Diff-SiamNet: Breast Ultrasound Image Classification with Forward Diffusion and Discriminative Embedding

Chenyi.Zhuang and Di Zhang(✉)

Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
di.zhang@xjtlu.edu.cn

Abstract. Existing breast ultrasound image classification algorithms frequently perform poorly in real-world clinical settings due to image deterioration, semantic drift, and insufficient supervision, significantly limiting tiny lesion recognition and practical deployment. To address this critical issue, this paper proposes Diff-SiamNet, a lightweight discriminative diffusion network. We present a novel forward diffusion perturbation mechanism to emulate scattering noise and boundary blurring, enhancing the model's feature robustness against low-quality images. Additionally, we develop a diffusion-aware attention gate (DAG) to dynamically integrate clear and degraded map features, thereby alleviating semantic discrepancies. The Triplet++ embedding strategy is designed to guide the model to form a discriminative space with inter-class separation and intra-class aggregation under weak annotation. On the BUSI dataset, Diff-SiamNet outperforms state-of-art ResNet, Swin Transformer and HGDF in five metrics, including accuracy (93.4%), AUC (94.2%), and F1-score (92.9%), which significantly improves the ability of recognizing fuzzy boundaries and tiny lesions. The method has good interpretability and deployment efficiency, and is expected to serve clinical intelligent screening in resource-constrained environments.

Keywords: Breast ultrasound, Image classification, Model, Discriminative Embedding.

1 Introduction

Breast cancer is one of the most frequent malignant tumors among women globally, with significant morbidity and mortality rates. It poses a serious danger to women's health and public health-care systems. Early screening and accurate classification are important to achieve individualized treatment and improve survival.

In clinical practice, breast ultrasound (BUS) has emerged as a crucial instrument for breast cancer screening and lesion evaluation, owing to its absence of radiation, cost-effectiveness, and suitability for high-density breast tissue [1]. In recent years, deep learning has achieved substantial advancements in medical image analysis, particularly in breast cancer categorization. CNN is proficient in capturing local textures, while Transformer improves the representation of intricate structures through global attention methods [2]. Furthermore, techniques such as embedding learning [3], multi-scale fusion [4], evolutionary feature selection [5], and interpretability enhancement have been gradually introduced to improve clinical usability and classification accuracy. Although these methods perform well on standard datasets, they still face challenges in real BUS images [6]. Limited by the imaging mechanism, BUS maps often contain scattered noise, low contrast, and fuzzy boundaries, which increase the subjectivity of manual interpretation and automatic classification. Most of the current mainstream models assume good image quality and lack modeling of degraded images, resulting in impaired feature representation and semantic stability [7]. Besides, the semantic differences between clear and degraded images make it difficult to achieve cross-conditional alignment for static fusion, which affects the discriminative effect. Lesions exhibiting intricate morphology, diverse scales, and indistinct boundaries are more prone to misidentification, particularly under suboptimal labeling settings [8]. High computational overhead limits the practicality of some high-performance models for edge device deployment.

In order to crack the semantic misalignment problem between clear and degraded images and improve the model's ability to recognize fuzzy regions, we propose Diff-SiamNet as a discriminative diffusion framework. By incorporating a forward diffusion perturbation mechanism, the model simulates real-world image degradation phenomena, including boundary blurring and speckle noise, thereby improving its robustness in challenging environments. In order to improve feature consistency and minimize semantic drift, we also build a diffusion-aware attention mechanism that dynamically modifies both clear and noisy features in both channel and spatial dimensions. Additionally, the twin embedding structure with auxiliary supervision helps the model learn to discriminate the representations with limited annotations, resulting in a compact semantic space. Main contributions of this paper are as follows:

1. We propose a feature enhancement mechanism based on diffusion modeling to simulate the degradation effect of ultrasound images during the imaging process by gradually injecting Gaussian noise into the original image through a forward diffusion process, which guides the model to learn robust and discriminative feature representations and enhances its adaptability to blurring and interference.
2. We design Diffusion-Aware Attention Gate (DAG), a lightweight fusion module that combines channel attention and spatial attention, to dynamically weight the features of clear and degraded maps, enhance the semantic perception of blurred regions, and improve the consistency and discriminative properties of the fused features.
3. The Triplet++ discriminative embedding mechanism is introduced, combined with online difficult sample mining and dynamic boundary adjustment strategy, to optimize the geometric relationship between samples under low-supervision conditions, and construct a low-dimensional embedding space with interclass separation and intraclass tightness, which significantly improves the model's classification robustness and generalization ability in complex scenarios.

2 Related work

2.1 Breast Ultrasound Image Classification

In recent years, breast ultrasound image (BUS) classification research has shifted from traditional manual feature methods to deep learning dominated paths. models such as ResNet [9] and Transformer [10] have achieved good Performance on BUSI dataset, and fusion strategy further improves the stability and generalization ability of the models, which constitutes the current mainstream SOTA route.

However, these approaches generally rely on high-quality images and lack the ability to model real degradation factors such as noise, blur, and low contrast, which leads to significant degradation of classification performance in complex scenarios [11]. Researchers pointed out that the existing models lack a quality-aware mechanism to deal with semantic shifts and feature confusions caused by degraded images. In addition, although the attention mechanism has been widely used in medical image modeling, it is mostly static, difficult to adapt to changes in image clarity, and prone to bias in the fusion process [12]. Although embedding methods such as Triplet Loss have improved the discriminative ability of fuzzy regions, they are still limited by fixed boundary settings and sample instability. Currently, the mainstream SOTA methods are still insufficient in degradation modeling, semantic alignment and fuzzy discrimination, which are difficult to cope with complex clinical scenarios, and there is an urgent need to construct a robust classification framework with degradation sensing, dynamic fusion, and weakly supervised discrimination to improve its practicability and robustness.

2.2 Diffusion Mechanisms in Medical Imaging

Diffusion models have attracted a lot of attention in the field of medical images in recent years due to their excellent noise modeling and image restoration capabilities. Med-DDPM [13] achieve high-quality image generation under low-resource conditions, while MC-DDPM [14] completes the multimodal structural complementation by using conditional guidance, which demonstrates its natural advantages in modeling image degradation. In the cross-modal generation and segmentation task, PathLDM [15] adeptly represents intricate anatomical features with the incorporation of spatial priors and conditional control. The forward diffusion process effectively simulates scattering and contrast loss in breast ultrasound pictures, enhancing the model's robustness in uncertain situations [16].

Despite the outstanding performance of the diffusion mechanism in generative tasks, existing classification methods still generally lack a degradation modeling module and have limited ability to adapt to noisy and blurred images. Attempts have been made to introduce diffusion into the classification process, but mostly at the auxiliary or preprocessing level without effective integration. It indicates that the current SOTA classification framework is insufficiently adaptive in image degradation scenarios, and the introduction of a forward diffusion mechanism can effectively enhance the model's perception of degraded semantics and improve its robustness and generalization ability.

2.3 Semantic Alignment and Dynamic Attention Modeling for Degraded Images

Degradation in breast ultrasound images not only weakens the low-level feature representation, but also triggers semantic misalignment between clear and fuzzy images, which seriously interferes with the stability of static fusion strategies. In order to improve the discriminative ability of critical regions, attention mechanisms have been widely introduced into medical image classification tasks [17].

In recent years, methods such as CBAM, nonlocal attention, and coordinate attention have advanced lesion localization and semantic enhancement [18]. However, these methods are generally based on the assumption of semantic consistency of the input images, which makes it difficult to deal with the semantic drift problem due to fluctuations in image quality. CoTNet [19]and Shuffle Attention [20] in recent years still exhibit insufficient targeted modeling capability to dynamically adjust to inputs of variable quality when addressing local semantic alterations induced by image degradation. Therefore, dynamic modeling approaches with quality-aware capabilities are needed because image deterioration causes semantic drift, making it difficult for conventional attention mechanisms to adjust. This study introduces the Diffusion Aware Attention Gate (DAG) module, which uses forward diffusion perturbation and attention fusion to achieve dynamic semantic alignment between clear and degraded images and increase feature fusion consistency and resilience.

2.4 Robust Metric Learning under Fuzzy Supervision

Breast ultrasound images often have overlapping features due to fuzzy boundaries, low contrast and weak labeling, which can easily lead to misclassification of fuzzy samples. For medical image classification, embedding learning techniques including Siamese network, Triplet Loss, and contrast learning have been presented to improve the discriminative qualities. These techniques are particularly well-suited for low-quality picture scenes and limited samples [21].

Conventional techniques such as ArcFace and ProxyAnchor have achieved notable advancements in inter-class separation[22]. However, the traditional Triplet Loss is constrained by fixed boundaries, presents challenges in negative sample selection, exhibits training instability, and struggles to accommodate the high variability of ambiguous samples in BUS environments [23]. Some studies have tried to alleviate the training difficulties through dynamic boundaries and online mining mechanisms, but most of them ignore the uncertainty characteristics of fuzzy samples and lack the structural design for fuzzy classification.

Therefore, we propose a fuzzy-aware discriminative embedding mechanism based on the Siamese framework, introduce the Triplet++ optimization strategy, integrate the quality distribution of fuzzy samples with the difficult case mining mechanism, and construct robust inter-class geometric boundaries in the low-dimensional space, so as to improve the discriminative robustness of the model in the conditions of fuzzy foci and weak annotation.

3 Methodology

The overall structure of the model is shown in Fig. 1, which consists of three main stages: forward diffusion perturbation generation, diffusion-aware attention fusion mechanism (DAG), and discriminative embedding learning path.

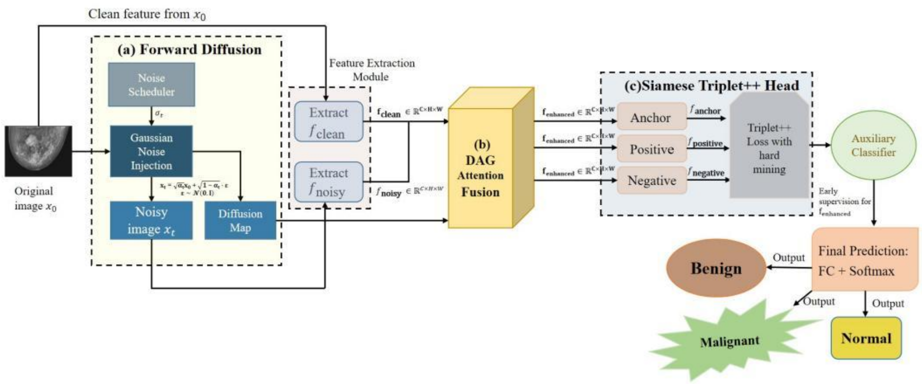

Fig. 1. Overview of Diff-SiamNet architecture. (a) Forward Diffusion simulates degradation,(b) DAG Fusion aligns clean/noisy features,(c) Triplet++ Head enhances discriminative embedding and guides classification.

At the input stage, the original ultrasound image is noted as x_0 , and subsequent to a diffusion perturbation process based on a predefined noise scheduling strategy, a degraded image is generated at x_t . The perturbation method replicates prevalent imaging degradation processes in clinical environments, including tissue scattering, fuzzy edges, and localized obstructions. This stage provides both semantic preservation and perturbation improvement by employing standard Gaussian noise and the time-step control variable t . In contrast to conventional diffusion models, the current framework omits inverse reconstruction paths, hence preventing the accumulation of reconstruction errors, minimizing unnecessary computations, and guaranteeing the model prioritizes discriminative information over pixel reduction.

The scrambled image x_t and the original image x_0 are simultaneously fed into a shared encoder $\mathcal{E}(\cdot)$ to extract their semantic features respectively f_{clean} with f_{noisy} . Owing to their substantial disparities in spatial configuration and textural intricacies, straight splicing or average fusion may result in semantic discrepancies. Considering the significant spatial and textural differences between clear and degraded features, direct splicing or average fusion is prone to trigger semantic bias. For this reason, we introduce a diffusion-aware attention gate (DAG) module for guided alignment and dynamic fusion of dual-stream features. The final fused feature representationf_{fused} has both stability and discriminative properties, and is used as the input for subsequent classification.

In the discriminative learning phase, we adopt the Siamese embedding network structure based on ternary inputs. For each set of input samples, we construct the ternary (x^a, x^p, x^n), which represents the Anchor, Positive and Negative samples respectively, and map them into the embedding vector (z^a, z^p, z^n) . This embedding space is learned from the shared network $\phi(\cdot)$ with the goal of minimizing the distance between similar samples and enlarging the interval between dissimilar samples. In order to optimize the separability of feature boundaries, we implement dynamic margins and an online hard sample mining mechanism to concentrate training on challenging samples in fuzzy regions. Meanwhile, an auxiliary classifier is added to oversee the fused features and improve category-level feedback via cross-entropy loss, thereby synergistically optimizing the embedding structure and classification performance.

3.1 Forward diffusion perturbation preprocessing

As shown in Fig. 2, a forward diffusion-based perturbation modeling strategy generates semantically consistent clear-degradation map pairs with the DDPM successive degradation mechanism and guides the model to learn degradation-invariant feature representations to improve BUS image noise and blur discrimination.

In diffusion modeling, the central goal of the forward process is to transform the original imagex_0 into the degraded imagex_t at step$t$ by gradually injecting Gaussian noise. This process can be formalized as the following expression:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

Where $\bar{\alpha}_t$ is the cumulative product of noise scheduling coefficients that controls the proportion of original information retained;ϵ is the noise sample in the standard Gaussian distribution; and t indicates the diffusion timestep, reflecting the degradation severity. The perturbation mechanism has controllability and continuity, and can simulate a wide range of image conditions from slight blurring to severe degradation, while maintaining structural integrity, avoiding semantic drift, and achieving more realistic and high-quality degradation modeling.

After constructing the degraded image x_t , we design a set of shared-parameter dual-path feature extraction networks$f_\theta(\cdot)$ that act on the original image x_0 and its degraded version x_t to obtain their respective deep representations:

$$f_{\text{clean}} = f_\theta(x_0), \quad f_{\text{noisy}} = f_\theta(x_t) \tag{2}$$

The design ensures that the features extracted from the two input images under the same network structure are in the same semantic space, which helps to model the modal differences and semantic consistency between the images and provides structural input for the subsequent attention fusion module.

From a task-driven perspective, this image pair construction approach offers three advantages. First, it can simulate the image degradation caused by differences in devices, operators or patients in real scenarios, and enhance the realism of the training sample distribution. Second, it enhances the model's adaptability to noise and blur by mixing clear and degraded samples, and improves the robustness of discrimination under blurred regions. Third, constructing pairs of semantically consistent but significantly differently observed samples provides effective anchor-positive pair support for Triplet++ embedding learning. Moreover, diffusion scheduling employs random time-step sampling to subject the model to varying levels of image deterioration throughout each training iteration, hence enhancing its generalization capacity. The feature extraction network may adeptly select lightweight or multi-scale architectures to optimize deployment efficiency and lesion detection capability.

3.2 Diffusion-Aware Attention Gate (DAG)

To address the issue of misalignment in semantic representation between clean and diffusion-perturbed maps, we develop a Diffusion-Aware Attention Gate (DAG) module to dynamically adjust the weighting of the joint features from both sources as illustrated in Fig. 2.

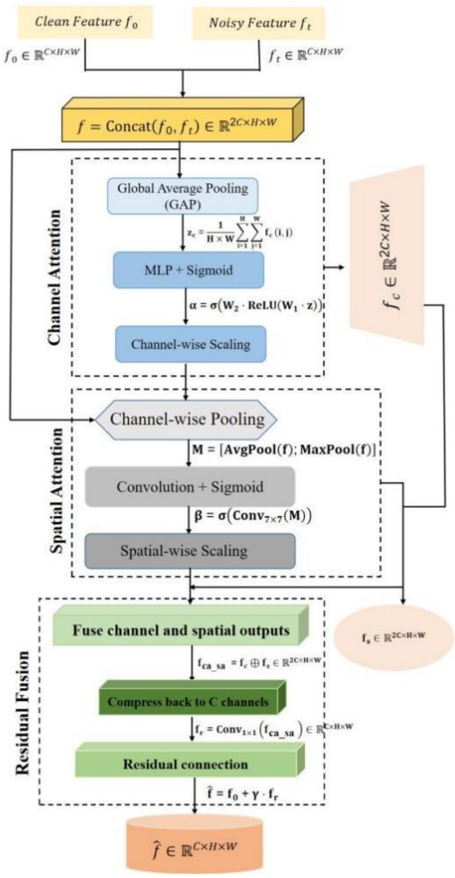

Fig. 2. Architecture of the Diffusion-Aware Attention Gate (DAG) Module.

The module estimates the importance distribution of features from both global and local scales based on both channel and spatial dimensions, and generates robust and structure-sensitive discriminative representations by combining residual connectivity, which effectively improves the model's adaptability to ambiguous regions and perturbation noise.

Specifically, let the input be a clear image feature $f_0 \in \mathbb{R}^{C\times H\times W}$ and its corresponding diffuse perturbation feature $f_t \in \mathbb{R}^{C\times H\times W}$, we first splice the two along the channel dimension to obtain the joint feature tensor:

$$f = \mathrm{Concat}(f_0, f_t) \in \mathbb{R}^{2C\times H\times W} \tag{3}$$

Where C is the original channel number and H, W are the spatial dimensions. This splicing operation preserves the semantic structure of the clean path while introducing the interference response information under diffuse perturbation, which provides the basis of dual-path fusion for the attention module.

To estimate the importance of each semantic channel from the global dimension, the DAG first performs a global average pooling operation on f to obtain a vector of channel-level statistics:

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} f\,(k,i,j), \quad z_c \in \mathbb{R}^{2C} \tag{4}$$

Where $f(k,i,j)$ denotes the response value of the kth channel at location (i,j), and $z_c(k)$ is the global average activation value of the channel, reflecting its semantic response strength.

Subsequently, we input this vector into a perceptron network consisting of two fully connected layers to model the nonlinear relationship between the channels and obtain the channel attention weights:

$$\alpha = \sigma\big(W_2 \cdot \mathrm{ReLU}(W_1 \cdot z_c)\big), \quad \alpha \in \mathbb{R}^{2C} \tag{5}$$

Where $W_1 \in \mathbb{R}^{\frac{2C}{r}\times 2C}, W_2 \in \mathbb{R}^{2C\times\frac{2C}{r}}$ is the learnable parameter, r is the channel compression rate, and σ denotes the Sigmoid activation function. This attention mechanism directs the model to focus on the channel dimensions that are more sensitive to the task and suppresses redundant responses that are not discriminative in the perturbation path.

The attentional weights α are applied to the original spliced features f by broadcast multiplication to obtain a channel-weighted feature representation $f_c \in \mathbb{R}^{2C\times H\times W}$ where each channel is scaled according to its importance.

In the spatial attention path, we further compute the discriminative weights of each position. We first perform channel average pooling and maximum pooling operations on f, respectively, and construct fusion features after channel dimensional splicing $M \in \mathbb{R}^{2\times H\times W}$. This preserves the background trends and strong response signals in localized regions. The spatial attention map is then extracted by a convolutional kernel of7×7 with Sigmoid activation:

$$\beta = \sigma\left(\mathrm{Conv}_{7\times7}(M)\right), \quad \beta \in \mathbb{R}^{1\times H\times W} \tag{6}$$

Where $\beta(i, j)$ denotes the significance score of the spatial location (i, j), with higher values indicating that the region is more discriminative. We expand and act on the original features f pixel by pixel to generate a spatially weighted representation $f_s \in \mathbb{R}^{2C\times H\times W}$ that enables the model to actively focus on anomalous boundaries, fuzzy structures, or disturbed regions.

Next, the channel attention result f_c and the spatial attention result f_s were summed element-by-element and fused into enhanced features:

$$f_{ca\_sa} = f_c + f_s \in \mathbb{R}^{2C\times H\times W} \tag{7}$$

This operation combines the weighted understanding of features in both pathways, capturing the information response in the dual mode of "channel importance × spatial focus".

To map the fused features back to the original channel dimension, we use a 1×1 convolution operation:

$$f_r = \mathrm{Conv}_{1\times1}\left(f_{ca\_sa}\right) \in \mathbb{R}^{C\times H\times W} \tag{8}$$

This transformation not only compresses the parameter dimensions, but also preserves the weighted feature semantics constructed by the previous attention mechanism.

Finally, to stabilize the training and enhance the consistency of the representation, we introduce residual connectivity:

$$\hat{f} = f_0 + \gamma \cdot f_r \tag{9}$$

Where γ is a learnable scaling factor to control the fusion ratio of the discriminative information carried by the perturbation paths. If $\gamma \approx 0$, the model relies more on clean representations; if $\gamma > 1$, it pays more attention to perturbation-sensitive features. The output result $\hat{f} \in \mathbb{R}^{C\times H\times W}$ has both structural stability and semantic perturbation sensing ability.

3.3 Siamese Triplet++ Discriminative Embedding Module

To improve the model's discriminative ability in fuzzy focus and small sample scenarios, Diff-SiamNet introduces a discriminative embedding module based on the Siamese architecture. This module, when combined with the optimized Triplet++ loss, creates a clear inter-class separation and intra-class aggregation in the low-dimensional semantic space, significantly improving classification robustness to fuzzy boundaries and uncertain regions.

Specifically, the input features $\hat{f} \in \mathbb{R}^{C\times H\times W}$ are firstly downscaled to a one-dimensional vector $v \in \mathbb{R}^d$ by Global Average Pooling (GAP) or flatten operation, and then enter the Embedding Head to generate the final discriminative embedding $e \in \mathbb{R}^D$. During the training process, the anchor samplee_a is sampled from each minibatch, and the corresponding positive samples e_p (similar but different images) and negative samples e_n (dissimilarities) together form the triad (e_a, e_p, e_n), which is optimized by the following Triplet++ loss:

$$\mathcal{L}_{triplet++} = \max\left(0, \|e_a - e_p\|_2^2 - \|e_a - e_n\|_2^2 + m_i\right) \tag{10}$$

Where $\|\cdot\|_2^2$ denotes the Euclidean distance squared, and m_i is the margin value for the i th triplet that is dynamically adjusted adaptively. Unlike the traditional Triplet loss which uses a fixed margin, a dynamic margin mechanism is introduced here:

$$m_i = m_0 \cdot \left(1 + \alpha \cdot \|e_a - e_p\|_2\right) \tag{11}$$

Where m_0 is the initial boundary threshold and α is the deflation factor. This strategy can be flexibly adjusted according to the actual distance between the anchor and its positive samples, effectively alleviating the impact of hard negative samples on the training stability.

To concentrate on representative challenging samples, we backpropagate solely the triples with a loss exceeding zero, therefore improving the embedding structure and augmenting the model's sensitivity to subtle distinctions. The module accomplishes intra-class aggregation and inter-class separation under poor supervision, and the inference phase executes discriminative prediction based on class centers using Euclidean distance.

For any test sample, its corresponding embedding e_t will be compared with each class center vector μ_k, and the class corresponding to the smallest distance will be selected as the prediction result:

$$\hat{y}_t = \mathrm{argmin}_k \|e_t - \mu_k\|_2 \tag{12}$$

Overall, the Siamese Triplet++ embedding module introduces a generalizable geometric discriminative space for Diff-SiamNet with high robustness, low supervised dependency, and enhanced boundary-awareness, which provides core support for complex scene classification in breast ultrasound images.

4 Experimental design and analysis of results

4.1 Experimental setting

We evaluated the proposed Diff-SiamNet model on the BUSI dataset [24], which contains 700 labeled ultrasound images covering normal, benign and malignant lesions. The database was divided into training, validation, and testing sets in an 8:1:1 ratio to ensure the scientific validity and reproducibility of the experiments. Evaluation metrics include Accuracy (Acc), Sensitivity (Sens), Specificity (Spec), F1-score, Area Under the Curve (AUC) to comprehensively measure classification performance. The model is trained on RTX 3090 GPUs based on PyTorch with Adam optimizer (learning rate $1e^{-4}$, batch size 16, maximum number of rounds 100), and an early-stopping mechanism is introduced to avoid overfitting with the validation set

AUC being accurate. All images are uniformly scaled to 224×224 and normalized to [0, 1] to improve training stability and generalization ability.

4.2 Comparative analysis with mainstream methods

We compare the performance of Diff-SiamNet with several mainstream models on BUSI: ResNet-50, BTS-ST, Twin-CNN, HGDF, Voting Ensemble, PSOCNN, Diff-SiamNet (Ours)

Table 1. Quantitative comparison of Diff-SiamNet against state-of-the-arts on the BUSI dataset.

Model	Acc	AUC	F1	Sens	Spec
ResNet-50[25]	0.837	0.872	0.864	0.849	0.871
BTS-ST[26]	0.862	0.901	0.875	0.853	0.875
Twin-CNN[27]	0.879	0.916	0.891	0.878	0.894
HGDF[28]	0.902	0.929	0.909	0.889	0.912
Voting Ensemble[29]	0.888	0.921	0.913	0.906	0.921
PSOCNN[30]	0.901	0.925	0.904	0.897	0.919
Diff-SiamNet (Ours)	**0.934**	**0.942**	**0.929**	**0.913**	**0.937**

Compared with existing methods, Diff-SiamNet achieves better overall performance on the BUSI dataset (AUC 0.942, F1 0.929), reflecting stronger discriminative ability for blurred lesions and low-quality images. Traditional methods such as ResNet-50 and BTS-ST rely on static feature extraction, which significantly degrades performance under image degradation conditions. We introduce a diffusion perturbation mechanism to simulate the real imaging degradation process, which enhances the feature robustness of the model and fits the advantages of diffusion modeling in anti-jamming tasks[31].

Meanwhile, the DAG module in Diff-SiamNet fuses channel and spatial attention to achieve dynamic semantic modeling, which performs better in dealing with fuzzy boundaries compared to the single-attention mechanism used in HGDF and PSOCNN. In addition, the Triplet++ embedding strategy combines dynamic boundary adjustment and difficult sample mining, overcoming the problems of difficult sample selection and unstable optimization of traditional Triplet Loss, and further improving the ability of inter-class differentiation[32]. Therefore, instead of relying on depth stacking or parameter expansion, Diff-SiamNet systematically solves the key performance bottlenecks of current methods in scenarios such as image degradation, semantic ambiguity, and weak supervision at the level of structural design, and shows stronger clinical adaptability and application potential.

4.3 Ablation experiments

To evaluate the performance contribution of each core module of Diff-SiamNet, we conducted systematic ablation experiments on the BUSI dataset. The impact on the overall discriminative performance is analyzed by gradually removing or replacing the forward diffusion perturbation, DAG attention gate and Triplet++ loss modules. The evaluation was performed using both accuracy and AUC metrics, and the results are shown below in percentage form:

Table 2. Ablation study results of Diff-SiamNet on the BUSI dataset

Model Configuration	Acc	AUC
Remove DAG Attention Gate	0.897	0.919
Replace Triplet++ loss with cross-entropy	0.875	0.893
Removal of forward diffusion perturbation mechanisms	0.862	0.874
Diff-SiamNet (Ours)	**0.934**	**0.942**

The ablation experiments indicate that the performance enhancement of Diff-SiamNet arises from the synergistic design of various core modules rather than the aggregation of a singular structure. Removing the DAG attention gate reduces the accuracy to 0.897 and the AUC to 0.919, indicating that it plays a key role in the dynamic fusion of clear and degraded image features, especially in the discrimination of fuzzy regions, which is better than the traditional static attention mechanism. Substituting the Triplet++ loss with cross-entropy results in accuracy and AUC decreasing to 0.875 and 0.893, respectively, suggesting that discriminative embedding is superior for creating a compact and semantically clear feature space, which is more appropriate for fine-grained and low-labeling tasks compared to pure classification supervision. After removing the diffusion perturbation mechanism, the accuracy decreases to 0.862 and the AUC decreases to 0.874, which is a significant performance degradation, verifying the unique value of image degradation modeling in enhancing feature robustness. Through multi-module synergy, Diff-SiamNet beats previous approaches in accuracy, robustness, and flexibility, proving its innovation and applicability in complicated medical image processing.

5 Conclusion

This research proposes Diff-SiamNet, a lightweight discriminative diffusion framework for classifying breast ultrasound images. By modeling the image degradation process and introducing the Triplet++ embedding mechanism, the model shows strong robustness and discriminative ability in difficult cases such as fuzzy boundaries and small lesions, and has the advantages of compact structure, efficient training, and friendly deployment[33].

Future endeavors will concentrate on three facets: firstly, augmenting the model to encompass multimodal medical imaging for enhanced recognition of intricate lesions; secondly, implementing self-supervised or cross-domain pre-training to diminish reliance on high-quality annotations; and thirdly, deploying the model in a multi-center clinical setting to assess its robustness and adaptability to varying equipment and patient conditions.

Acknowledgments. All authors acknowledge the financial support provided by XJTLU Academic Enhancement Fund, Center of Excellence for Syntegrative Education (CoESE) Fund (ref: CoESE242502) and China-BRICS Artificial Intelligence Development and Cooperation Center's 2025 BRICS Industry Ecosystem Think Tank Network and Open Project.

6 Reference

1. G. Ayana and S. Choe, "Vision transformers-based transfer learning for breast mass classification from multiple diagnostic modalities", *J. Electr. Eng. Technol.*, vol. 19, no. 5, Art. no. 5, Jul. 2024, doi: 10.1007/s42835-024-01904-w.
2. A. Batool and Y.-C. Byun, "Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm".
3. T. Islam, "Predictive modeling for breast cancer classification in the context of bangladeshi patients by use of machine learning approach with explainable AI".
4. T. A. Rashid *et al.*, "NSGA-II-DL: metaheuristic optimal feature selection with deep learning framework for HER2 classification in breast cancer".
5. K. Atrey, "Multimodal classification of breast cancer using feature level fusion of mammogram and ultrasound images in machine learning paradigm", *Multimed. Tools Appl.*, 2024.
6. A. Yaqoob, "Improving breast cancer classification with mRMR + SS0 + WSVM: a hybrid approach", *Multimed. Tools Appl.*.
7. E. A. Rakha, G. M. Tse, and C. M. Quinn, "An update on the pathological classification of breast cancer", 2022.
8. Z. Zhu, Y. Sun, and B. H. S. Asli, "Early breast cancer detection using artificial intelligence techniques based on advanced image processing tools".
9. A. C. Yadav, M. H. Kolekar, and M. K. Zope, "ResNet-101 empowered deep learning for breast cancer ultrasound image classification", 2024.
10. W. Lee, H. Lee, H. Lee, E. K. Park, H. Nam, and T. Kooi, "Transformer-based deep neural network for breast cancer classification on digital breast tomosynthesis images".
11. C. Mao, "Discrimination between human epidermal growth factor receptor 2 (HER2)-low-expressing and HER2-overexpressing breast cancers: a comparative study of four MRI diffusion models".
12. S. E. Rhabori, "Exploring innovative strategies for identifying anti-breast cancer compounds by integrating 2D/3D-QSAR, molecular docking analyses, ADMET predictions, molecular dynamics simulations, and MM-PBSA approaches".

13. Y. Shokrollahi, S. Yarmohammadtoosky, M. M. Nikahd, P. Dong, and L. Gu, “A comprehensive review of generative AI in healthcare”.
14. M. A. Bahloul, S. Jabeen, S. Benoumhani, H. A. Alsaleh, Z. Belkhatir, and A. Al-Wabil, “Advancements in synthetic CT generation from MRI: a review of techniques, and trends in radiation therapy planning”.
15. S. Yellapragada, A. Graikos, P. Prasanna, T. Kurc, J. Saltz, and D. Samaras, “PathLDM: text conditioned latent diffusion model for histopathology”.
16. G. Müller-Franzes, “Diffusion probabilistic”, 2024.
17. N. Meng, “A comparative study of the value of amide proton transfer-weighted imaging and diffusion kurtosis imaging in the diagnosis and evaluation of breast cancer”.
18. H. Ni *et al.*, “High-content stimulated raman histology of human breast cancer”, vol. 14, no. 4, 2024.
19. R. E. Lane, D. Korbie, K. K. Khanna, A. Mohamed, M. M. Hill, and M. Trau, “Defining the relationship between cellular and extracellular vesicle (EV) content in breast cancer via an integrative multi-omic analysis”.
20. H. Liu *et al.*, “Assessment of the reliability and quality of breast cancer related videos on TikTok and bilibili: cross-sectional study in China”.
21. S. Hussain, “A discriminative level set method with deep supervision for breast tumor segmentation”, *Comput. Biol. Med.*, 2022.
22. Z. Ren, X. Liu, J. Xu, Y. Zhang, and M. Fang, “LittleFaceNet: a small-sized face recognition method based on RetinaFace and AdaFace”.
23. Y. Liu *et al.*, “Enhanced dual contrast representation learning with cell separation and merging for breast cancer diagnosis”, 2024.
24. W. Al-Dhabyani, “Dataset of breast ultrasound images”.
25. T. Das, D. S. K. Nayak, and A. Kar, “ResNet-50: the deep networks for automated breast cancer classification using MR images”.
26. A. Iqbal and M. Sharif, “BTS-ST: swin transformer network for segmentation and classification of multimodality breast cancer images”, 2023.
27. O. N. Oyelade, “A twin convolutional neural network with hybrid binary optimizer for multimodal breast cancer digital image classification”.
28. D. Addo, “A hybrid lightweight breast cancer classification framework using the histopathological images”, *Biocybern. Biomed. Eng.*, 2024.
29. A. Batool and Y.-C. Byun, “Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm”.
30. K. Atrey, “Multimodal classification of breast cancer using feature level fusion of mammogram and ultrasound images in machine learning paradigm”, *Multimed. Tools Appl.*, 2024.
31. G. Müller-Franzes, “Diffusion probabilistic”, 2024.
32. B. Abhisheka, “HBNet: an integrated approach for resolving class imbalance and global local feature fusion for accurate breast cancer classification”.
33. F. Taheri, “Improving breast cancer classification in fine-grain ultrasound images through feature discrimination and a transfer learning approach”, *Biomed. Signal Process. Control*, 2025.