



Research article

BGDNet: boundary-guided dual-backbone encoder–decoder network for dermoscopic lesion segmentation

Jiayu Zhang, Haiyang Liu, Ying Ye and Xiaoming Huang*

School of Computer Science, Beijing Information Science and Technology University, Beijing 100192, China

* **Correspondence:** Email: huangxm18@bistu.edu.cn; Tel: +8613426005342.

Abstract: Accurate segmentation of dermoscopic lesions is essential for computer-aided melanoma assessment, but reliable delineation remains challenging because lesions often exhibit irregular shapes, low contrast, and ambiguous boundaries. This study aimed to improve both lesion-region segmentation and contour localization by better coordinating complementary local and global image representations. We developed BGDNet, a boundary-guided dual-backbone encoder–decoder network that combines TransXNet and Swin Transformer. The model integrates local structural information with long-range contextual information and uses boundary priors to guide feature interaction. Its performance was evaluated on the ISIC-2016, ISIC-2018, and PH2 datasets and compared with representative convolutional, Transformer-based, and hybrid segmentation methods using both region-based and boundary-sensitive metrics. On ISIC-2018, BGDNet achieved an intersection over union (IoU) of 0.8813 ± 0.0014 , a Dice similarity coefficient (DSC) of 0.9330 ± 0.0009 , a 95th percentile Hausdorff distance (HD95) of 5.84 ± 0.11 pixels, an average surface distance (ASD) of 1.34 ± 0.03 pixels, and a boundary F-score at a 2-pixel tolerance (BF@2) of 0.868 ± 0.003 . A recall (sensitivity) of 0.9602 ± 0.0021 and a specificity of 0.9847 ± 0.0007 further characterize foreground–background behavior under class imbalance. Additional in-dataset and cross-dataset experiments, together with component-wise ablation and stage-wise boundary-prior visualization, provide complementary evidence for the model design. These findings show that boundary-guided interaction between complementary encoders can improve both lesion overlap and contour accuracy, particularly for lesions with irregular or weakly defined margins. BGDNet therefore provides a promising framework for accurate dermoscopic lesion segmentation.

Keywords: dermoscopic image segmentation; skin lesion segmentation; boundary-aware segmentation; dual-backbone network; transformer-based models

1. Introduction

The incidence of melanoma continues to increase globally, driven by environmental factors, greater ultraviolet exposure, and population aging, with a rising incidence among younger cohorts [1, 2]. Melanoma is highly aggressive, and early detection substantially improves survival outcomes [3]. Early-stage lesions often present with subtle or atypical signs, which can lead to misdiagnosis or diagnostic delay. Dermoscopy, a noninvasive imaging modality, plays a central role in early screening and clinical assessment of cutaneous lesions [4, 5]. Nonetheless, dermoscopic interpretation relies on clinician expertise and is subject to inter- and intra-observer variability; manual delineation of lesion boundaries is time-consuming and inconsistent [6]. These limitations motivate automated segmentation systems that are accurate, efficient, and clinically reliable.

Deep learning has driven substantial progress in medical image segmentation. Convolutional neural networks (CNNs), notably U-Net and its variants, effectively combine low-level spatial detail with high-level semantic context through encoder–decoder structures and skip connections [7, 8]. However, convolutional operators are intrinsically local, which constrains the effective receptive field and limits the modeling of long-range dependencies. This limitation is consequential for dermoscopic images, where lesions commonly exhibit irregular shapes, weak contrast at margins, and heterogeneous textures. Approaches such as dilated convolutions and multiscale fusion partially alleviate this issue [9, 10], yet they do not provide a principled mechanism for capturing global context.

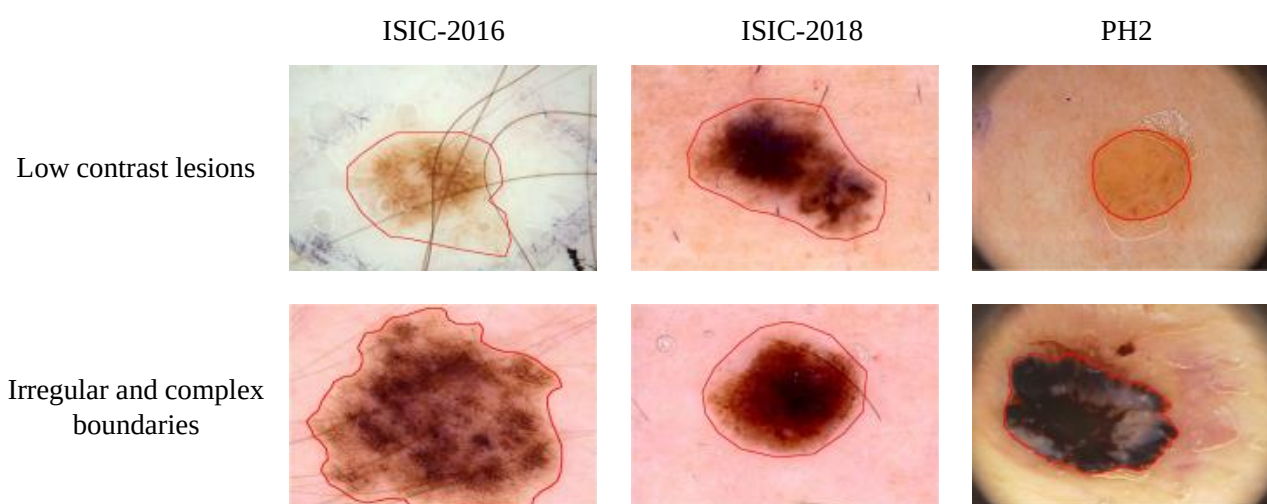


Figure 1. Representative dermoscopic images from ISIC-2016, ISIC-2018, and PH2 with ground-truth lesion boundaries overlaid in red, highlighting low contrast and irregular lesion margins that challenge accurate segmentation.

Vision Transformers, including the Vision Transformer [11], model long-range interactions through self-attention [12]. Hierarchical designs such as the Swin Transformer [13] further balance local and global representations through shifted-window attention. Despite these strengths, Transformer-based

models may insufficiently preserve fine structural details, often yielding over-smoothed boundaries that are critical for clinical interpretation. Hybrid CNN–Transformer architectures attempt to combine convolutional spatial precision with attention-driven context [14,15], yet explicit boundary exploitation and multiscale alignment across heterogeneous encoders remain challenging. Although boundary-aware loss functions can complement region-based objectives, their effectiveness depends on reliable boundary-cue extraction from intermediate features [16].

As illustrated in Figure 1, dermoscopic lesions from widely used public benchmarks, including ISIC-2016, ISIC-2018, and PH2, frequently exhibit weak contrast against surrounding skin and highly irregular or ambiguous boundaries. These characteristics substantially increase the difficulty of accurate lesion delineation, particularly along lesion margins, and highlight the need for segmentation frameworks that explicitly account for boundary information rather than relying solely on region-based feature learning.

Despite recent progress, aligning heterogeneous representations while preserving reliable boundary information remains challenging. Direct concatenation or unconstrained attention between local-structure and Transformer branches may introduce semantic and statistical mismatch, particularly near low-contrast lesion margins. Furthermore, boundary information is frequently used only as an auxiliary decoder output or a training objective rather than as a signal that directly regulates cross-branch feature interaction.

To address these limitations, we introduce BGDNet, a boundary-guided dual-backbone network for dermoscopic lesion segmentation. The model combines TransXNet [17] and Swin Transformer to capture complementary local-detail and global-context representations. Rather than using Sobel operators, attention, and deformable convolution as isolated components, BGDNet organizes them within a boundary-prior-conditioned fusion framework. Feature-level boundary priors derived from fixed Sobel responses and a learnable projection directly regulate bidirectional interaction between the two heterogeneous branches, making feature fusion more reliable around ambiguous lesion contours.

The principal contributions of this work are as follows:

- (1) We construct a heterogeneous dual-backbone encoder using a four-stage Swin Transformer branch and a TransXNet-XXS branch to provide complementary hierarchical representations. Because the two branches produce features with different channel characteristics and statistical distributions, stage-specific 1×1 projections are applied before boundary extraction and cross-branch interaction, providing compatible feature spaces for subsequent boundary-guided fusion.
- (2) We introduce a boundary extraction module (BEM) that applies fixed channel-wise Sobel operators followed by a learnable projection to derive feature-level boundary priors. These priors provide additional boundary guidance from intermediate representations and are not interpreted as pixel-accurate clinical lesion boundaries.
- (3) We introduce boundary-guided cross-attention (BGCA) to regulate bidirectional information exchange between the two encoder branches. BEM-derived boundary priors are introduced into the query and key pathways, whereas the value pathways retain the corresponding semantic features. The boundary priors therefore influence cross-branch correspondence without replacing the semantic content being transferred.

2. Related work

2.1. CNN-based segmentation methods

Since the introduction of U-Net [7], CNNs have remained a core paradigm for medical image segmentation. The encoder–decoder design with skip connections mitigates spatial information loss induced by downsampling, improving localization and optimization stability. Building on this foundation, V-Net [18], U-Net++ [19], and attention U-Net [8] strengthen multiscale representation learning through volumetric convolutions, densely nested pathways, and attention gates that suppress irrelevant activations. Subsequent designs such as residual U-Net++ (ResUNet++) [20] and parallel reverse attention network (PraNet) [21] further incorporate residual learning, multiscale context modules, and reverse-attention or edge-aware cues, yielding improved contour delineation in dermoscopic and gastrointestinal segmentation tasks.

Beyond architectural variants, CNN-based methods commonly adopt deep supervision, feature recalibration, and auxiliary edge streams to stabilize training and recover fine structures. Nevertheless, the locality of convolution restricts the effective receptive field and limits the modeling of long-range dependencies. This constraint is pronounced for dermoscopic lesions with irregular morphology, low contrast, and weak boundaries.

Although dilated convolutions and pyramid-style multiscale aggregation expand contextual coverage, they may introduce gridding artifacts and require careful tuning of dilation rates. Moreover, these mechanisms do not provide an explicit formulation for image-wide relational reasoning. CNNs can also be sensitive to domain shifts across devices and clinical centers. In the absence of global context, learned representations may overemphasize texture while under-representing structural cues, which degrades generalization to small or ambiguous lesions. From an efficiency perspective, maintaining high-resolution decoding to preserve boundary detail increases memory cost and latency, posing challenges for deployment under practical clinical constraints.

2.2. Transformer-based segmentation methods

The Vision Transformer [11] introduces self-attention for capturing global dependencies, enabling direct modeling of long-range interactions among visual tokens. Transformer U-Net (TransUNet) [14] extends this paradigm by embedding a Vision Transformer into an encoder–decoder framework, combining token-level global context with convolutional low-level features. The Swin Transformer [13] and Swin-Unet [22] improve scalability through hierarchical representations with shifted-window attention, achieving favorable efficiency while retaining strong representational capacity. Medical Transformer (MedT) [23] further explores gated axial attention to refine feature interactions and reduce redundant correlations.

Despite their advantages in contextual modeling, Transformer-based segmentation methods face several limitations in boundary-critical settings. First, tokenization and window partitioning can attenuate high-frequency information. Coarse patching or aggressive downsampling tends to smooth thin structures and weaken sharp contours. Second, attention-based models often require large-scale and diverse training data. In medical imaging, limited annotations and substantial appearance variation can lead to overfitting without stronger priors or explicit regularization.

Third, although hierarchical attention reduces the quadratic complexity of global attention,

memory usage still grows rapidly at the high resolutions typical of dermoscopic images, limiting the use of fine-grained, boundary-preserving features. Finally, window partitioning and positional encoding may introduce discontinuities at window boundaries. Without dedicated alignment mechanisms, these artifacts can conflict with anatomical contours, reducing boundary accuracy even when global consistency is improved.

2.3. Hybrid segmentation methods

Hybrid architectures aim to unify the strengths of CNNs and Transformers by combining convolutional locality with attention-based global reasoning. Representative methods, including U-Net with Channel Transformer (UCTransNet) [24], Dual Swin Transformer U-Net (DS-TransUNet) [25], and TransFuse [26], adopt parallel convolutional and Transformer branches and integrate them using dedicated fusion designs.

In general, convolutional components provide precise spatial localization and texture cues, whereas Transformer components contribute holistic context and long-range consistency. Fusion can be performed at early, intermediate, or late stages and is frequently implemented through cross-attention, gating, or feature-recalibration modules to adaptively exchange information between branches.

Although hybrid methods have reported strong performance on International Skin Imaging Collaboration (ISIC) benchmarks, several issues remain. Many approaches prioritize region-level overlap metrics and pixel-wise accuracy while under-specifying boundary modeling, which can yield ambiguous or jagged contours in heterogeneous regions. In addition, heterogeneous encoders produce features with different resolutions, receptive fields, and statistical distributions. Naive concatenation or additive fusion may blur edges or misalign fine structures without explicit alignment.

Moreover, dual-pathway designs and deep fusion stacks often increase parameter counts and inference latency, limiting applicability on resource-constrained platforms. These observations motivate boundary-aware alignment and multiscale fusion mechanisms that can effectively translate global context into clinically meaningful contour precision.

Recent lightweight alternatives further investigate efficient local–global modeling. SM-UNet combines axial shift with a spatial state-space model to learn local and global lesion representations [27], whereas frequency-aware directional shift modeling adapts long-range interactions to different frequency components [28].

Recent hybrid architectures further emphasize structured fusion of complementary representations. Hybrid serial–parallel fusion network (HSPF-Net) adopts a serial–parallel hybrid CNN–Transformer architecture and introduces multi-receptive-field feature fusion to integrate features obtained from different receptive fields [29]. Progressive multi-granularity attention fusion network (PMGAF-Net), although developed for polyp segmentation, integrates Pyramid Vision Transformer and CNN-based representations and employs progressive multi-granularity attention together with weighted hybrid attention fusion to improve cross-layer feature integration [30]. These studies further demonstrate the importance of coordinating complementary local and global representations through dedicated fusion mechanisms. In contrast, BGDNet focuses specifically on using feature-derived boundary priors to regulate bidirectional correspondence between heterogeneous encoder representations rather than relying on general feature fusion alone.

2.4. Boundary-aware segmentation methods

Boundary-aware methods seek to improve contour fidelity by introducing explicit objectives or architectural components that emphasize edge information. Boundary loss [16] addresses class imbalance by supervising the surface distance between predicted and reference contours, encouraging accurate boundary localization even when interior pixels dominate. Boundary-enhanced segmentation network (BESNet) [31] adopts separate region and boundary streams to enable mutual enhancement, whereas Hausdorff-based supervision [32] improves robustness to shape variability by penalizing large contour deviations. More recent designs such as body and edge aware network (BEA-Net) [33] integrate edge cues into backbone features to better preserve sharp structural details.

Recent work has also explored frequency–spatial collaborative modeling for improving boundary representation in low-contrast dermoscopic images. Self-attention-based cross-layer fusion network (SAFNet) jointly models frequency- and spatial-domain features and introduces a wavelet-deformable attention gate to exploit high-frequency edge cues while adapting spatial feature sampling [34]. Its cross-layer fusion mechanism further promotes interaction between features at different levels. This frequency–spatial strategy provides a complementary approach to boundary enhancement. In contrast, BGDNet does not rely on frequency decomposition to enhance high-frequency image components. Instead, fixed Sobel responses extracted from intermediate encoder features are transformed into feature-level boundary priors and used to condition cross-branch feature interaction.

Boundary-aware segmentation remains difficult because lesion contours are sparse, frequently ambiguous, and susceptible to annotation uncertainty. Many existing approaches use boundary information as an auxiliary prediction target, an edge-focused decoder cue, or an additional loss term. Although these strategies can improve contour fidelity, they do not explicitly address the statistical and semantic mismatch that may arise when heterogeneous local and global representations are fused. BGDNet instead uses boundary priors as conditioning signals for bidirectional cross-branch interaction, allowing boundary information to regulate the alignment between TransXNet and Swin Transformer features before decoder-side reconstruction.

3. Materials and method

3.1. Overall architecture

Figure 2 presents the overall architecture of BGDNet. The network consists of a Swin Transformer encoder $\Phi_{\text{Swin}}(\cdot)$, a TransXNet encoder $\Phi_{\text{TransXNet}}(\cdot)$, stage-specific channel-projection layers, BEMs, BGCA, a multi-scale deformable decoder (MSDC), and two prediction heads.

Given an input image $I \in \mathbb{R}^{3 \times 352 \times 352}$, the two encoders produce four-stage hierarchical features. At the i th stage, the heterogeneous features are first projected into a common feature space. Boundary priors are then extracted, followed by boundary-guided bidirectional interaction:

$$\begin{aligned} \bar{S}_i &= \pi_i^S(\Phi_{\text{Swin}}(I)_i), & \bar{T}_i &= \pi_i^T(\Phi_{\text{TransXNet}}(I)_i), \\ P_i^S &= \text{BEM}_i^S(\bar{S}_i), & P_i^T &= \text{BEM}_i^T(\bar{T}_i), \\ (S'_i, T'_i) &= \text{BGCA}_i(\bar{S}_i, \bar{T}_i, P_i^S, P_i^T), \\ Z_i &= [S'_i, T'_i], & i &= 0, \dots, 3. \end{aligned} \tag{3.1}$$

The native Swin feature dimensions are [48, 96, 192, 384], whereas the TransXNet feature dimensions are [64, 128, 320, 512]. The projection layers map both branches to the common channel dimensions [64, 128, 320, 512] before boundary-guided interaction.

For an input resolution of 352×352 , the four fused feature maps have approximate dimensions $128 \times 176 \times 176$, $256 \times 88 \times 88$, $640 \times 44 \times 44$, and $1024 \times 22 \times 22$. These hierarchical features are progressively aligned and aggregated by the decoder. The final prediction heads produce the lesion-mask logits M and the auxiliary boundary logits B .

The BEM-derived boundary priors and the auxiliary boundary supervision serve different purposes. The former are generated directly from intermediate encoder features and participate in BGCA during both training and inference. In contrast, the auxiliary boundary target is derived from the ground-truth segmentation mask and is used only to supervise the boundary prediction head during training.

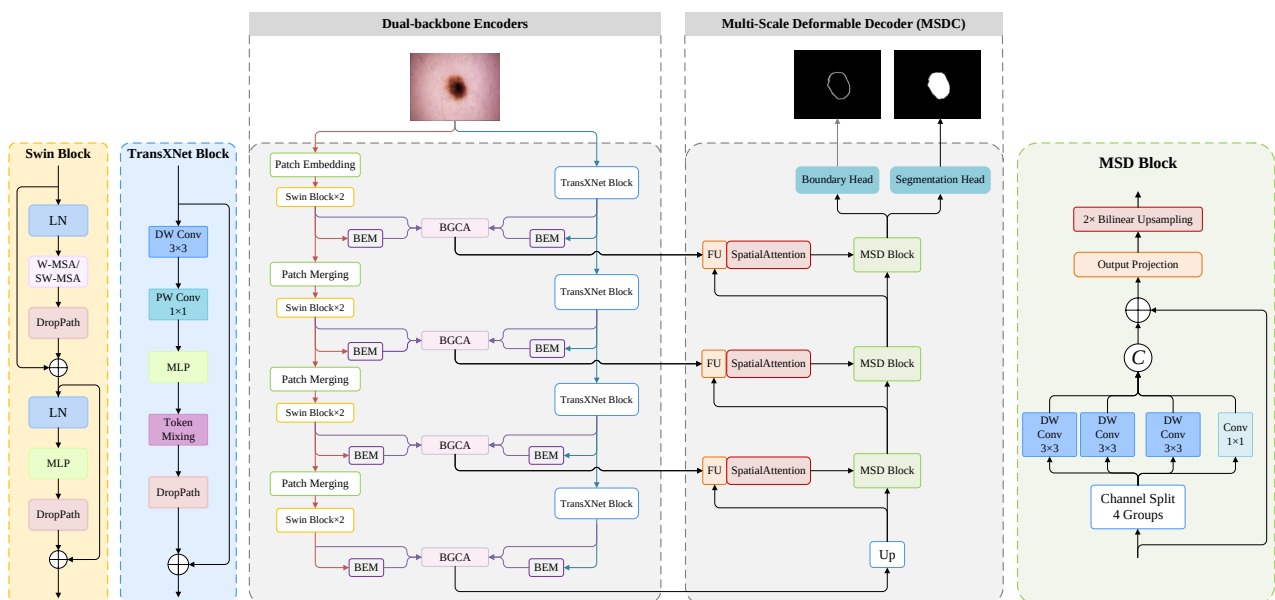


Figure 2. Overall architecture of BGDNet. The Swin Transformer and TransXNet branches provide complementary hierarchical representations. At each encoder stage, the corresponding features are first mapped into compatible feature spaces. BEM then extracts feature-level boundary priors, which are used by BGCA to guide bidirectional cross-branch interaction. The enhanced features from the two branches are concatenated and progressively reconstructed by MSDC through deformable feature alignment, spatial attention, and multi-scale deformable aggregation. Two independent 1×1 prediction heads generate the lesion-mask logits M and auxiliary boundary logits B . The internal MSD block subgraph illustrates its channel projection, parallel multi-scale branches, residual fusion, output projection, and bilinear upsampling.

3.2. Dual-backbone encoders

The global-context branch is implemented using a four-stage Swin Transformer. In the present configuration, the initial embedding dimension is 24, the patch size is 2×2 , and the window size is 7×7 . The four stages contain (2, 2, 2, 2) Swin blocks with (3, 6, 12, 24) attention heads, respectively. This branch produces hierarchical features with channel dimensions [48, 96, 192, 384].

The local-structure branch is implemented using TransXNet-XXS [17], which produces four feature maps with channel dimensions [64, 128, 320, 512]. Its convolution-oriented operations and dynamic token mixing preserve local continuity and fine structural details, complementing the long-range contextual representation learned by the Swin Transformer.

Because the two encoders have different channel dimensions and feature statistics, independent 1×1 projections are used at every stage. These projections provide a common feature space before boundary extraction and cross-branch interaction.

3.3. Boundary extraction module

BEM extracts feature-level contour responses using fixed Sobel operators. Figure 3 presents its structure.

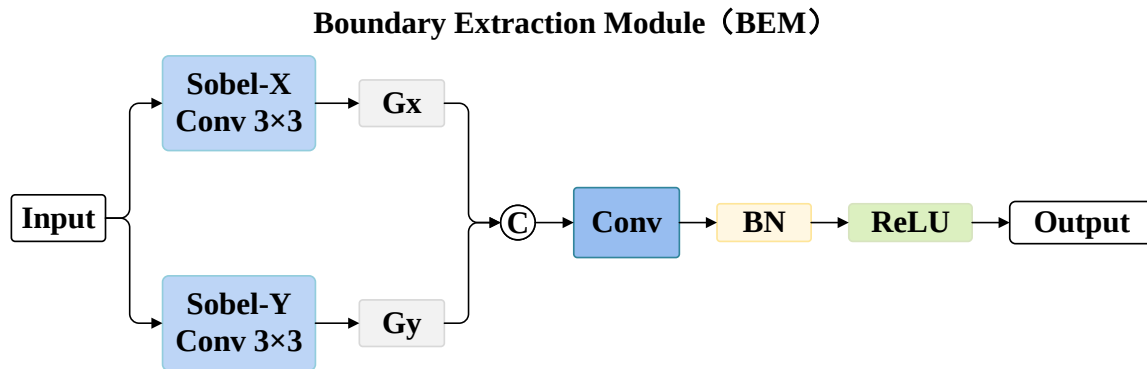


Figure 3. Boundary extraction module (BEM). Fixed Sobel-X and Sobel-Y kernels are applied independently to every feature channel using depthwise convolution. The resulting horizontal and vertical responses are concatenated and projected by a learnable 1×1 convolution followed by batch normalization and rectified linear unit activation.

Let $X_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ denote either a projected Swin feature or a projected TransXNet feature. The fixed horizontal and vertical Sobel kernels [35] are defined as

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \quad (3.2)$$

The two kernels are applied independently to each feature channel by depthwise convolution with groups = C_i . The two Sobel responses are then concatenated and directly projected to obtain the boundary prior:

$$P_i = \text{ReLU}\left[\text{BN}_i\left(\text{Conv}_{1\times 1,i}\left([X_i * K_x, X_i * K_y]\right)\right)\right]. \quad (3.3)$$

Here, $*$ denotes convolution, and $[\cdot, \cdot]$ denotes channel-wise concatenation. The Sobel kernels are fixed during training, whereas the 1×1 projection and batch-normalization layer are learnable. Independent BEM instances are used for the two encoder branches and for all feature stages.

The BEM output is a feature-level boundary prior, not a direct pixel-level lesion boundary. Priors from shallow stages keep finer spatial changes, while priors from deeper, lower-resolution features become coarser. They may also respond to non-lesion structures or feature artifacts. Therefore, BEM is used to guide feature interaction rather than to predict a precise clinical boundary.

3.4. Boundary-guided cross-attention

The Swin Transformer and TransXNet branches provide complementary features, but their feature distributions are different. Direct cross-attention can therefore produce unreliable correspondence near weak or ambiguous lesion boundaries. BGCA uses the BEM priors in the query and key pathways so that boundary information can guide cross-branch matching, as illustrated in Figure 4.

At stage i , let $\bar{S}_i$ and $\bar{T}_i$ denote the projected Swin and TransXNet features, and let P_i^S and P_i^T denote their BEM priors. For the TransXNet $\rightarrow$ Swin direction, the Swin feature provides the query, while the TransXNet feature provides the key and value:

$$\begin{aligned} Q_i^S &= W_q^S(\bar{S}_i + P_i^S), & K_i^T &= W_k^T(\bar{T}_i + P_i^T), & V_i^T &= W_v^T\bar{T}_i, \\ Q_i^T &= W_q^T(\bar{T}_i + P_i^T), & K_i^S &= W_k^S(\bar{S}_i + P_i^S), & V_i^S &= W_v^S\bar{S}_i. \end{aligned} \quad (3.4)$$

The second row gives the reverse Swin $\rightarrow$ TransXNet direction. The conditioned query, key, and value representations are passed to the corresponding stage-specific cross-attention operators. To keep the formulation focused on the proposed boundary-conditioning mechanism, the internal attention realization is denoted abstractly by $\mathcal{A}_i^{T\rightarrow S}$ and $\mathcal{A}_i^{S\rightarrow T}$ rather than expanded into a particular dense token-level implementation.

After cross-branch interaction, an output projection and a residual connection update the two branches:

$$\begin{aligned} S'_i &= \text{BN}_i^S\left[\bar{S}_i + W_o^S\mathcal{A}_i^{T\rightarrow S}(Q_i^S, K_i^T, V_i^T)\right], \\ T'_i &= \text{BN}_i^T\left[\bar{T}_i + W_o^T\mathcal{A}_i^{S\rightarrow T}(Q_i^T, K_i^S, V_i^S)\right]. \end{aligned} \quad (3.5)$$

The operators $W_q^{S/T}$, $W_k^{S/T}$, $W_v^{S/T}$, and $W_o^{S/T}$ are bias-free 1×1 convolutions. Boundary priors are added only when constructing the query and key. The value features remain $\bar{S}_i$ or $\bar{T}_i$, so the boundary prior guides cross-branch correspondence without replacing the semantic information being transferred. The two updated features S'_i and T'_i are then concatenated to form Z_i as defined in Equation (3.1).

Here, $\bar{S}_i$ and $\bar{T}_i$ are the channel-projected encoder features, whereas P_i^S and P_i^T are obtained after Sobel filtering, learnable 1×1 projection, batch normalization, and activation. Therefore, the addition in Equation (3.4) is used only to guide query and key construction; it is not a direct addition of raw Sobel gradients to encoder features.

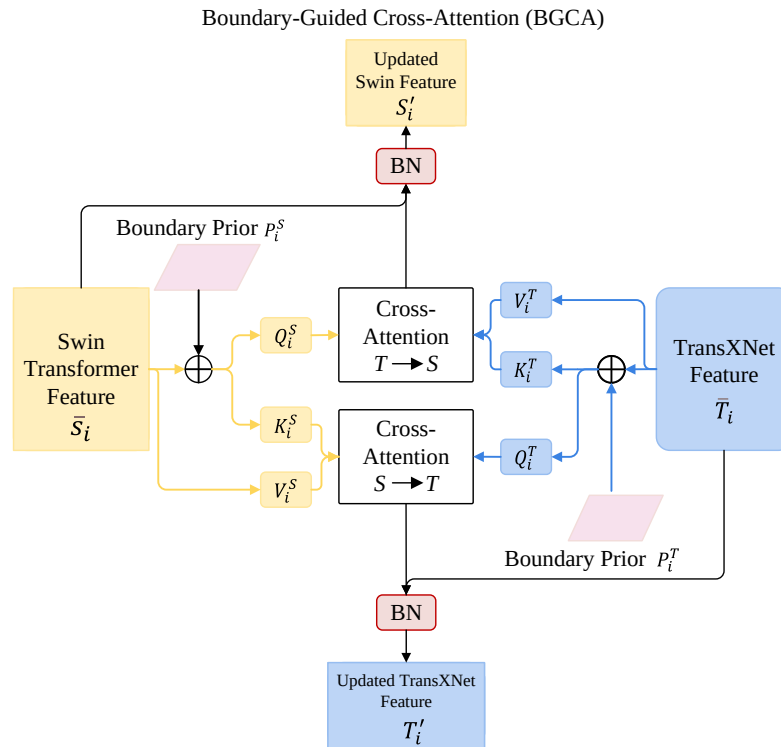


Figure 4. Boundary-guided cross-attention (BGCA). In the TransXNet→Swin direction, the Swin feature and its boundary prior construct the query, while the TransXNet feature and its boundary prior construct the key and the unconditioned TransXNet feature provides the value. The reverse Swin→TransXNet branch updates the TransXNet feature analogously. Boundary priors therefore condition only the query and key pathways. Each attention response is projected, combined with the target-branch feature through residual addition, and normalized. The two updated representations are finally concatenated.

3.5. Multi-scale deformable decoder

The MSDC used in BGDNet is adapted from the decoder introduced in MSDUNet [36]. It is retained as the reconstruction pathway of BGDNet and is not claimed as an independent methodological contribution.

The decoder combines deformable skip-feature alignment [37] with multi-scale deformable aggregation. Let Z_0 , Z_1 , Z_2 , and Z_3 denote the fused BGCA features from high to low spatial resolution.

For the deepest decoder stage, $X_3 = \text{Up}_3(Z_3)$, whereas $X_j = D_{j+1}$ for $j = 1, 2$. The alignment kernel size is $k_j = 2j + 1$, giving 3×3 , 5×5 , and 7×7 kernels from the high-resolution to the low-resolution stage. At stage j , X_j predicts the sampling offsets used to align Z_{j-1} , and the aligned feature is then refined by spatial attention:

$$\begin{aligned} \widehat{Z}_{j-1} &= \text{DeformConv}_{k_j \times k_j, d=2}(Z_{j-1}; \text{Conv}_{k_j \times k_j, d=2}^{\text{off}}(X_j)), \\ D_j &= \mathcal{D}_j(X_j, \text{SA}(\widehat{Z}_{j-1})), \quad j = 1, 2, 3. \end{aligned} \quad (3.6)$$

The spatial-attention operator $SA(\cdot)$ follows the channel-pooling design of the convolutional block attention module (CBAM) [38]: It uses channel-wise average and maximum pooling, followed by a 7×7 convolution and sigmoid activation.

Within each decoder block $\mathcal{D}_j$, the current decoder feature and the aligned skip feature are concatenated, normalized, and projected using a 1×1 convolution with Gaussian error linear unit activation. The projected feature is then evenly divided into four channel groups.

The first group is processed using a regular 1×1 convolution. The remaining three groups are processed by independent depthwise deformable convolutions with dilation rates 1, 2, and 3. Each deformable branch predicts its own offset field from the corresponding channel group. The four branch outputs are concatenated and added to the projected feature through a residual connection.

A second 1×1 projection is subsequently applied. Bilinear upsampling and layer-scale refinement then produce the output of the current decoder stage. These operations are represented collectively by $\mathcal{D}_j(\cdot)$ in Equation (3.6), because they are standard internal operations of the decoder block rather than separate methodological contributions.

The four parallel branches of the MSD block are shown explicitly in the internal MSD block subgraph of Figure 2, so that the text and architecture diagram describe the same inference sequence.

For a 3×3 deformable kernel, each internal offset predictor outputs 18 channels, corresponding to horizontal and vertical offsets for the nine sampling locations. The offset predictors are not shared across decoder stages or dilation branches. No modulation mask is used, and the offset weights are trained jointly with the remaining network parameters without explicit zero initialization.

The offset-prediction layers contain approximately 3.663 million learnable parameters in total. At an input resolution of 352×352 , they contribute approximately 10.817 giga multiply–accumulate operations (GMACs), corresponding to approximately 21.633 giga floating-point operations (GFLOPs) when one multiply–accumulate operation is counted as two floating-point operations.

Finally, two independent 1×1 convolutional heads map the highest-resolution decoder feature D_1 to the lesion-mask logits M and the auxiliary boundary logits B , respectively.

3.6. Boundary-aware training objective

The boundary target is generated directly from the ground-truth lesion mask M_{gt} . Using the Sobel kernels in Equation (3.2), the gradient magnitude is computed as

$$G_{\text{gt}} = \sqrt{(M_{\text{gt}} * K_x)^2 + (M_{\text{gt}} * K_y)^2}. \quad (3.7)$$

The gradient magnitude is then normalized to obtain the soft boundary target:

$$B_{\text{gt}} = \frac{G_{\text{gt}}}{\max_{u,v} G_{\text{gt}}(u, v) + \epsilon}. \quad (3.8)$$

For lesion segmentation, the Dice loss is defined as

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum \sigma(M) M_{\text{gt}} + \epsilon}{\sum \sigma(M)^2 + \sum M_{\text{gt}}^2 + \epsilon}. \quad (3.9)$$

The final training objective combines binary cross-entropy (BCE) for lesion segmentation, Dice loss, and auxiliary boundary supervision:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(M, M_{\text{gt}}) + \mathcal{L}_{\text{Dice}} + \lambda_{\text{bnd}} \mathcal{L}_{\text{BCE}}(B, B_{\text{gt}}). \quad (3.10)$$

Here, $\mathcal{L}_{\text{BCE}}$ denotes BCE computed directly from logits, and $\sigma(\cdot)$ denotes the sigmoid function. The maximum is calculated over all spatial locations (u, v) in each training sample. The constant ϵ ensures numerical stability. The boundary-loss weight is set to $\lambda_{\text{bnd}} = 0.4$. The Dice term directly emphasizes foreground overlap and reduces the influence of background dominance.

No threshold is applied during boundary-target generation. Therefore, $B_{\text{gt}} \in [0, 1]$ represents a normalized soft contour map rather than a fixed-width binary morphological boundary. This soft target retains variations in local edge strength and avoids the need to introduce an additional boundary-width hyperparameter.

Because the soft boundary target is derived from a manually annotated mask, it may be affected by annotation variability, resizing, imaging artifacts, and local boundary uncertainty. A Sobel response of the annotation should therefore not be interpreted as a unique biological lesion boundary. The boundary term remains auxiliary, while the BCE–Dice segmentation objective provides the primary supervision.

3.7. Experimental settings and datasets

All experiments were implemented in PyTorch [39] and conducted on a single NVIDIA Tesla V100 graphics processing unit with 16 GB of memory. Images were resized to 352×352 and normalized using the ImageNet channel means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225). The corresponding lesion masks were resized to the same spatial resolution and converted to floating-point tensors.

The models were trained for 200 epochs with a batch size of 4. AdamW [40] was used with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} . Parameter gradients were clipped element-wise to the interval $[-0.5, 0.5]$. The learning-rate decay factor was 0.1, and the decay epoch was set to 200.

No random data augmentation was enabled in the principal experiments. The training loader shuffled the samples and discarded the final incomplete batch. The Swin branch was initialized using the Swin-Tiny ImageNet-22K pretrained checkpoint, whereas the TransXNet branch was initialized from the corresponding pretrained TransXNet checkpoint. Newly introduced BEM, BGCA, decoder, deformable-offset, and prediction-head parameters were optimized jointly with the two encoders.

The reported optimization and loss hyperparameters were pre-specified and kept fixed across the reported BGDNet experiments. The 200 epochs specify the maximum optimization horizon rather than the epoch automatically used for testing. Checkpoint selection was based on the mean validation Dice similarity coefficient (DSC) of the lesion-mask prediction. The checkpoint with the highest mean validation DSC was retained for final evaluation, and the held-out test sets did not participate in hyperparameter tuning or checkpoint selection. Weight decay and gradient clipping were used as regularization and optimization-stabilization measures.

During inference, the segmentation and boundary logits were bilinearly interpolated to the spatial resolution of the corresponding ground-truth mask. A sigmoid function was then applied, and a threshold of 0.5 was used to obtain the binary lesion prediction. No morphological post-processing or test-time augmentation was used.

Experiments were conducted on ISIC-2016 [41], ISIC-2018 [42], and PH2 [43]. For ISIC-2018, the official 1000-image test set was retained for final evaluation, while the original training pool was

divided into 2334 training images and 260 validation images. For ISIC-2016, the official 379-image test set was retained, while the original training pool was divided into 810 training images and 90 validation images. For PH2, the 200 images were randomly divided into 140 training, 20 validation, and 40 test images. The resulting PH2 split was then kept fixed across random seeds and model configurations.

All repeated BGDNet experiments used the same three random seeds: 2026, 2027, and 2028. Repeated-run results are reported as mean $\pm$ standard deviation (SD).

The models were trained independently on the corresponding training subset of each dataset. In particular, the PH2 results were obtained from models trained on PH2 and therefore represent in-dataset evaluation rather than zero-shot testing using ISIC-trained weights.

The comparison results of previously published segmentation methods were collected from the corresponding original publications or from published comparative studies reporting the respective methods and datasets. These baseline values were not obtained by retraining all comparison methods under the BGDNet experimental pipeline. Consequently, differences in dataset partitioning, preprocessing, training strategy, pretrained initialization, and evaluation implementation among the source studies should be taken into account when interpreting the literature-based comparisons.

Cross-dataset experiments were conducted separately by selecting the source-domain model using its source validation subset and applying the selected model directly to the held-out PH2 test subset without target-domain fine-tuning or target-label-based model selection.

3.8. Evaluation metrics

The region-based evaluation includes intersection over union (IoU), DSC, recall (sensitivity), and specificity. For every test image, the prediction logits are bilinearly interpolated to the ground-truth resolution, converted to probabilities using the sigmoid function, and binarized using a threshold of 0.5.

Let TP, FP, FN, and TN denote the numbers of true-positive, false-positive, false-negative, and true-negative pixels, respectively. The region-based metrics are defined as

$$\begin{aligned} \text{IoU} &= \frac{\text{TP} + 1}{\text{TP} + \text{FP} + \text{FN} + 1}, & \text{DSC} &= \frac{2\text{TP} + 1}{2\text{TP} + \text{FP} + \text{FN} + 1}, \\ \text{Recall (Sensitivity)} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, & \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}. \end{aligned} \quad (3.11)$$

All region-based metrics are calculated independently for each image and subsequently averaged over the complete test set. Consequently, the reported mean IoU and mean DSC do not necessarily satisfy the exact algebraic conversion that applies when pixel counts are pooled globally.

The additive constant 1 in IoU and DSC follows the evaluation implementation and prevents undefined values for empty foreground masks.

Recall, also termed sensitivity in binary segmentation, quantifies the proportion of lesion pixels correctly recovered, whereas specificity measures the proportion of background pixels correctly rejected. These two metrics are additionally reported on ISIC-2018 to characterize foreground–background behavior under class imbalance.

Because BGDNet is designed to improve lesion contours, the 95th percentile Hausdorff distance (HD95), average surface distance (ASD), and boundary F-score at a 2-pixel tolerance (BF@2) are additionally reported. For boundary-sensitive evaluation, both the predicted binary mask and the

ground-truth mask are resized to 352×352 before their contours are extracted. Nearest-neighbor interpolation is used for this binary-mask resizing. HD95 and ASD are therefore reported in pixels at the 352×352 evaluation resolution.

Let ∂P and ∂G denote the predicted and ground-truth boundaries. For a boundary point x and a boundary set S , the point-to-set distance is

$$d(x, S) = \min_{y \in S} \|x - y\|_2.$$

Let $\mathcal{D}$ denote the multiset containing distances from every point in ∂P to ∂G and from every point in ∂G to ∂P . Repeated distance values are retained.

For a tolerance τ , P_{bnd} denotes the proportion of predicted boundary pixels located within τ pixels of the ground-truth boundary, whereas R_{bnd} denotes the proportion of ground-truth boundary pixels located within τ pixels of the predicted boundary. The boundary-sensitive metrics are defined as

$$\begin{aligned} \text{HD95} &= \text{Percentile}_{95}(\mathcal{D}), \\ \text{ASD} &= \frac{1}{2} \left[\frac{1}{|\partial P|} \sum_{p \in \partial P} d(p, \partial G) + \frac{1}{|\partial G|} \sum_{g \in \partial G} d(g, \partial P) \right], \\ \text{BF@2} &= \frac{2P_{\text{bnd}}R_{\text{bnd}}}{P_{\text{bnd}} + R_{\text{bnd}} + \epsilon}, \quad \tau = 2. \end{aligned} \quad (3.12)$$

Lower HD95 and ASD and higher BF@2 indicate better boundary localization. The constant ϵ in BF@2 is used only for numerical stability when both boundary precision and boundary recall are zero.

4. Results

4.1. Quantitative comparison

Table 1 reports the main comparison on ISIC-2018 using both region-based and boundary-sensitive metrics.

BGDNet achieves strong performance in both region-overlap and boundary-sensitive evaluation. On ISIC-2018, BGDNet obtains an IoU of 0.8813 ± 0.0014 , a DSC of 0.9330 ± 0.0009 , an HD95 of 5.84 ± 0.11 pixels, an ASD of 1.34 ± 0.03 pixels, and a BF@2 of 0.868 ± 0.003 . Its recall (sensitivity) is 0.9602 ± 0.0021 , while the specificity is 0.9847 ± 0.0007 . The comparison values for previously published methods are literature-reported values collected from their original publications or published comparative studies where available. Because these values may follow different dataset partitions, preprocessing procedures, and evaluation protocols, the cross-method comparison should be interpreted as a literature-based performance comparison rather than a strictly controlled reimplementation study.

To examine whether the performance observed on ISIC-2018 is maintained on additional datasets, Table 2 reports the BGDNet results together with literature-reported comparisons on ISIC-2016 and PH2.

As shown in Table 2, BGDNet achieves the best listed region-overlap and boundary-localization results on both ISIC-2016 and PH2. Together with the ISIC-2018 results, these findings show that the performance improvements are maintained across the three evaluated datasets.

Table 1. Quantitative comparison on ISIC-2018. HD95 and ASD are measured in pixels. Higher IoU, DSC, BF@2, recall (sensitivity), and specificity are better, whereas lower HD95 and ASD are better. BGDNet is reported as mean $\pm$ SD over three runs.

Method	IoU $\uparrow$	DSC $\uparrow$	HD95 $\downarrow$	ASD $\downarrow$	BF@2 $\uparrow$	recall $\uparrow$	specificity $\uparrow$
U-Net [7]	0.7733	0.8545	12.84	3.12	0.741	0.8813	0.9643
nnU-Net [44]	0.8379	0.9016	9.31	2.34	0.789	0.9102	0.9736
BATFormer [45]	0.8464	0.9076	8.74	2.10	0.801	0.9122	0.9748
TransUNet [14]	0.8462	0.9102	9.02	2.18	0.796	0.9132	0.9751
TransFuse [26]	0.8465	0.9108	8.46	2.03	0.808	0.9139	0.9756
PVT-GCASCADe [46]	0.8553	0.9146	7.82	1.86	0.821	0.9266	0.9762
D-LKA [47]	0.8564	0.9164	7.44	1.79	0.828	0.9194	0.9773
MSDUNet [36]	0.8553	0.9168	7.16	1.71	0.834	0.9203	0.9781
SM-UNet [27]	0.8250	0.9023	9.68	2.41	0.782	0.8942	0.9721
FMLP-UNet [28]	0.8618	0.9258	6.42	1.52	0.847	0.9186	0.9802
BGDNet (ours)	0.8813 ± 0.0014	0.9330 ± 0.0009	5.84 ± 0.11	1.34 ± 0.03	0.868 ± 0.003	0.9602 ± 0.0021	0.9847 ± 0.0007

Table 2. In-dataset evaluation on ISIC-2016 and PH2. HD95 and ASD are measured in pixels. Higher IoU, DSC, and BF@2 are better, whereas lower HD95 and ASD are better. BGDNet is reported as mean $\pm$ SD over three runs.

Method	ISIC-2016					PH2				
	IoU $\uparrow$	DSC $\uparrow$	HD95 $\downarrow$	ASD $\downarrow$	BF@2 $\uparrow$	IoU $\uparrow$	DSC $\uparrow$	HD95 $\downarrow$	ASD $\downarrow$	BF@2 $\uparrow$
TransUNet [14]	0.8489	0.9126	8.21	1.96	0.814	0.8411	0.9098	7.44	1.82	0.826
TransFuse [26]	0.8619	0.9203	7.52	1.78	0.827	0.8546	0.9183	6.81	1.65	0.841
D-LKA [47]	0.8715	0.9283	6.98	1.63	0.839	0.8529	0.9165	6.94	1.69	0.837
MSDUNet [36]	0.8763	0.9308	6.61	1.55	0.846	0.8630	0.9240	6.32	1.52	0.849
BGDNet (ours)	0.9091 ± 0.0015	0.9450 ± 0.0011	4.98 ± 0.14	1.15 ± 0.03	0.883 ± 0.003	0.9246 ± 0.0019	0.9564 ± 0.0013	4.08 ± 0.15	0.99 ± 0.04	0.898 ± 0.004

4.2. Cross-dataset generalization

Table 3 reports cross-dataset performance on the held-out PH2 test subset using the evaluation protocol described in Section 3.7.

Cross-dataset performance is lower than the in-dataset PH2 result, as expected under dataset shift. Nevertheless, both source-trained models retain useful lesion-overlap and boundary-localization performance on the unseen PH2 test subset without target-domain fine-tuning. These results support partial cross-dataset robustness rather than complete domain invariance.

Table 3. Cross-dataset generalization to the held-out PH2 test subset without target-domain fine-tuning. Results are mean $\pm$ SD over three runs.

Source	Target	IoU $\uparrow$	DSC $\uparrow$	HD95 $\downarrow$	ASD $\downarrow$	BF@2 $\uparrow$
ISIC-2018	PH2 test	0.8578 ± 0.0027	0.9234 ± 0.0018	7.66 ± 0.21	1.79 ± 0.05	0.833 ± 0.004
ISIC-2016	PH2 test	0.8462 ± 0.0031	0.9167 ± 0.0022	8.14 ± 0.27	1.92 ± 0.06	0.821 ± 0.005

4.3. Boundary-focused ablation and prior analysis

Table 4 evaluates the effects of the dual encoders, bidirectional cross-attention, BEM, and BGCA while keeping MSDC fixed as the common reconstruction pathway in all ablation configurations. The baseline consists of a U-Net-based single-encoder segmentation framework coupled with the same MSDC used in the remaining configurations. This design keeps the reconstruction pathway unchanged so that the subsequent comparisons primarily examine the effects of encoder configuration and cross-branch boundary-guided interaction. “Bi-CA” denotes the same bidirectional interaction structure as BGCA but without BEM-derived boundary-prior conditioning. In the BEM-only configuration, each BEM prior is added to its corresponding projected encoder feature before dual-branch concatenation.

Table 4. Boundary-focused ablation study on ISIC-2018. Higher IoU, DSC, and BF@2 are better, whereas lower HD95 and ASD are better.

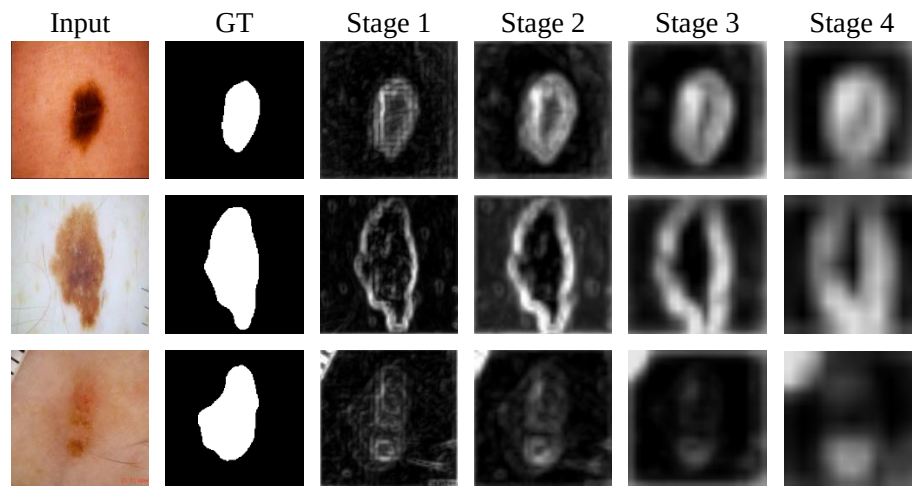
Setting	Dual Enc.	Bi-CA	BEM	BGCA	IoU ↑	DSC ↑	HD95 ↓	ASD ↓	BF@2 ↑
U-Net + MSDC baseline	–	–	–	–	0.8142 ± 0.0008	0.8882 ± 0.0005	11.10 ± 0.15	2.76 ± 0.19	0.749 ± 0.013
Dual encoders	✓	–	–	–	0.8389 ± 0.0009	0.9059 ± 0.0006	9.30 ± 0.12	2.27 ± 0.12	0.781 ± 0.009
Dual + Bi-CA	✓	✓	–	–	0.8265 ± 0.0009	0.8964 ± 0.0005	10.04 ± 0.11	2.49 ± 0.08	0.766 ± 0.007
Dual + BEM	✓	–	✓	–	0.8669 ± 0.0011	0.9241 ± 0.0007	6.91 ± 0.13	1.630 ± 0.009	0.839 ± 0.002
Full BGDNet	✓	–	✓	✓	0.8813 ± 0.0014	0.9330 ± 0.0009	5.84 ± 0.11	1.34 ± 0.03	0.868 ± 0.003

Compared with the U-Net + MSDC baseline, introducing the heterogeneous dual encoders improves both region-based and boundary-sensitive performance. Conventional Bi-CA does not improve the dual-encoder configuration, whereas direct incorporation of BEM priors produces a clear gain. The complete BGDNet achieves the best results, indicating that the improvement does not arise only from adding boundary priors; using the priors to guide cross-branch interaction provides a further benefit.

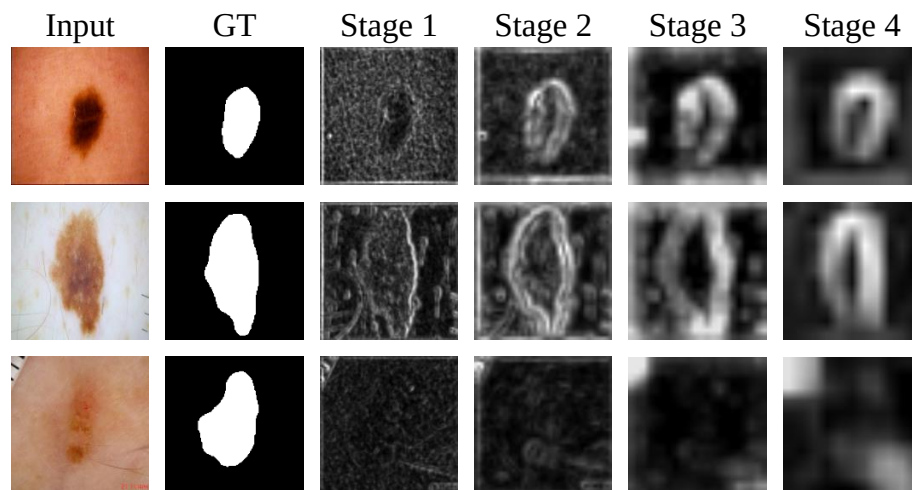
To examine the fixed Sobel design, the Sobel branches in BEM were replaced by learnable depthwise 3×3 convolutions while keeping the remaining architecture, data split, training protocol, and random seeds unchanged. The learnable version achieved an IoU of 0.8765 ± 0.0006 , a DSC of 0.9297 ± 0.0004 , an HD95 of 6.18 ± 0.07 pixels, an ASD of 1.44 ± 0.02 pixels, and a BF@2 of 0.859 ± 0.002 . The fixed Sobel configuration retained slightly better performance under this controlled setting, supporting the adopted fixed prior without implying that fixed operators are universally superior.

To examine how the BEM priors change with encoder depth, responses from all four stages of both encoder branches are visualized in Figure 5. For visualization only, the multi-channel responses are averaged across channels, independently normalized, and upsampled to the input resolution; these display operations do not alter the inference pipeline.

The displayed cases show finer lesion-related transitions in shallow stages and progressively coarser responses with increasing depth. Stage 4 may also contain artifact-related activations. These observations support interpreting deep BEM outputs as feature-level boundary priors for attention guidance rather than pixel-accurate lesion contours.



(a) Swin Transformer branch



(b) TransXNet branch

Figure 5. Stage-wise visualization of BEM boundary priors for (a) the Swin Transformer branch and (b) the TransXNet branch. Each row shows the input image, ground-truth (GT) mask, and projected BEM responses from Stages 1–4. For display, the responses are averaged across channels, independently normalized, and upsampled to the input resolution. Shallow stages retain finer spatial transitions, whereas deeper stages show coarser boundary-prior responses.

4.4. Qualitative results

Figure 6 shows visual comparisons on representative cases.

Compared with the baselines, BGDNet better preserves lesion extent and produces cleaner contours, especially in low-contrast regions with ambiguous boundaries. The improvements are most visible along irregular margins and thin protrusions. These visual observations are consistent with the quantitative and ablation results.

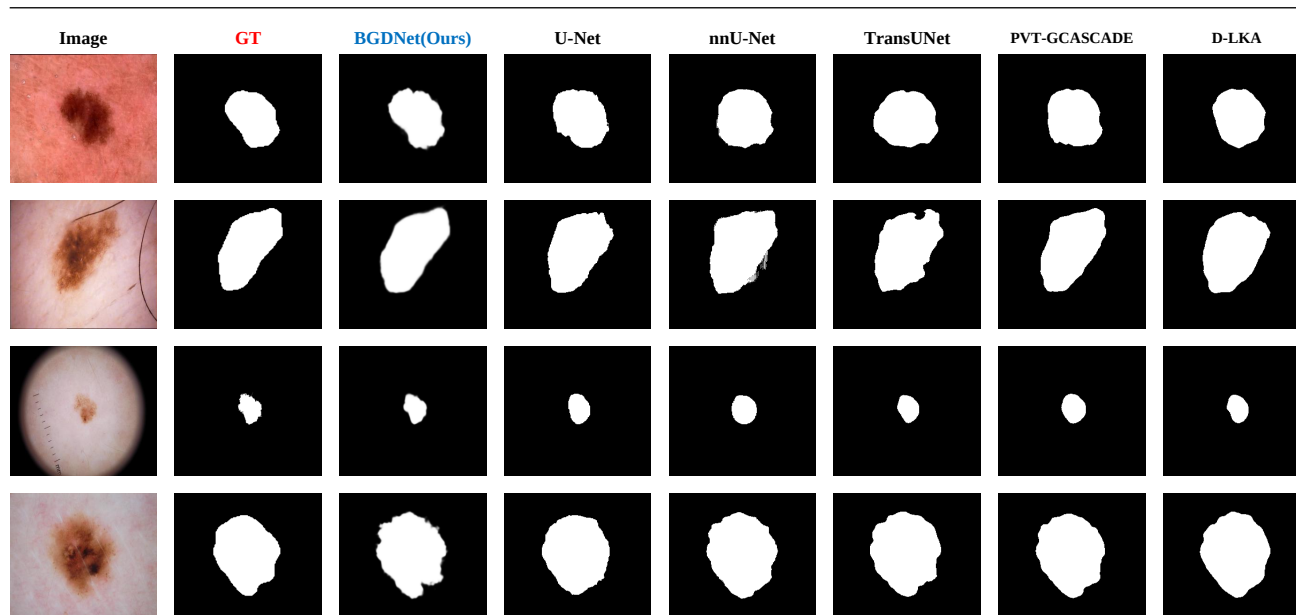


Figure 6. Qualitative comparison of lesion segmentation. From left to right: input image, ground truth, BGDNet, U-Net, nnU-Net, TransUNet, PVT-GCASCADe, and D-LKA.

5. Discussion

Across the reported experiments, BGDNet shows consistent improvements in both lesion-region overlap and boundary localization. The joint improvements in region-based and boundary-sensitive metrics indicate that the proposed design benefits not only the overall lesion area but also the accuracy of lesion contours.

The principal architectural contribution of BGDNet lies in the use of feature-derived boundary priors to regulate bidirectional interaction between heterogeneous encoders. BEM applies fixed channel-wise Sobel operators to intermediate features and transforms the resulting responses through a learnable 1×1 projection. BGCA adds these priors to the query and key inputs, thereby influencing correspondence weights while retaining the original projected branch features as values.

The component-wise ablation further clarifies the role of boundary guidance. The dual-encoder configuration improves over the U-Net + MSDC baseline, whereas conventional bidirectional cross-attention does not improve the heterogeneous feature interaction. Direct incorporation of BEM priors produces a clear gain, and BGCA further improves performance by using the same priors to guide cross-branch attention. The fixed-versus-learnable comparison also supports the adopted Sobel prior in the present architecture without implying that fixed operators are universally superior.

The additive conditioning in BGCA should not be interpreted as adding raw Sobel gradients directly to semantic encoder features. The encoder features are first projected into compatible channel spaces, whereas the BEM priors are obtained after Sobel filtering, learnable projection, batch normalization, and activation. Boundary information therefore modifies query–key correspondence while the value pathways retain the semantic features being transferred.

The stage-wise visualization further supports this interpretation. Shallow BEM responses retain relatively fine spatial transitions, whereas deeper responses become coarser and more semantic. Deep

BEM outputs should therefore be interpreted as feature-level boundary priors for attention guidance rather than as pixel-accurate lesion contours.

The boundary-supervision target is generated using a normalized Sobel gradient magnitude. This soft target avoids the need to select a fixed morphological boundary width, but it remains dependent on the quality and interpolation of the original segmentation annotation. Annotation uncertainty, gradual lesion transitions, and strong artifact edges may therefore influence the auxiliary boundary signal.

The recall (sensitivity) and specificity metrics provide complementary information under foreground–background imbalance. The reported recall indicates strong recovery of lesion pixels, while the high specificity reflects effective rejection of background pixels. The Dice component further emphasizes foreground overlap during optimization.

Cross-dataset evaluation provides a complementary assessment under dataset shift. Although performance decreases relative to PH2-trained models, the ISIC-trained models retain useful region-overlap and boundary-localization performance on the held-out PH2 test subset without target-domain fine-tuning. These results support partial cross-dataset robustness rather than complete domain invariance.

The deformable reconstruction pathway should be distinguished from the newly proposed components. MSDC is adapted from MSDUNet [36] and is retained as the reconstruction pathway in the component-wise comparisons; it is therefore not regarded as an independent methodological contribution of BGDNet.

6. Limitations and future work

Several limitations remain. First, fixed Sobel operators provide a simple first-order boundary prior but cannot adapt their local gradient operators to every lesion appearance. Deep-stage BEM responses are also spatially coarse and may include non-lesion or artifact-related activations.

Second, the annotation-derived soft boundary target cannot explicitly represent inter-observer uncertainty. Multi-annotator labels, probabilistic contours, signed-distance representations, or uncertainty-aware targets may provide more faithful supervision.

Third, bidirectional cross-attention and deformable reconstruction increase parameter count, computation, and memory requirements. Efficient or linear attention, lighter encoders, knowledge distillation, structured pruning, and reduced deformable computation are therefore relevant future directions.

Fourth, the cross-dataset experiment uses PH2 as the unseen target and therefore provides only a limited assessment of dataset shift. The three public datasets do not cover the full diversity of acquisition devices, clinical centers, skin appearances, and annotation practices. External multi-center validation with fixed trained models remains necessary before making claims of clinical readiness.

Fifth, the quantitative comparisons with previously published methods include literature-reported results rather than uniformly retrained implementations of all competing architectures. Differences in dataset partitioning, preprocessing, pretrained initialization, training strategy, and metric implementation among the source studies may therefore affect direct numerical comparability. A fully controlled benchmark in which all competing methods are retrained and evaluated under an identical protocol would provide a stronger basis for future comparisons.

7. Conclusion

This work proposed BGDNet, a boundary-guided dual-backbone encoder–decoder network for dermoscopic lesion segmentation. The model combines a hierarchical Swin Transformer with TransXNet and uses channel-wise Sobel feature responses to condition bidirectional cross-branch attention. Deformable skip-feature alignment and multi-scale deformable aggregation are then used to progressively reconstruct the lesion mask and auxiliary boundary map.

Experiments on ISIC-2016, ISIC-2018, and PH2 demonstrate strong in-dataset region-overlap and boundary-localization performance for BGDNet. Cross-dataset experiments further show that ISIC-trained models retain useful performance on the held-out PH2 test subset without target-domain fine-tuning.

The ablation results indicate that the benefit of BGDNet does not arise only from introducing boundary priors. Using these priors to guide bidirectional interaction between the heterogeneous encoders provides a further improvement over direct boundary-prior incorporation and conventional bidirectional cross-attention. The stage-wise visualization also supports interpreting deep BEM responses as feature-level boundary priors rather than precise lesion contours.

Overall, the results support feature-derived boundary conditioning as an effective strategy for coordinating complementary local and contextual representations. Future work will investigate more efficient cross-attention, uncertainty-aware boundary supervision, stronger robustness to artifact edges, lightweight dual-backbone variants, and external multi-center validation.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

Jiayu Zhang: Conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft preparation, and visualization.

Haiyang Liu: Validation, data curation, and writing—review and editing.

Ying Ye: Validation and writing—review and editing.

Xiaoming Huang: Supervision, project administration, and writing—review and editing.

All authors have read and agreed to the published version of the manuscript.

Data availability statement

The ISIC-2016, ISIC-2018, and PH2 datasets used in this study are publicly available from their respective repositories. The exact dataset-split files, evaluation configuration, and implementation details are available from the corresponding author upon reasonable request.

References

1. Sung H, Ferlay J, Siegel RL, et al. (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 71: 209–249. <https://doi.org/10.3322/caac.21660>
2. Garbe C and Leiter U (2009) Melanoma epidemiology and trends. *Clin Dermatol* 27: 3–9. <https://doi.org/10.1016/j.clindermatol.2008.09.001>
3. Gershenwald JE and Scolyer RA (2018) Melanoma staging: American Joint Committee on Cancer (AJCC) 8th edition and beyond. *Ann Surg Oncol* 25: 2105–2110. <https://doi.org/10.1245/s10434-018-6513-7>
4. Kittler H, Pehamberger H, Wolff K, et al. (2002) Diagnostic accuracy of dermoscopy. *Lancet Oncol* 3: 159–165. [https://doi.org/10.1016/S1470-2045\(02\)00679-4](https://doi.org/10.1016/S1470-2045(02)00679-4)
5. Mihulecea CRJ, Iancu GM, Leventer M, et al. (2023) The many roles of dermoscopy in melanoma detection. *Life* 13: 477. <https://doi.org/10.3390/life13020477>
6. Codella NCF, Gutman D, Celebi ME, et al. (2018) Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging, hosted by the International Skin Imaging Collaboration, 2018 IEEE 15th International Symposium on Biomedical Imaging, USA, 168–172. <https://doi.org/10.1109/ISBI.2018.8363547>
7. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, Cham: Springer, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
8. Oktay O, Schlemper J, Le Folgoc L, et al. (2018) Attention U-Net: learning where to look for the pancreas. *arXiv*: 1804.03999. Available from: <https://arxiv.org/abs/1804.03999>.
9. Yu F and Koltun V (2016) Multi-scale context aggregation by dilated convolutions. *arXiv*: 1511.07122. Available from: <https://arxiv.org/abs/1511.07122>.
10. Lin TY, Dollár P, Girshick R, et al. (2017) Feature pyramid networks for object detection, 2017 IEEE Conference on Computer Vision and Pattern Recognition, USA, 936–944. <https://doi.org/10.1109/CVPR.2017.106>
11. Dosovitskiy A, Beyer L, Kolesnikov A, et al. (2021) An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv*: 2010.11929. Available from: <https://arxiv.org/abs/2010.11929>.
12. Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need, In: Guyon I, Von Luxburg U, Bengio S, et al., *Advances in Neural Information Processing Systems* 30, 5998–6008. Available from: <https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need>.
13. Liu Z, Lin Y, Cao Y, et al. (2021) Swin transformer: hierarchical vision transformer using shifted windows, 2021 IEEE/CVF International Conference on Computer Vision, Canada, 9992–10002. <https://doi.org/10.1109/ICCV48922.2021.00986>
14. Chen J, Mei J, Li X, et al. (2024) TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med Image Anal* 97: 103280. <https://doi.org/10.1016/j.media.2024.103280>

15. Hatamizadeh A, Tang Y, Nath V, et al. (2022) UNETR: transformers for 3D medical image segmentation, *2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, USA, 1748–1758. <https://doi.org/10.1109/WACV51458.2022.00181>
16. Kervadec H, Bouchtiba J, Desrosiers C, et al. (2021) Boundary loss for highly unbalanced segmentation. *Med Image Anal* 67: 101851. <https://doi.org/10.1016/j.media.2020.101851>
17. Lou M, Zhang S, Zhou HY, et al. (2025) TransXNet: learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *IEEE Trans Neural Netw Learn Syst* 36: 11534–11547. <https://doi.org/10.1109/TNNLS.2025.3550979>
18. Milletari F, Navab N, Ahmadi SA (2016) V-Net: fully convolutional neural networks for volumetric medical image segmentation, *2016 Fourth International Conference on 3D Vision*, USA, 565–571. <https://doi.org/10.1109/3DV.2016.79>
19. Zhou Z, Siddiquee MMR, Tajbakhsh N, et al. (2018) UNet++: a nested U-Net architecture for medical image segmentation, In: Stoyanov D, Taylor Z, Carneiro G, et al., *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Cham: Springer, 3–11. https://doi.org/10.1007/978-3-030-00889-5_1
20. Jha D, Smedsrud PH, Riegler MA, et al. (2019) ResUNet++: an advanced architecture for medical image segmentation, *2019 IEEE International Symposium on Multimedia*, USA, 225–230. <https://doi.org/10.1109/ISM46123.2019.00049>
21. Fan DP, Ji GP, Zhou T, et al. (2020) PraNet: parallel reverse attention network for polyp segmentation, In: Martel AL, Abolmaesumi P, Stoyanov D, et al., *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2020*, Cham: Springer, 263–273. https://doi.org/10.1007/978-3-030-59725-2_26
22. Cao H, Wang Y, Chen J, et al. (2023) Swin-Unet: unet-like pure transformer for medical image segmentation, In: Karlinsky L, Michaeli T, Nishino K, *Computer Vision—ECCV 2022 Workshops*, Cham: Springer, 205–218. https://doi.org/10.1007/978-3-031-25066-8_9
23. Valanarasu MJ, Oza P, Hacihaliloglu I, et al. (2021) Medical transformer: gated axial-attention for medical image segmentation, In: de Bruijne M, Cattin PC, Cotin S, et al., *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2021*, Cham: Springer, 36–46. https://doi.org/10.1007/978-3-030-87193-2_4
24. Wang H, Cao P, Wang J, et al. (2022) UCTransNet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer. *Proc AAAI Conf Artif Intell* 36: 2441–2449. <https://doi.org/10.1609/aaai.v36i3.20144>
25. Lin A, Chen B, Xu J, et al. (2022) DS-TransUNet: dual swin transformer U-Net for medical image segmentation. *IEEE Trans Instrum Meas* 71: 4005615. <https://doi.org/10.1109/TIM.2022.3178991>
26. Zhang Y, Liu H, Hu Q (2021) TransFuse: fusing Transformers and CNNs for medical image segmentation, In: de Bruijne M, Cattin PC, Cotin S, et al., *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2021*, Cham: Springer, 14–24. https://doi.org/10.1007/978-3-030-87193-2_2
27. Li G, Huang Q, Wang W, et al. (2025) Skin lesion segmentation by fusing local and global features using axial shift and spatial state model. *Appl Soft Comput* 178: 113261. <https://doi.org/10.1016/j.asoc.2025.113261>

28. Li G, Huang Q, Dong XF (2026) Enhancing long-range contextual interactions via frequency-aware directional shifts for medical image segmentation. *Appl Soft Comput* 199: 115395. <https://doi.org/10.1016/j.asoc.2026.115395>
29. Fang H, Sun Y, Zhang S, et al. (2026) HSPF-Net: hybrid CNN-transformer with serial-parallel fusion for skin lesion segmentation. *Digit Signal Process* 172: 105891. <https://doi.org/10.1016/j.dsp.2026.105891>
30. Zhang S, Fang H, Teng X, et al. (2026) PMGAF-Net: progressive multi-granularity attention fusion network for accurate polyp segmentation. *Biomed Signal Process Control* 127: 111179. <https://doi.org/10.1016/j.bspc.2026.111179>
31. Oda H, Roth HR, Chiba K, et al. (2018) BESNet: boundary-enhanced segmentation of cells in histopathological images, In: Frangi A, Schnabel J, Davatzikos C, *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2018*, Cham: Springer, 228–236. https://doi.org/10.1007/978-3-030-00934-2_26
32. Karimi D and Salcudean SE (2020) Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Trans Med Imaging* 39: 499–513. <https://doi.org/10.1109/TMI.2019.2930068>
33. Kuang H, Wang Y, Liang Y, et al. (2023) BEA-Net: body and edge aware network with multi-scale short-term concatenation for medical image segmentation. *IEEE J Biomed Health Inform* 27: 4828–4839. <https://doi.org/10.1109/JBHI.2023.3304662>
34. Fang H, Zhang S, Teng X, et al. (2025) SAFNet: a self-attention based cross layer fusion network with frequency enhancement for accurate skin lesion segmentation. *Expert Syst Appl* 305: 130823. <https://doi.org/10.1016/j.eswa.2025.130823>
35. Sobel I and Feldman G (1968) A 3×3 isotropic gradient operator for image processing. Presented at the Stanford Artificial Intelligence Project, Stanford University, Stanford, CA, USA.
36. Li X, Li L, Xing X, et al. (2025) MSDUNet: a model based on feature multi-scale and dual-input dynamic enhancement for skin lesion segmentation. *IEEE Trans Med Imaging* 44: 2819–2830. <https://doi.org/10.1109/TMI.2025.3549011>
37. Dai J, Qi H, Xiong Y, et al. (2017) Deformable convolutional networks, *2017 IEEE International Conference on Computer Vision*, Italy, 764–773. <https://doi.org/10.1109/ICCV.2017.89>
38. Woo S, Park J, Lee JY, et al. (2018) CBAM: convolutional block attention module, In: Ferrari V, Hebert M, Sminchisescu C, et al., *Computer Vision—ECCV 2018*, Cham: Springer, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
39. Paszke A, Gross S, Massa F, et al. (2019) PyTorch: an imperative style, high-performance deep learning library, In: Wallach H, Larochelle H, Beygelzimer A, et al., *Advances in Neural Information Processing Systems* 32, 8024–8035. Available from: <https://proceedings.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library>.
40. Loshchilov I and Hutter F (2019) Decoupled weight decay regularization. *arXiv*: 1711.05101. <https://doi.org/10.48550/arXiv.1711.05101>

41. Gutman D, Codella NCF, Celebi ME, et al. (2016) Skin lesion analysis toward melanoma detection: a challenge at ISBI 2016, hosted by the International Skin Imaging Collaboration. *arXiv*: 1605.01397. <https://doi.org/10.48550/arXiv.1605.01397>
42. Codella NCF, Rotemberg V, Tschandl P, et al. (2019) Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration. *arXiv*: 1902.03368. <https://doi.org/10.48550/arXiv.1902.03368>
43. Mendonça T, Ferreira PM, Marques JS, et al. (2013) PH2—A dermoscopic image database for research and benchmarking, *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Japan, 5437–5440. <https://doi.org/10.1109/EMBC.2013.6610779>
44. Isensee F, Jaeger PF, Kohl SAA, et al. (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 18: 203–211. <https://doi.org/10.1038/s41592-020-01008-z>
45. Lin X, Yu L, Cheng KT, et al. (2023) BATFormer: towards boundary-aware lightweight Transformer for efficient medical image segmentation. *IEEE J Biomed Health Inform* 27: 3501–3512. <https://doi.org/10.1109/JBHI.2023.3266977>
46. Rahman MM and Marculescu R (2024) G-CASCADE: efficient cascaded graph convolutional decoding for 2D medical image segmentation, *2024 IEEE/CVF Winter Conference on Applications of Computer Vision*, USA, 7713–7722. <https://doi.org/10.1109/WACV57701.2024.00755>
47. Azad R, Niggemeier L, Hüttemann M, et al. (2024) Beyond self-attention: deformable large kernel attention for medical image segmentation, *2024 IEEE/CVF Winter Conference on Applications of Computer Vision*, USA, 1276–1286. <https://doi.org/10.1109/WACV57701.2024.00132>



AIMS Press

© 2026 Zhang et al., licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)