geometrically grounded "semantic bias", which can be understood as perturbations. First, leveraging mean-field theory, we parameterize probability measures on tangent spaces to estimate semantic direction distributions. Subsequently, geodesic-aware features are sampled via a specific indicator function and integrated into the feature pool. During inference, we identify semantic directions that may disrupt the structure of medical images and avoid zero bias to significantly reduce the likelihood of generating meaningless semantic transformations. We then optimize a reparameterized variational posterior to approximate the true distribution on manifold. This constructs a differentially constrained semantic augmentation space preserving anatomical integrity. To further optimize the variational distribution within the semantic augmentation space, we derive a closed-form upper bound on the expected cross-entropy (CE) loss for variational Bayesian inference in limiting cases. This upper bound exhibits a monotonic relationship with the negative evidence lower bound (ELBO). Therefore, minimizing it directly promotes the maximization of the ELBO, which essentially constitutes a novel robust surrogate loss function.

Augmentation invariance is a crucial property of the data distribution, reflecting the intrinsic relationship between data transformations and the original samples. Research (Surner, Khelil, and Bothmann 2025; Chen et al. 2019) demonstrates that networks with the ability to learn augmentation invariance exhibit enhanced robustness to variations in augmentation parameters, thereby showing superior generalization performance. To constrain the augmented distribution and facilitate network learning of augmentation invariance (Zhang and Ma 2022), we design a hybrid loss function, AiHLoss, based on robust surrogate loss. Specifically, since AILS performs class identity-preserving semantic transformations, this should not affect the model's predictions for categories. Therefore, while performing data augmentation, we minimize the Kullback-Leibler (KL) divergence between the predictions of augmented features and original features.

Our contributions can be summarized as follows: (1) An efficient implicit augmentation-invariant learning strategy (AILS) is proposed. It applies manifold-constrained variational Bayesian inference to medical images, ensuring anatomical plausibility. (2) A novel composite surrogate loss function (AiHLoss) is designed. It incorporates a Bayesian discriminative boundary for robustness and a bidirectional semantic divergence term for consistency. (3) Extensive experiments demonstrate that networks trained using our algorithm indeed acquire better augmentation transferability and robustness.

Method

Semantic Transformation in Deep Feature Space

In the realm of medical imaging, the concept of "semantics" transcends intuitive descriptions of image content to encompass a profound comprehension of the underlying biological processes, pathological traits, and their clinical implications (Qureshi et al. 2023). The semantic information embedded within these images ranges from the identification of fundamental anatomical structures to the evaluation of intricate pathological conditions, mirroring an abstract progression from low-level visual features to high-level medical cognition (Song and Chong 2024). Deep networks are adept at extracting high-level representations during the convolutional process, thereby constructing a structured "deep feature space" (Huang et al. 2019). Within this space, each feature vector corresponds to a high-order semantic representation derived from the original medical images, with the directional shifts of vectors within the space reflecting specific semantic transformations (Chao, Zhang, and Yan 2022).

Figure 2 illustrates the schematic representation of the categorical distribution and directional migration of seman-

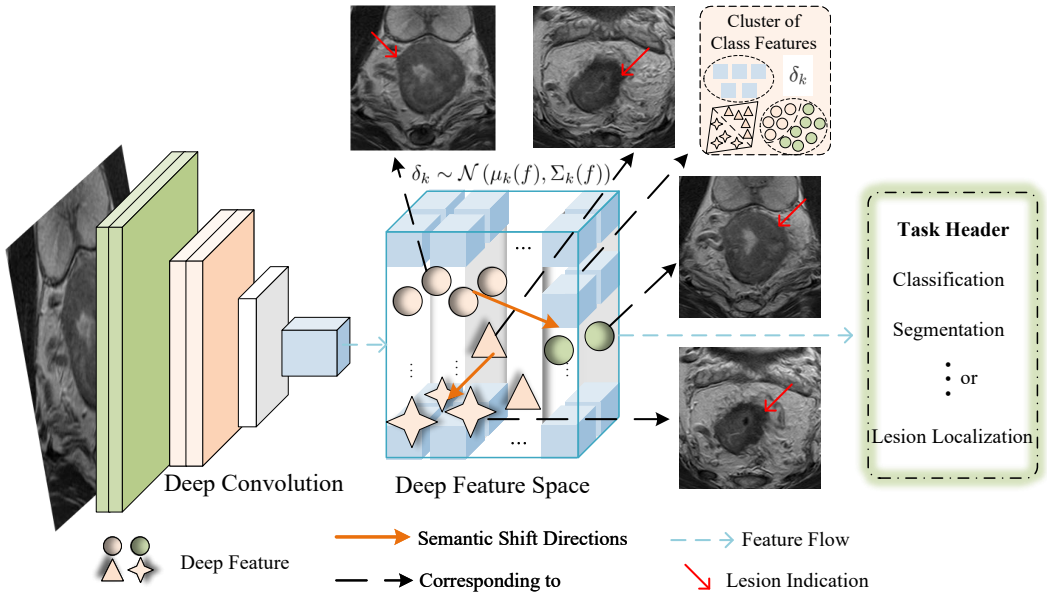

Figure 2: The schematic diagram of categorical distribution and orientation shift characteristics for semantic feature clusters in deep convolutional feature space of medical images.

tic feature clusters in the deep convolutional feature space of medical images. Building upon the theoretical framework of spatial transformation, we can achieve controllable manipulation of semantic information in medical images by adjusting the spatial positioning and transformation intensity of feature vectors.

Given an original input x and a deep convolutional network G, the extracted raw features are denoted as $f = G(x)$. The augmented features f' after undergoing semantic transformation are then obtained:

$$f' = T(f; \delta; \alpha) \tag{1}$$

where $T(\cdot; \delta; \alpha)$ denotes the semantic transformation function defined by the parameters δ and α. δ represents the direction of semantic augmentation and α denotes the magnitude of semantic augmentation.

The transformation T can be further decomposed into specific operations:

$$f' = T(f; \delta; \alpha) = f + \alpha * d_\delta * f \tag{2}$$

where $d \in \{0, 1\}^k$ is a binary vector initially set to all ones, and its final value is then determined based on Bayesian probabilities. The symbol $*$ denotes element-wise multiplication.

Semantic Direction Sampling Based on Bayesian Inference

Based on the assumption of linear semantic properties within the deep feature space, we propose a method for semantic direction inference. This method aims to implicitly construct a diverse semantic augmentation space without the need for explicitly generating samples. The core of the method presented in this section lies in estimating the probability distribution of potential semantic directions in the feature space through Bayesian inference and optimizing the enhanced feature generation process based on variational inference. Figure 3 provides a detailed exposition of the theoretical underpinnings of the proposed algorithm. AILS establishes a framework-level continual learning strategy by systematically integrating Bayesian inference within a geometrically constrained manifold $\mathcal{M} \subseteq R^d$. At its core, this section presents the process of transforming semantic directions into direction-aware random variables.

Probabilistic Modeling of Semantic Directions on $\mathcal{M}$. Given a primitive feature vector $f \in \mathcal{M}$, we model its corresponding semantic direction δ as residing within the tangent space $T_f\mathcal{M}$ at f. We posit that δ follows an implicit conditional probability distribution $p(\delta|f)$. To facilitate inference under the manifold constraint, we leverage mean-field theory, decomposing the joint distribution into independent components along latent semantic axes.

$$p(\delta \mid f) = \prod_{k=1}^{K} p\left(\delta_k \mid f\right) \tag{3}$$

where K represents the number of latent semantic directions. To balance tractability with expressive power, we assume each component δ_k follows a Gaussian distribution parametrized by deep networks, respecting the local geometry of $\mathcal{M}$.

$$\delta_k \sim \mathcal{N}\left(\mu_k(f), \Sigma_k(f)\right) \tag{4}$$

where $\mu_k(f) \in T_f\mathcal{M}$ is the mean direction, and $\Sigma_k(f)$ is a positive semi-definite covariance operator acting on $\mathcal{M}$.

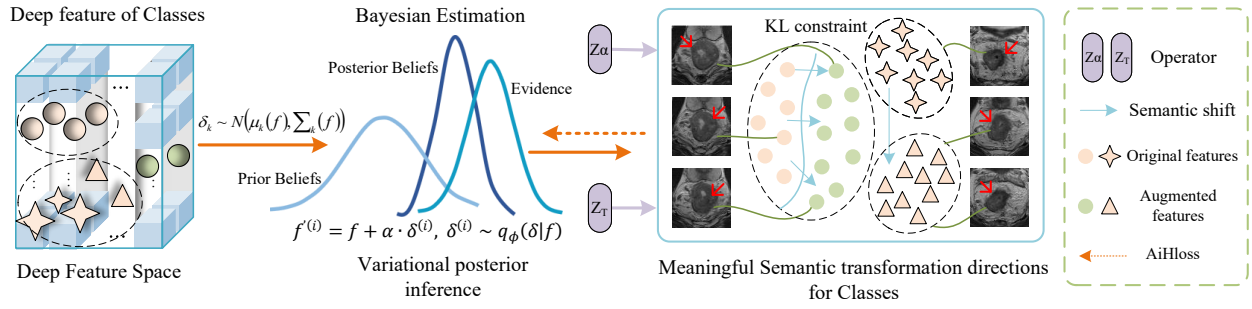

Figure 3: The detailed illustration of the underlying theoretical foundation of the proposed AILS.

This parameterized form allows the statistical characteristics of semantic directions to be learned implicitly through deep networks.

Variational Posterior Inference. The true posterior $p(\delta|f)$ is difficult to solve directly. We therefore introduce a variational family of distributions $q_\phi(\delta|f)$ defined on $T_f\mathcal{M}$ to approximate it, where ϕ is the variational parameter. According to the variational Bayesian framework, the optimization objective is to maximize the Evidence Lower Bound (ELBO), which links Bayesian inference to generalization performance:

$$\mathcal{L}_{ELBO} = E_{q_\phi(\delta|f)}[\log p(y \mid f+\delta)] - \mathrm{KL}\left(q_\phi(\delta \mid f) \| p(\delta)\right) \tag{5}$$

where $p(y|f+\delta)$ is a classification-like function, and $p(\delta)$ is the prior distribution (set as an isotropic Gaussian). Through the reparameterization trick, the gradient can be propagated to the variational parameter ϕ , to achieve end-to-end optimization.

Incorporation of Robust Augmentation Features. In the training stage, based on the inferred variational distribution $q_\phi(\delta|f)$, we sample the semantic offset $\delta^{(i)}$ to generate augmented features:

$$f_{aug}^{(i)} = f + \alpha \cdot \delta^{(i)}, \quad \delta^{(i)} \sim q_\phi(\delta \mid f) \tag{6}$$

where α is the intensity coefficient, controlling the amplitude of the semantic offset. To ensure the semantic validity and manifold consistency of the augmentation, thereby further improving robustness and preventing detrimental perturbations, we introduce a semantic effectiveness constraint based on the manifold geometry:

$$C(\delta) = \mathbf{1}\{sim(\Pi_M(f), \Pi_M(f+\alpha\delta)) \geq \tau\} \tag{7}$$

where $\mathbf{1}(\cdot)$ is the indicator function and $sim(\cdot)$ is the cosine similarity metric. The threshold τ controls the semantic consistency between augmented and original features. Initially derived from feature distribution statistics, it undergoes dynamic decay during training, ensuring a principled trade-off between augmentation diversity and stability while rigorously preserving anatomical plausibility. Ineffective offsets will be removed to avoid the generation of meaningless semantic transformations.

AiHLoss

As the cornerstone of our AILS, the Augmentation-invariant Hybrid Loss (AiHloss) function provides the essential theoretical and operational framework for enforcing semantic consistency across augmented views. By jointly optimizing variational Bayesian feature augmentation and cross-view distributional alignment, it constitutes the critical control mechanism that binds the entire learning system into a cohesive, invariant architecture.

Progressive Optimization of the Variational Distribution. To achieve fine-grained control of the augmentation space, we derive the upper bound of the cross-entropy loss:

$$\mathcal{L}_{CE-UB} = E_{q_\phi(\delta|f)}[-\log p(y \mid f+\delta)] + \beta \cdot KL\left(q_\phi||p\right) \tag{8}$$

where β is the weighting coefficient. By minimizing $\mathcal{L}_{CE-UB}$, we indirectly optimize the discriminative ability of augmented features while constraining the deviation of the variational distribution from the prior. Through theoretical analysis, when $\beta \to 1$, the upper bound has a monotonic relationship with the ELBO, ensuring the stability of the optimization process.

Augmentation invariant Hybrid Loss Function. For two semantic-features f_1' and f_2' derived from the same sample (sharing label), let p_1 and p_2 denote their prediction probability vectors respectively. To enforce feature consistency across different augmented views of the same sample in neural networks, a loss function that effectively captures the similarity between feature distributions rather than merely point-wise distances is required. Kullback-Leibler (KL) divergence not only accounts for the disparity between distributions but also provides a more comprehensive measure of distributional alignment in feature space. This property makes it particularly suitable for constraining semantic consistency in high-dimensional feature representations. Consequently, we adopt symmetric bidirectional KL divergence to quantify the similarity across multiple probability distributions, though we primarily analyze the pairwise case for

Methods	Blood	Chest	Breast	Derma	OCT	Path	Pneumonia	Retina	Tissue
Mixup	87.3±0.8	77.6±0.7	77.8±0.9	70.3±1.3	80.5±1.2	83.4±0.8	81.2±0.9	53.4±1.1	70.7±0.7
CutMix	87.8±0.7	78.1±0.6	78.3±0.8	70.8±1.2	81.0±1.1	83.9±0.7	81.7±0.8	54.0±1.0	71.2±0.6
ADA	88.4±0.9	78.7±0.5	78.9±0.7	71.4±1.1	81.6±1.0	84.5±0.6	82.3±0.7	54.6±0.9	71.8±0.5
DDPM	89.2±0.8	79.5±0.4	79.7±0.6	72.2±1.0	82.4±0.9	85.3±0.5	83.1±0.6	55.4±0.8	72.6±0.4
AdvProp	90.1±0.7	80.4±0.3	80.6±0.5	73.1±0.9	83.3±0.8	86.2±0.4	84.0±0.5	56.3±0.7	73.5±0.3
MedAugment	91.2±0.6	81.5±0.2	81.7±0.4	74.2±0.8	84.4±0.7	87.3±0.3	85.1±0.4	57.4±0.6	74.6±0.2
BSDA	98.8±0.1	86.4±0.5	86.1±1.5	76.4±0.8	88.8±1.3	91.9±3.2	88.8±1.2	57.2±0.1	73.3±1.0
AILS(Ours)	**98.9±0.2**	**87.7±0.4**	**87.7±0.8**	**79.8±2.3**	**90.8±1.1**	**94.6±0.5**	**92.2±0.1**	**60.6±1.0**	**79.3±0.2**

Table 1: Comparing ACC (%) performance with state-of-the-art methods.

clarity. Formally:

$$\begin{aligned}\mathcal{L}_{inv}(p_1, p_2) &= KL(p_1 \| p_2) + KL(p_2 \| p_1) \\ &= \sum_{k=1}^{c} \left(p_1^k \log \frac{p_1^k}{p_2^k} + p_2^k \log \frac{p_2^k}{p_1^k} \right)\end{aligned} \tag{9}$$

where $\mathcal{L}_{inv}$ measures the logarithmic difference between p_1 and p_2, and it is always non-negative, with $L(p_1, p_2) = 0$ if and only if $p_1 = p_2$. This kind of difference belongs to redundant augmentation and can be regarded as an ineffective shift. Consequently, employing it as the training objective directly drives the model to learn augmentation-invariant feature representations, which is crucial for improving generalization performance.

Finally, we jointly use $\mathcal{L}_{CE-UB}$ and $\mathcal{L}_{inv}$ as the constraints of the mixed loss function in the entire network training process:

$$\mathcal{L}_{AiH} = \mathcal{L}_{CE-UB} + \lambda \cdot \mathcal{L}_{inv} \tag{10}$$

where λ is a hyperparameter that adjusts the role of the augmentation invariance constraint term in the overall loss function. $\mathcal{L}_{CE-UB}$ ensures the model's discriminative power does not degrade under augmentation perturbations, while $\mathcal{L}_{inv}$ ensures this discriminative power is built upon core semantic features robust to these augmentations.

AiHLoss constitutes the central optimization mechanism and core constraint framework of the algorithm AILS. Optimizing it enables the model to simultaneously learn highly discriminative features while maintaining invariance to predefined augmentation transformations. This significantly enhances the overall robustness and generalization ability of the model, forming a complete, loss-driven augmentation-invariant learning framework. We present the pseudocode of AILS in Algorithm 1. The rigorous mathematical derivation of our AILS algorithm can be found in **Appendix**.

Algorithm 1: The AILS Algorithm

Input: Training dataset D, hyperparameters $\alpha, \tau, \beta, \lambda$ network parameters Θ, variational parameters ϕ
Output: Trained network parameters Θ^* and variational parameters ϕ^*

Initialize Θ and ϕ
for $t = 0$ to T **do**
  Sample a mini-batch $\{(x_i, y_i)\}_{i=1}^{B}$ from D
  Extract deep features $\mathbf{f}_i = G(x_i; \Theta)$
  Bayesian manifold modeling, estimate $\mu_i, \Sigma_i = \text{ManifoldHead}(f_i; \phi)$ $(\mu_i \in T_{f_i}\mathcal{M}, \Sigma_i \succ 0)$
  Compute augmented features $\delta_i = \mu_i + \Sigma_i^{1/2} \odot \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, I)$; $\mathbf{f}_i' = \text{Vector}_{f_i}(\alpha\delta_i) = \mathbf{f}_i + \alpha\delta_i$
  Filter invalid offsets δ_i $(sim(\Pi_M(\mathbf{f}_i), \Pi_M(\mathbf{f}_i')) < \tau)$
  Compute cross-entropy loss $\mathcal{L}_{\text{CE}} = \frac{1}{B}\sum_{i=1}^{B} \text{CE}(\text{Classifier}(\mathbf{f}_i'), y_i)$
  Compute KL divergence regularization term $\mathcal{L}_{KL} = \frac{1}{B}\sum_{i=1}^{B} \text{KL}(q(\delta_i|\mathbf{f}_i) \| p(\delta_i))$
  Compute bidirectional divergence constraint $\mathcal{L}_{inv} = \frac{1}{B}\sum_{i=1}^{B} (\text{KL}(p_1 \| p_2) + \text{KL}(p_2 \| p_1))$
  Compute total loss $\mathcal{L}_{AiH} = \mathcal{L}_{CE} + \beta\mathcal{L}_{KL} + \lambda\mathcal{L}_{inv}$
  Update parameters $\Theta, \phi \leftarrow \text{Adam}(\nabla_{\Theta,\phi}\mathcal{L}_{AiH})$
end for
return Θ^*, ϕ^*

Experiments and Results

Datasets and Settings

We conducted comprehensive experiments on the publicly available MedMNIST+ (Yang et al. 2023), a standardized benchmark for biomedical image computing. This dataset comprises 12 pre-processed 2D datasets and 6 pre-processed 3D datasets, encompassing diverse biomedical imaging modalities, task types, and data scales. To rigorously evaluate the proposed AILS, nine representative 2D subsets are selected (PathMNIST, ChestMNIST, DermaMNIST, OCTMNIST, PneumoniaMNIST, RetinaMNIST, BreastMNIST, BloodMNIST, TissueMNIST; MNIST suffix omitted for brevity in subsequent text). As detailed in Table 3, these subsets cover seven different medical imaging modalities, collectively containing 600,338 image samples. Each subset is pre-partitioned into training, validation, and test sets, with the specific sample counts for each split provided the table. In this study, we leverage a combination of ResNeSt-101 (Zhang et al. 2022) and DenseNet-121 (Huang et al. 2017) as the backbone. Traditional augmentation methods (Mixup (Zhang et al. 2018), CutMix (Yun et al. 2019)), explicit generative model-based method (StyleGAN2-ADA

Backbones	Baseline Accuracy(%)	+AILS	Increase(%)	FDI of AILS	Epochs to convergence
MobileNetV3	70.1±0.9	76.3±0.7	**+6.2**	1.21	$78 \rightarrow 62$
ResNeSt-101	80.3±0.6	85.1±0.5	**+4.8**	1.42	$65 \rightarrow 53$
DenseNet-121	81.9±0.5	85.6±0.4	**+3.7**	1.39	$68 \rightarrow 55$
EfficientNet-B4	82.7±0.6	86.2±0.5	**+3.5**	1.51	$60 \rightarrow 48$
Swin-Tiny	80.5±0.7	84.4±0.6	**+3.9**	1.56	$80 \rightarrow 64$

Table 2: Cross-architecture adaptability of AILS algorithm: accuracy improvements and training efficiency.

Name	Data Modality	Training/Validation/Test
Path	Colon Pathology	89,996/10,004/7,180
Chest	Chest X - Ray	78,468/11,219/22,433
Derma	Dermatoscope	7,007/1,003/2,005
OCT	Retinal OCT	97,477/10,832/1,000
Pneumonia	Chest X - Ray	4,708/524/624
Retina	Fundus Camera	1,080/120/400
Breast	Breast Ultrasound	546/78/156
Blood	Microscope	11,959/1,712/3,421
Tissue	Microscope	165,466/23,640/47,280

Table 3: Detailed information of experimental datasets.

(Karras et al. 2020)), Diffusion-based (DDPM (Ho, Jain, and Abbeel 2020)), implicit augmentation method AdvProp (Xie et al. 2020), and the medical-specific methods (MedAugment (Liu et al. 2023), BSDA (Zhu et al. 2024)) are selected as comparative baselines. To mitigate the effects of stochastic bias in model selection, we conduct ten repeated trials for each experimental configuration and report both mean performance metrics and standard deviations. We implement all experiments in PyTorch and train it on the Huawei Cloud AI Training Service, utilizing Ascend 910B clusters. During the training phase, we use the AdamW (Loshchilov and Hutter 2019) optimizer with an initial learning rate of 0.01 and implement a learning rate warm-up strategy for the first five epochs to stabilize the dynamics at the initial training stage.

Additionally, to achieve "soft augmentation", we incorporate a temperature coefficient T during network inference, ensuring that AiHLoss term does not overlook categories with probabilities close to zero. T governs the dynamic threshold for filtering ineffective semantic shifts. Early in training, semantic safety is prioritized (T=0.8), while later stages explore more refined feature manifolds (T=0.6). Experimental results show that the optimal value of λ is 0.5.

Comparison Experiments

Verification of Generalization Ability in Cross-Modal Tasks. Table 1 summarizes the classification accuracy (ACC) of various data augmentation methods across CT, ultrasound, and X-ray modalities. The data presented indicate that AILS consistently outperforms all baseline methods, achieving an average accuracy of 85.74% on 2D datasets. Notably, AILS achieves an average accuracy improvement of 7.39% compared to explicit methods (ADA, DDPM), demonstrating that avoiding data generation in pixel space effectively reduces the risk of anatomical distortion.

The paired t-test confirm that the performance improvement of AILS is not a random occurrence, but rather a systematic enhancement with statistical significance ($p < 0.01$). For instance, on the OCT dataset, it achieves a notably high accuracy of 90.8%, which strongly provides compelling evidence for this. Collectively, these results provide both theoretical and empirical validation for the efficacy of the AILS.

Verification of Data Efficiency. To systematically evaluate data efficiency, we conduct subsampling experiments on the original 2D training dataset at three distinct ratios (20%, 30%, and 50%), simulating real-world scenarios with limited data availability. Subsequently, our AILS is comprehensively compared against seven algorithms on these subsampled datasets to assess their robustness under varying data scales, ranging from extreme to moderate scarcity conditions. As evidenced in Table 4, the experimental results

Method | Ratio	20%	30%	50%
Mixup	55.3±3.1	60.1±1.8	68.2±1.6
CutMix	57.8±2.4	62.4±1.9	70.5±1.8
StyleGAN2-ADA	54.2±7.3	61.7±5.2	72.3±3.9
DDPM	58.0±3.2	63.9±2.3	73.6±2.2
AdvProp	59.6±2.9	64.2±2.6	71.8±1.7
MedAugment	60.1±1.5	65.3±1.3	73.8±1.2
BSDA	67.2±1.7	70.0±2.2	74.1±1.2
AILS	**69.4±3.2**	**75.5±2.1**	**81.8±1.7**

Table 4: Average classification ACC (%) under different training data ratios.

demonstrate a consistent performance degradation across all compared algorithms when facing data scarcity. Notably, AILS exhibits significantly less performance deterioration compared to other methods. This compelling evidence validates the effectiveness of our feature-space augmentation mechanism in mitigating overfitting with limited samples. We attribute AILS's superior data efficiency to its variational feature enrichment mechanism. By sampling directions from the learned distribution $q_\phi(\delta|f)$, it generates diverse yet anatomically plausible feature variations, thereby effectively expanding the training distribution without requiring additional labeled data.

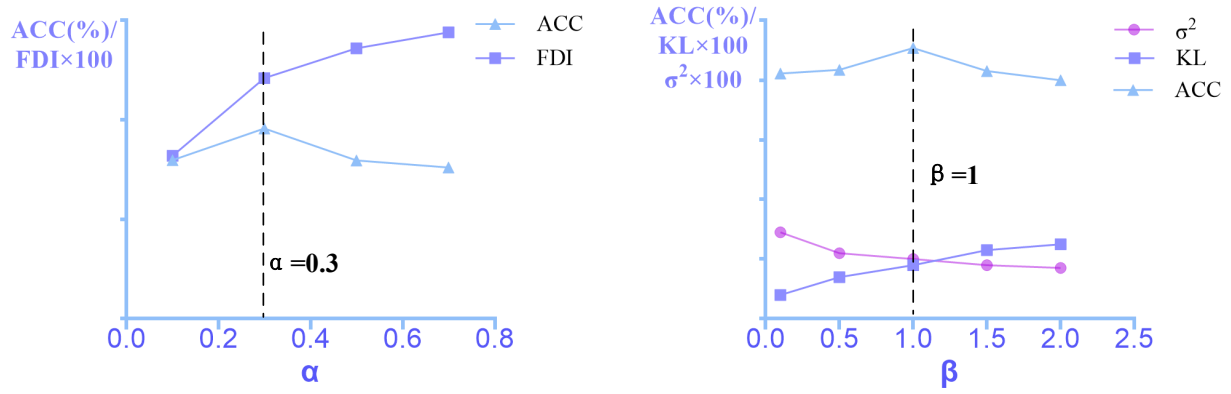

Figure 4: Line graph of experimental data for hyperparameter analysis. The left subfigure reflects the augmentation intensity α, while the right subfigure reflects the trade-off parameter β.

Ablation Studies

Analysis of the Augmentation Intensity α. The augmentation intensity α controls the degree of semantic shift in the feature space. To analyze the impact of the parameter α, an evaluation is performed in $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$ while keeping other hyperparameters fixed ($\beta = 1.0, \lambda = 0.5$). We quantify the impact of α through three metrics: classification accuracy (ACC), feature diversity index (FDI), and distribution alignment (KL). The formula for the feature diversity index is expressed as: $FDI = \frac{1}{N}\sum_{i=1}^{N} \|f_i' - f_i\|_2$, where f represents the original features, while f' denotes the augmented features.

As shown in Table 5, the best average ACC occurs at the moderate level of FDI, while striving to strike a balance between diversity and consistency. Therefore, AILS achieves the optimal performance when $\alpha = 0.3$. It is evident that when $\alpha > 0.5$, an excessive semantic shift will lead to feature misalignment; when $\alpha < 0.1$, it will result in insufficient augmentation. The left subfigure in Figure 4 provides an intuitive visualization of ACC metric variations across different α values, demonstrating a clear parametric dependence.

Values of α	ACC	AUC	FDI	KL	selected
0.1	79.84±1.6	83.33±3.1	0.82	–	
0.3	85.74±0.5	92.97±2.4	1.21	0.18	✓
0.5	79.64±1.4	82.89±3.3	1.36	–	
0.7	76.04±2.7	80.46±1.1	1.44	–	

Table 5: Impact of augmentation intensity α on algorithm performance.

Analysis of the KL Trade-off Parameter β. The parameter β is utilized to balance the cross-entropy loss and KL regularization in the $\mathcal{L}_{CE-UB}$. We evaluate $\beta \in \{0.1, 0.5, 1.0, 1.5, 2\}$ on nine 2D datasets. Evaluation metrics include the average ACC, the KL divergence between the variational distribution and the prior distribution, and the variance σ^2 of the augmented features.

As shown in Table 6, the highest accuracy and the optimal balance are achieved when $\beta = 1$. Furthermore, Figure 4 shows that higher values of β lead to reduced variance of the characteristics, confirming the role of KL regularization in stabilizing the training process—the importance of stability of the characteristics. Overall, $\beta = 1$ represents the optimal trade-off parameter, achieving equilibrium between constraint intensity and feature diversity.

Values of β	ACC	AUC	KL	σ^2	selected
0.1	78.24±2.7	82.15±3.4	0.08	0.29	
0.5	79.44±1.9	82.07±2.8	0.14	0.22	
1.0	85.74±0.5	92.97±2.4	0.18	0.2	✓
1.5	78.04±1.1	81.88±3.3	0.23	0.18	
2	75.02±1.5	79.91±2.0	0.25	0.17	

Table 6: Impact of trade-off parameter β on model performance and feature stability.

Adaptation Analysis of the AILS in Diverse Deep Network Architectures. To validate the architectural adaptability of the proposed AILS, we evaluate its performance across five distinct deep network backbones, spanning diverse depths and design philosophies. As shown in Table 2, AILS consistently enhances baseline performance across all architectures. Notably, MobileNetV3 exhibits the most significant gain (a 6.2% accuracy increase), indicating that AILS effectively compensates for the limited representation ability of lightweight models.

In addition, AILS demonstrates universal convergence acceleration of 17-23% (measured by epochs needed to achieve 95% peak accuracy) across architectures of varying complexity. These results validate the universal validity of the linear semantic properties of deep features across

(a) (b) (c)

Figure 5: Visualization of deep features on DermaMNIST (7-Class) using t-SNE. (a), (b) and (c) respectively demonstrate the effects of 7,000 samples with different data augmentation methods. (a) Cutmix. (b) BSDA. (c) Our AILS.

heterogeneous architectures, which accounts for AILS's robust performance, independent of network design variations. Critically, the algorithm demonstrates architecture-agnostic behavior, providing both theoretical guarantees and practical adaptability for clinical model selection.

A t-SNE Visualization on DermaMNIST. As evidenced in Figure 5, we conduct a comparative t-SNE visualization study on the DermaMNIST dataset (7-class). Feature embeddings are extracted from the penultimate layer after training with three augmentation methods. Compared to CutMix (Yun et al. 2019) and BSDA (Zhu et al. 2024), AILS achieves optimal feature topology characterized by compact intra-class formations, distinct inter-class boundaries, indicating semantic consistency and anatomical plausibility in augmented medical feature space.

Specific Discussion on Novelty and Contributions of AILS

Building upon the foundational insights laid by BSDA (Zhu et al. 2024), which pioneered the application of Bayesian principles through localized magnitude sampling, our proposed AILS constitutes a paradigm shift rather than mere improvement, achieving comprehensive theoretical, mechanistic and functional transcendence over prior arts. Distinct from BSDA with fundamental theoretical limitations, AILS establishes a mathematically rigorous probabilistic learning system. BSDA operates as a augmentation tool that locally applies Bayesian stochasticity for magnitude estimation, primarily serving to expand feature diversity through sampling. In stark contrast, AILS constitutes a framework-level continual learning strategy that systematically embeds Bayesian inference within a geometrically constrained manifold. Our approach transforms semantic directions into direction-aware random variables, implements end to end variational inference through parameterized priors, posterior reparameterization, and an autonomously designed bidirectional constraint. Consequently, AILS constitutes not a collection of isolated sampling steps, but a cohesive, probabilistic learning system designed for persistent adaptation and generalization.

This foundational divergence manifests in three critical dimensions: 1) BSDA's magnitude sampling functions as transient data augmentation, while AILS' directional enhancement acts as persistent feature regularization. Our geometric validity filtering and distributional symmetry enforcement intrinsically preserve semantic topology-mechanisms absent in BSDA. 2) Where BSDA uses Bayesian estimation merely as a magnitude calculator, AILS elevates it to a learning strategy driver through its manifold-constrained variational formulation, with probable Lipschitz continuity in L_{inv}. 3) Manifold-constrained variational optimization integrates uncertainty modeling directly into the core objective of learning robust feature representations.

This systematic unification of Bayesian inference with differential geometry marks a substantial leap in continual learning theory. The direction-aware invariance loss and theoretical completeness modeling constitute mathematical novelties that fundamentally reconfigure how Bayesian principles interact with deep representations, thus intrinsically enhancing model robustness and generalization against perturbations and distribution shifts.

Conclusion

In this paper, we propose an efficient implicit semantic augmentation learning strategy (AILS), providing a novel perspective for improving model feature representation and robust learning in the deep feature space of medical images. We believe that the theoretical framework constructed by AILS also holds significant implications for the interpretability research in deep learning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62376183, 62476190 and U21A20469, the special fund for Science and Technology Innovation Teams of Shanxi Province under Grant 202304051001009.

References

Chao, H.; Zhang, J.; and Yan, P. 2022. Regression metric loss: learning a semantic representation space for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 427–436. Springer.

Chen, H.; Fu, Y.; Jiang, X.; Chen, Y.; Li, W.; Zhou, Y.; and Zheng, F. 2023. Gradient Learning With the Mode-Induced Loss: Consistency Analysis and Applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Chen, W.; Tian, L.; Fan, L.; and Wang, Y. 2019. Augmentation invariant training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2963–2971.

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

El Jiani, L.; El Filali, S.; et al. 2022. Overcome medical image data scarcity by data augmentation techniques: A review. In *2022 International Conference on Microelectronics (ICM)*, 21–24. IEEE.

Goceri, E. 2023. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11): 12561–12605.

Harris, E.; Marcu, A.; Painter, M.; Niranjan, M.; Prügel-Bennett, A.; and Hare, J. 2020. Fmix: Enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, G.; Liu, Z.; Pleiss, G.; Van Der Maaten, L.; and Weinberger, K. Q. 2019. Convolutional networks with dense connectivity. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 8704–8716.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kim, D.; Geenjaar, E.; and Calhoun, V. 2024. Auxiliary objectives improve generalization performance but reduce model specification for low-data neuroimaging-based brain age prediction. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*.

Li, X.; Wu, Y.; Tang, C.; Fu, Y.; and Zhang, L. 2024. Explicitly learning augmentation invariance for image classification by Consistent Augmentation. *Engineering Applications of Artificial Intelligence*, 130: 107541.

Liu, G.; Li, H.; He, Z.; and Zhong, S. 2024. Enhancing Generalization in Medical Visual Question Answering Tasks via Gradient-Guided Model Perturbation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2220–2224. IEEE.

Liu, Z.; Lv, Q.; Li, Y.; Yang, Z.; and Shen, L. 2023. Medaugment: Universal automatic data augmentation plug-in for medical image analysis. *arXiv preprint arXiv:2306.17466*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Qureshi, I.; Yan, J.; Abbas, Q.; Shaheed, K.; Riaz, A. B.; Wahid, A.; Khan, M. W. J.; and Szczuko, P. 2023. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90: 316–352.

Song, Y.; and Chong, N. Y. 2024. S-cyclegan: Semantic segmentation enhanced ct-ultrasound image-to-image translation for robotic ultrasonography. In *2024 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 115–120. IEEE.

Surner, M.; Khelil, A.; and Bothmann, L. 2025. Invariance Pair-Guided Learning: Enhancing Robustness in Neural Networks. *arXiv preprint arXiv:2502.18975*.

Upadhyay, A. K.; and Bhandari, A. K. 2024. Advances in deep learning models for resolving medical image segmentation data scarcity problem: A topical review. *Archives of Computational Methods in Engineering*, 31(3): 1701–1719.

Upchurch, P.; Gardner, J.; Pleiss, G.; Pless, R.; Snavely, N.; Bala, K.; and Weinberger, K. 2017. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7064–7073.

Wang, Y.; Huang, G.; Song, S.; Pan, X.; Xia, Y.; and Wu, C. 2021. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3733–3748.

Xie, C.; Tan, M.; Gong, B.; Wang, J.; Yuille, A. L.; and Le, Q. V. 2020. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 819–828.

Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; and Ni, B. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1): 41.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. 2022. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2736–2746.

Zhang, J.; and Ma, K. 2022. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16650–16659.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13001–13008.

Zhu, Y.; Cai, X.; Wang, X.; Chen, X.; Fu, Z.; and Yao, Y. 2024. BSDA: bayesian random semantic data augmentation for medical image classification. *Sensors*, 24(23): 7511.