SDG-UNet: Medical Image Segmentation Based on Structure-Guided Dynamic Convolution and Dual-Path Hybrid Attention

Jiajian Hao[1], Jiaqing Mo[1,*], Gang Zhou[1], Zifang Zhao[1]
[1]*Key Laboratory of Signal Detection and Processing,*
Department of Computer Science and Technology, Xinjiang University, Urumqi, China
[*]Corresponding author. Email: 2226966386@qq.com

***Abstract*—Medical image segmentation requires both global contextual reasoning and precise boundary preservation, while shallow skip features often introduce semantic noise. To address this, we propose SDG-UNet, which combines a dual-path context encoder for global-local representation learning with a structure-guided dynamic filtering module for skip calibration. We introduce a Dual-Path Context Encoder (DPCE), which integrates a Global-Cross Collaborative Module (GCCM) to capture long-range dependencies and an Edge-aware Local Activation (ELA) module to reinforce fine-grained boundaries, thereby jointly modeling multi-scale global semantics and fine-grained boundary information. Furthermore, we propose a Structure-Guided Hybrid Dynamic Filtering (SG-HDF) module in the decoder. This module utilizes high-level structural priors to generate adaptive kernels that calibrate shallow skip features, effectively suppressing noise and aligning semantics before fusion. We evaluated the proposed method on three different medical image segmentation benchmarks. Experiments on three medical segmentation benchmarks show that SDG-UNet achieves competitive or superior performance with a favorable accuracy-efficiency trade-off.**

***Index Terms*—Medical image segmentation, hybrid attention, dynamic convolution, multi-scale representation learning, encoder-decoder networks.**

I. Introduction

Medical image segmentation remains challenging due to low contrast, imaging artifacts, large morphological variations, and ambiguous boundaries. These characteristics require models to jointly capture long-range context and fine structural details.

CNN-based U-shaped architectures remain strong baselines for medical image segmentation, but their locality-biased receptive fields limit long-range dependency modeling. To address this issue, Vision Transformers (ViT) [1] and hybrid architectures such as TransUNet [2] have been introduced to improve global context modeling. However, these methods often incur higher computational cost and stronger dependence on large-scale annotated data, which limits their practicality in resource-constrained medical scenarios.

Meanwhile, many existing attention mechanisms still process spatial and channel dimensions in a relatively decoupled manner. For example, CBAM [3] applies channel and spatial attention sequentially, while CPCA [4] further improves structural feature modeling through channel-prioritized interaction. Nevertheless, fully exploiting cross-dimensional semantic dependencies in complex medical images remains difficult. In addition, shallow encoder features delivered through skip connections often contain semantic noise and irrelevant background textures, which can lead to error propagation during decoding.

To address these issues, we propose SDG-UNet, a hybrid segmentation framework that combines dual-path contextual encoding with structure-guided skip calibration. Specifically, the proposed Dual-Path Context Encoder (DPCE) enhances encoder representations by jointly modeling global contextual dependencies and fine-grained structural details, while the Structure-Guided Hybrid Dynamic Filtering (SG-HDF) module uses deep structural priors to recalibrate shallow skip features before fusion. In summary, this work makes three main contributions: 1) we propose DPCE to improve global-local representation learning in the encoder; 2) we introduce SG-HDF to suppress semantic noise and enhance skip-feature consistency in the decoder; and 3) extensive experiments on three medical image segmentation benchmarks demonstrate that SDG-UNet achieves competitive or superior performance with a favorable accuracy-efficiency trade-off.

II. Related Work

A. Multi-scale Feature Modeling and Efficient Attention Mechanisms

The effective encoding of multi-semantic spatial information is fundamental to constructing robust model representations. Foundational architectures, such as the early InceptionNets, initiated the practice of feature fusion through multi-branch parallel convolutions, which laid the foundation for subsequent adaptive feature modeling strategies. While Transformer-based architectures (e.g., ViT [1]) demonstrate superior proficiency in global modeling through Multi-Head Self-Attention, they are frequently hampered by prohibitive computational burdens. To mitigate these overheads, recent research has focused on efficient attention mechanisms that leverage factorization or feature reshaping strategies. For example, ELA [5] preserves

This work was supported by the Xinjiang Uygur Autonomous Region Natural Science Foundation under Grant 2019D01C072, the Major Science and Technology Initiative of the Xinjiang Uygur Autonomous Region, China (Project No. 2022A01007), and the Tianshan Talent Training Project, the Technology Innovation Team Program (2023TSYCTD0012).

structural integrity by decoupling features along spatial dimensions. Seeking to further optimize the trade-off between efficiency and representation, LiteFANet [6] introduced a multi-branch fusion module designed to preserve long-range dependencies with a minimized parameter footprint. Nevertheless, the majority of existing hybrid mechanisms (e.g., CBAM [3]) predominantly rely on the simple serial or parallel stacking of spatial and channel attention. Consequently, these approaches often fail to fully exploit deep, cross-dimensional semantic correlations for mutual calibration, thereby limiting their robustness in challenging tasks that require the precise identification of fine-grained lesions.

B. Dynamic Convolution

Dynamic convolution has emerged as a promising paradigm in segmentation tasks, primarily due to its capacity to adaptively modulate kernel parameters in response to specific input characteristics [7]. However, a significant limitation of this approach is its reliance on a solitary attention scalar; consequently, all output filters are constrained to share an identical attention value, a design choice that neglects spatial and channel-specific distinctions and thereby restricts the depth of feature exploration. Recent studies have explored more diverse kernel generation strategies: Pinwheel [8] innovates by reshaping the convolution range into a pinwheel configuration, whereas FDConv [9] approaches the problem from a frequency domain perspective. In the specific context of medical imaging, DPConv [10] leverages reconstruction priors to optimize feature fitting capabilities. Nevertheless, effectively modeling the subtle, gradual transitions characteristic of physiological tissues in medical imagery remains a persistent challenge. This necessitates the development of more sophisticated, structure-guided kernel exploration strategies to ensure precise and reliable segmentation.

III. Method

SDG-UNet is built upon a standard U-shaped architecture (Fig. 1). To address limited receptive fields, we integrate a Dual-Path Context Encoder (DPCE) at each encoder stage, which employs parallel paths to simultaneously capture long-range global dependencies and reinforce local high-frequency boundary details. In the decoder, to bridge the semantic gap, we introduce the Structure-Guided Hybrid Dynamic Filtering (SG-HDF) module. Located at skip connections, SG-HDF leverages deep bottleneck features as structural priors to dynamically calibrate shallow features, effectively suppressing background noise before fusion. Finally, a segmentation head outputs pixel-level predictions.

A. Dual-Path Context Encoder (DPCE)

Medical image segmentation demands a balance between two conflicting objectives: locating large-scale anatomical structures (which requires global context) and delineating fuzzy lesion boundaries (which relies on local high-frequency details). Standard encoders often struggle to satisfy both

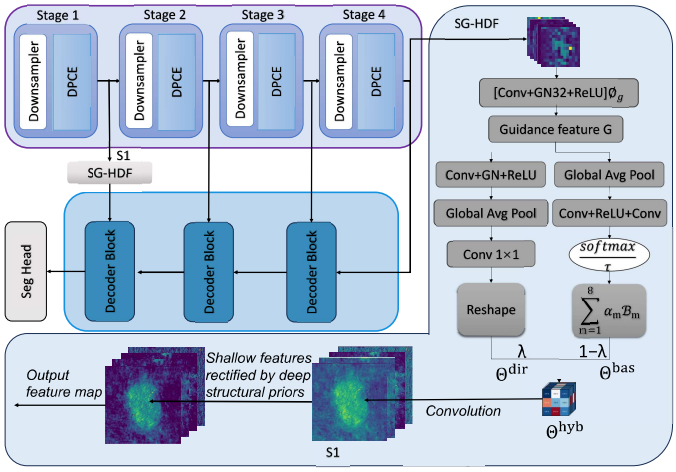

Fig. 1. The overall architecture of the proposed SDG-UNet and the details of the SG-HDF module.

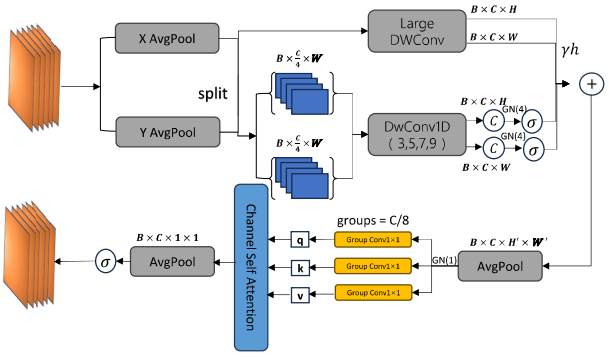

Fig. 2. Illustration of the Global-Cross Collaborative Module (GCCM).

requirements simultaneously. Therefore, we employ a parallel design at each stage of the encoder, integrating the GCCM and the ELA module. **Global-Cross Collaborative Modeling (GCCM).** The first path (Fig. 2) overcomes the limited receptive field of CNNs via an *Orthogonal Context Decomposition* strategy. This spatial decomposition branch is partially inspired by ELA [5]; however, GCCM further extends it with multi-scale axial processing, a parallel large-kernel depthwise branch, and lightweight channel interaction modeling for richer global-local representation.**(1) Axial Decomposition & Interaction:** Let $X \in \mathbb{R}^{C\times H\times W}$ denote the input. We first compress global context into orthogonal axis-vectors X^H and X^W via spatial pooling. To capture multi-scale features, these vectors are split into groups and processed by depthwise convolutions with varying kernel sizes ($\mathcal{K} = \{3, 5, 7, 9\}$). Crucially, a parallel branch with a large kernel ($k = 9$) injects a super-long-range background prior. The aggregated height feature $\tilde{X}^H$ is formulated as:

$$\tilde{X}^H = \text{Concat}_{k\in\mathcal{K}}(\mathcal{F}^k_{dw}(X^H_k)) + \gamma_h \cdot \mathcal{F}^{large}_{dw}(X^H) \quad (1)$$

where γ_h is a learnable scalar. Symmetrical operations yield $\tilde{X}^W$. **(2) Spatial Modulation:** These axial features are expanded and fused to generate a spatial attention map A_{sp}.

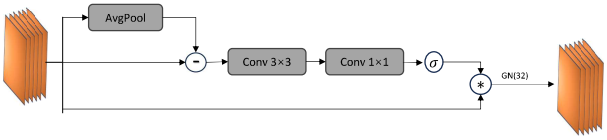

Fig. 3. Illustration of the Edge-Aware Local Activation (ELA) module.

The input is initially modulated as $X_{sp} = X \odot A_{sp}$.

(3) Lightweight Channel Self-Attention ($\mathcal{M}_{csa}$): To further refine inter-channel dependencies, the spatially-enhanced feature X_{sp} is fed into $\mathcal{M}_{csa}$. As shown in Fig. 2 (Right), we propose a parameter-efficient attention mechanism. First, to reduce computational overhead, X_{sp} is downsampled to Y. Then, Query (Q), Key (K), and Value (V) projections are generated via **Grouped Pointwise Convolutions** (groups=$C/8$), rather than standard linear layers. This design significantly reduces parameters while preserving channel subgroup interactions. Unlike spatial attention, we calculate the **Channel-to-Channel Cross-Covariance** to capture semantic dependencies:

$$\begin{aligned} Q, K, V &= \mathcal{F}^{1\times1}_{GConv}(Y), \\ Map_{attn} &= \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_{head}}}\right) \cdot V \end{aligned} \tag{2}$$

where Map_{attn} encodes the channel correlation matrix. Finally, a global average pooling is applied to extract a global channel weight vector $A_{channel} \in \mathbb{R}^{C\times1\times1}$, which modulates the features to yield the final output:

$$X_{gccm} = X_{sp} \odot \sigma(A_{channel}) \tag{3}$$

Edge-Aware Local Activation (ELA). Complementing the global path, ELA (Fig. 3) explicitly extracts high-frequency boundary cues. We subtract the local average from the input to obtain a high-pass residual signal $X_{high} = X - \mathcal{P}^{3\times3}_{avg}(X)$. This signal is transformed into a weight map to locally enhance boundaries:

$$X_{ela} = \text{GN}\big(X \odot \sigma(\mathcal{F}_{1\times1}(\mathcal{F}_{3\times3}(X_{high})))\big) \tag{4}$$

The outputs of GCCM and ELA are fused via a residual block, ensuring the encoder captures both semantic richness and structural precision.

B. Structure-Guided Hybrid Dynamic Filtering (SG-HDF)

To address the inherent "semantic gap" between shallow spatial details and deep semantic features in U-Net architectures, we propose the SG-HDF module (as shown in the decoder path of Fig. 1). Shallow features (s_1), while rich in boundary information, are often corrupted by background noise and irrelevant textures; conversely, deep features (s_4) are semantically explicit but lack spatial resolution. Direct concatenation leads to noise propagation and reduced segmentation accuracy. SG-HDF innovatively utilizes deep bottleneck features as Structural Priors to dynamically generate adaptive kernels that cleanse and calibrate shallow features.

The core innovation of SG-HDF lies in its kernel generation strategy. Unlike standard dynamic convolutions that rely solely on direct regression, SG-HDF employs a hybrid mechanism to synthesize kernels. This combines the generalization capability of Basis Decomposition with the instance-specific precision of Direct Regression. Let $G = \Phi(s_4)$ be the structural guidance feature derived from the deep bottleneck, and let $\Theta \in \mathbb{R}^{C\times1\times k\times k}$ be the target dynamic kernel intended to filter the shallow features. First, to capture common structural patterns shared across the dataset, we establish a bank of M learnable static basis kernels $\{\mathcal{B}_m\}_{m=1}^{M}$. A projection head predicts mixing coefficients $\alpha \in \mathbb{R}^M$ from the guidance feature G. These coefficients determine how to linearly combine the basis kernels to form a general filter:

$$\Theta_{basis} = \sum_{m=1}^{M} \frac{\exp(\alpha_m/\tau)}{\sum_j \exp(\alpha_j/\tau)} \cdot \mathcal{B}_m \tag{5}$$

where τ is a temperature parameter that regulates the sharpness of the attention distribution. In this work, we set $\tau = 0.4$ to impose a lower-temperature constraint, which forces the softmax distribution to become sharper. This *sparse selection strategy* encourages the model to select only a few most relevant basis kernels (or even a single dominant one) rather than averaging all bases. Consequently, the generated filters become highly specific to the current anatomical structure, effectively suppressing irrelevant background noise.

Complementing this basis-driven generalization, we employ a parallel branch to handle unique, sample-specific variations (e.g., irregular lesion shapes) that cannot be perfectly represented by a fixed basis bank. This branch directly regresses a residual kernel Θ_{dir} from G via a lightweight perceptron. The final dynamic kernel is constructed by adaptively fusing these two components, controlled by a learnable gating factor λ that balances generalization and specificity:

$$\tilde{\lambda} = \sigma(\lambda) \tag{6}$$

$$\Theta_{hybrid} = \tilde{\lambda} \cdot \Theta_{dir} + (1 - \tilde{\lambda}) \cdot \Theta_{basis} \tag{7}$$

Dynamic kernels generated in this manner can sometimes suffer from magnitude instability during early training, leading to convergence difficulties. To stabilize the optimization process, we apply **instance-wise kernel normalization** before the convolution operation. This ensures that the filters maintain a consistent scale:

$$\hat{\Theta} = \frac{\Theta_{hybrid} - \mu(\Theta_{hybrid})}{\|\Theta_{hybrid} - \mu(\Theta_{hybrid})\|_2 + \epsilon} \tag{8}$$

where $\mu(\cdot)$ computes the spatial mean of the kernel. Finally, this normalized dynamic kernel acts as a content-aware filter. It is applied to the noisy shallow skip feature s_1 via a batch-wise depthwise convolution:

$$\tilde{s}_1 = s_1 + \beta \cdot (s_1 * \hat{\Theta}) \tag{9}$$

where β is a learnable scale factor initialized to 0.15. Through this operation, deep semantic information effectively guides the filtering of shallow features, enhancing structure-relevant details (such as organ boundaries) while suppressing background noise. This ensures that the features passed to the decoder are semantically consistent and spatially precise.

TABLE I

QUANTITATIVE COMPARISON ON ISIC 2018 AND SEGPC 2021 DATASETS. PARAMETERS AND FLOPS ARE MEASURED AT AN INPUT RESOLUTION OF 224×224. BEST RESULTS ARE BOLDED. METHODS MARKED WITH * ARE REPORTED AS MEAN $\pm$ STANDARD DEVIATION OVER THREE INDEPENDENT RUNS.

Methods	Par.↓	FLOPs↓	ISIC 2018			SegPC 2021		
			AC↑	DSC↑	mIoU↑	AC↑	DSC↑	mIoU↑
TransUNet	94.03	25.83	94.52	84.99	83.65	97.02	82.33	83.38
SwinUNet	43.25	**8.96**	94.14	89.45	80.92	97.20	83.92	72.29
UCTransNet*	72.47	45.39	$94.45_{\pm 0.11}$	$90.01_{\pm 0.31}$	$81.84_{\pm 0.42}$	$98.02_{\pm 0.08}$	$89.25_{\pm 0.27}$	$88.58_{\pm 0.33}$
CPP-Net	**1.15**	9.67	93.18	87.69	84.54	97.21	83.99	84.70
ScribFormer	47.95	34.68	93.40	84.69	83.47	97.47	86.77	87.79
YOHO*	20.06	10.86	$94.15_{\pm 0.13}$	$86.03_{\pm 0.38}$	$86.24_{\pm 0.36}$	$97.90_{\pm 0.09}$	$88.41_{\pm 0.29}$	$88.75_{\pm 0.31}$
HTC-Net	58.95	12.59	93.90	88.31	85.57	97.95	88.55	88.61
Ours*	18.27	15.73	$\mathbf{95.01}_{\pm 0.09}$	$\mathbf{90.86}_{\pm 0.24}$	$\mathbf{87.31}_{\pm 0.28}$	$\mathbf{98.05}_{\pm 0.07}$	$\mathbf{89.77}_{\pm 0.21}$	$\mathbf{88.84}_{\pm 0.24}$

C. Decoder and Feature Reconstruction

The decoder mirrors the encoder stages to recover spatial resolution. In each stage, deep features are upsampled and concatenated with encoder counterparts. To align channel dimensions and reduce complexity, the fused features are processed by a 1×1 convolution followed by Group Normalization (GN) and ReLU activation. Subsequently, a Residual Block—comprising two stacked 3×3 convolutions (each with GN and ReLU) and an identity shortcut—is employed for feature refinement. Crucially, to mitigate the semantic gap, the shallowest skip feature s_1 is explicitly calibrated by the proposed SG-HDF module (Section III-B) prior to fusion, ensuring noise suppression. Finally, a 1×1 convolution maps the reconstructed features to the segmentation mask.

D. Loss Function

During training, we employ a hybrid loss function combining Cross-Entropy loss ($\mathcal{L}_{ce}$) and Dice loss ($\mathcal{L}_{dice}$) to address potential class imbalance and ensure region overlap accuracy:

$$\mathcal{L}_{total} = \gamma \mathcal{L}_{ce} + (1 - \gamma)\mathcal{L}_{dice} \tag{10}$$

where γ is set to 0.5. The Cross-Entropy loss penalizes pixel-level classification errors, while the Dice loss optimizes the intersection-over-union between the predicted and ground-truth masks, which is particularly effective for medical image segmentation tasks involving small or irregular targets.

IV. EXPERIMENTAL

A. Implementation Details

We implemented our model in PyTorch and trained it on a single NVIDIA GeForce RTX 3090 GPU with 24 GB memory. All input images were resized to 224×224 for consistency; for the 3D UCSF-PDGM dataset, 2D axial slices were extracted before resizing. For fair complexity comparison, the parameter counts and FLOPs of all methods were measured on the full models under the same input resolution of 224×224. Standard data augmentation, including random horizontal/vertical flipping and random rotation, was applied during training, and all input intensities were normalized to $[0, 1]$.

The networks were optimized using AdamW with an initial learning rate of 6×10^{-5} and a batch size of 6. The temperature parameter τ in the SG-HDF module was set to 0.4, and a Cosine Annealing scheduler was adopted to adjust the learning rate during training. According to the convergence behavior of different datasets, the maximum number of training epochs was set to 20 for UCSF-PDGM, 100 for ISIC 2018, and 80 for SegPC 2021.

For dataset partitioning, we used the official training/testing splits for ISIC 2018 and SegPC 2021, while a 5-fold cross-validation strategy was adopted for UCSF-PDGM. Although all images were resized to a unified resolution of 224×224 for fair comparison, the boundary-related conclusions in this work should be interpreted as relative improvements under the same low-resolution setting rather than full preservation of native high-resolution boundary details. All quantitative results are reported as the averages of three independent runs with different random seeds.

B. Comparative Analysis

In this section, we present experimental details and results on three datasets.We compare our model with state-of-the-art (SOTA) methods, including CNN-based CPPNet [11] and HTCNet [12], Transformer-like ScribFormer [13], and hybrid architectures like SwinUNet [14], TransUNet [2], YOHO [15], and UCTransNet [16]. All compared methods were reimplemented and trained under the same preprocessing protocol, input resolution, and dataset splits for fair comparison. We use Dice coefficient and mean Intersection-over-Union (mIoU) as primary metrics. For tasks where boundary quality is critical, we also consider Hausdorff Distance (HD). Qualitative comparisons are visualized in Fig. 4. As observed, SDG-UNet produces masks with smoother boundaries and better structural consistency across all three tasks.

1) ISIC 2018 Skin Lesion Segmentation: The ISIC 2018 dataset [17] contains 3,694 RGB dermoscopic images with ambiguous lesion boundaries and complex artifacts such as hair and illumination variations, making it a challenging benchmark for skin lesion segmentation. As shown in Table I, SDG-UNet achieves superior performance with a Dice score

TABLE II

QUANTITATIVE COMPARISON ON THE UCSF-PDGM DATASET. BEST RESULTS ARE BOLDED. METHODS MARKED WITH * ARE REPORTED AS MEAN ± STANDARD DEVIATION OVER THREE INDEPENDENT RUNS. EACH RUN FOLLOWS THE SAME 5-FOLD CROSS-VALIDATION PROTOCOL, AND THE REPORTED SCORE IS THE AVERAGE OVER THE FIVE FOLDS.

Methods	**mIoU↑**	**DSC↑**	**HD↓**	**Target mIoU↑**		
				NCR	**ED**	**ET**
TransUNet	65.11	73.64	9.04	60.47	70.34	64.51
SwinUNet	66.86	75.45	10.72	59.35	69.14	72.09
UCTransNet	66.34	76.04	4.65	57.00	70.94	71.07
CPP-Net*	68.61±0.52	77.00±0.46	**2.44**±0.24	61.47±0.91	**76.14**±0.78	68.24±1.02
ScribFormer	65.71	74.72	5.06	58.38	72.31	66.44
YOHO*	67.94±0.49	76.74±0.41	5.97±0.31	62.12±0.84	73.89±0.73	67.81±0.96
HTC-Net	64.67	74.87	5.12	57.43	69.52	67.06
Ours*	**70.45**±0.37	**78.09**±0.34	3.54±0.22	**64.47**±0.76	73.21±0.69	**73.67**±0.88

of 90.86% and an mIoU of 87.31% compared with state-of-the-art methods. Qualitatively, SDG-UNet produces masks that adhere more closely to ground-truth contours on lesions with irregular boundaries and subtle shape variations. This benefit mainly comes from the complementary design of ELA and GCCM: ELA captures fine high-frequency details for smoother boundary delineation, while GCCM models long-range dependencies to suppress background artifacts such as normal skin texture and hair, thereby reducing under-segmentation. Overall, the proposed global-local hybrid mechanism is particularly effective for lesions with fuzzy boundaries.

2) SegPC 2021 Plasma Cell Segmentation: The SegPC 2021 dataset [18] contains 498 bone marrow smear images with densely packed, low-contrast plasma cells, making small-object segmentation challenging. As shown in Table I, SDG-UNet achieves the best performance, with a Dice coefficient of 89.77% and an mIoU of 88.84%, outperforming methods like YOHO (88.41%) and ScribFormer (86.77%). SDG-UNet also strikes a favorable accuracy-efficiency balance, using only 18.27M parameters—fewer than most Transformer and hybrid baselines—while achieving the best segmentation results on both ISIC 2018 and SegPC 2021. Although its FLOPs are not the lowest, they are still lower than strong baselines like TransUNet, UCTransNet, and ScribFormer. These gains are attributed to the SG-HDF module, which enhances boundary-aware features and suppresses ambiguous background responses, enabling clearer separation in dense and overlapping cell regions.

3) UCSF-PDGM Brain Tumor Segmentation: The UCSF-PDGM dataset [19] contains 500 3D MRI volumes (155 slices, 240×240) with annotations for three tumor regions: Enhancing Tumor (ET), Necrotic (NCR), and Edema (ED). These tumors are often small, low-contrast, and morphologically variable, making them susceptible to detail loss during downsampling. As summarized in Table II, SDG-UNet achieves a Dice score of 78.09%, outperforming TransUNet (73.64%) and Swin-UNet (75.45%), while also obtaining a competitive Hausdorff Distance of 3.54, which is markedly lower than several representative baselines such as TransUNet (9.04). The model also delivers strong mIoU performance on the challenging ET and NCR regions. This improvement can be attributed to the complementary design of DPCE and SG-HDF: the dual-path encoder captures both long-range anatomical context and fine boundary cues, while SG-HDF leverages deep structural priors to recalibrate shallow features and better preserve small tumor sub-regions during decoding. It is worth noting that, under our preprocessing protocol, the UCSF-PDGM experiments are conducted on 2D axial slices rather than full 3D volumes; therefore, the reported results should be interpreted as slice-wise 2D segmentation performance.

C. Ablation Studies

We validate the effectiveness of the proposed DPCE (comprising GCCM and ELA) and SG-HDF module. Ablation results on UCSF-PDGM and ISIC 2018, measured by DSC and mIoU, are detailed in Table III.

1) Overall Module Ablation: We further analyze the contributions of the proposed modules on the UCSF-PDGM and ISIC 2018 datasets, as summarized in Table III. Row 1 corresponds to the plain baseline without ELA, GCCM, or SG-HDF. Rows 2 and 3 evaluate two asymmetric variants that retain SG-HDF while enabling only one encoder branch, i.e., ELA or GCCM. Row 4 employs the complete dual-path encoder (DPCE) but replaces SG-HDF with direct skip fusion, while Row 5 represents the full SDG-UNet.

From Table III, both encoder-side contextual modeling and decoder-side skip calibration contribute to the final performance, but their effects are not uniformly additive in isolation. Rows 2 and 3 show that enabling either ELA or GCCM together with SG-HDF already improves over the plain baseline on both datasets, indicating that both local boundary enhancement and global contextual aggregation are beneficial. However, Row 4 shows that using the complete dual-path encoder without SG-HDF does not consistently improve performance, and even leads to a slight drop on ISIC 2018. This suggests that the gains introduced by DPCE are more effectively realized when shallow skip features are further calibrated by structure-guided dynamic filtering. Overall, the full model (Row 5) achieves the best performance on both datasets, demonstrating that DPCE and SG-HDF are complementary.

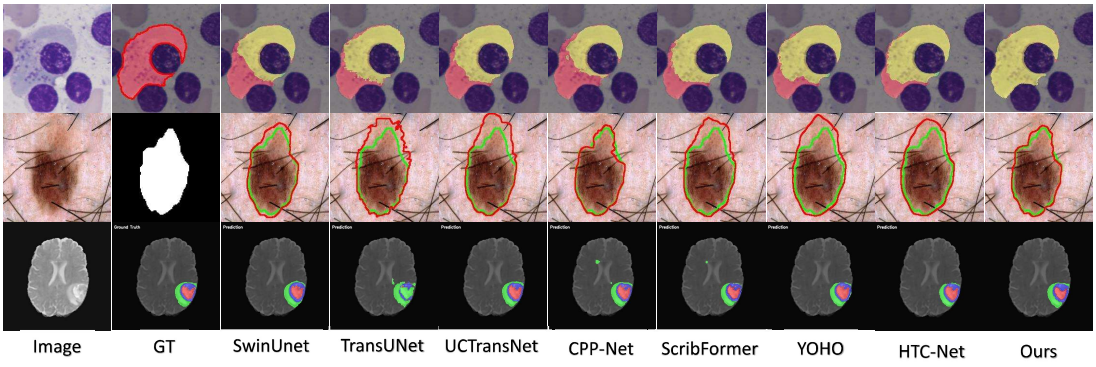

Fig. 4. Visualization of SDG-UNet results and comparison with different methods on the three benchmark datasets.

TABLE III
MODULE-WISE ABLATION OF ELA, GCCM, AND SG-HDF ON UCSF-PDGM AND ISIC 2018 DATASETS.

Modules			UCSF-PDGM		ISIC 2018	
ELA	GCCM	SG-HDF	DSC↑	mIoU↑	DSC↑	mIoU↑
×	×	×	75.60	66.84	89.80	85.46
✓	×	✓	77.02	68.88	90.20	86.15
×	✓	✓	77.95	70.24	90.35	86.42
✓	✓	×	77.25	69.22	89.45	84.86
✓	✓	✓	**78.09**	**70.45**	**90.86**	**87.31**

V. CONCLUSION

We presented SDG-UNet, which combines dual-path contextual encoding and structure-guided skip calibration for medical image segmentation. Experiments on dermoscopy, microscopy, and MRI benchmarks demonstrate competitive accuracy with favorable efficiency. The current evaluation is conducted under a resized 2D setting, and validating the method at native resolution or with richer volumetric context will be part of future work.

REFERENCES

[1] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[3] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[4] H. Huang, Z. Chen, Y. Zou, M. Lu, C. Chen, Y. Song, H. Zhang, and F. Yan, "Channel prior convolutional attention for medical image segmentation," *Computers in Biology and Medicine*, vol. 178, p. 108784, 2024.

[5] W. Xu and Y. Wan, "Ela: Efficient local attention for deep convolutional neural networks," *arXiv preprint arXiv:2403.01123*, 2024.

[6] K. Xu, X. Guo, B. Pan, Y. Zhang, Y. Liu, X. Zhang, X. Zeng, and Y. Liu, "Litefanet: A lightweight unet-based fusion-attention segmentation network for 2d and 3d medical images," *Biomedical Signal Processing and Control*, vol. 112, p. 108632, 2025.

[7] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 408–14 419.

[8] J. Yang, S. Liu, J. Wu, X. Su, N. Hai, and X. Huang, "Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9202–9210.

[9] L. Chen, L. Gu, L. Li, C. Yan, and Y. Fu, "Frequency dynamic convolution for dense image prediction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 30 178–30 188.

[10] B. Nian, F. Tang, J. Ding, J. Yang, Z. Zheng, S. K. Zhou, and W. Liu, "Srs: Siamese reconstruction-segmentation network based on dynamic-parameter convolution," *IEEE Transactions on Image Processing*, 2025.

[11] S. Chen, C. Ding, M. Liu, J. Cheng, and D. Tao, "Cpp-net: Context-aware polygon proposal network for nucleus segmentation," *IEEE Transactions on Image Processing*, vol. 32, pp. 980–994, 2023.

[12] H. Tang, Y. Chen, T. Wang, Y. Zhou, L. Zhao, Q. Gao, M. Du, T. Tan, X. Zhang, and T. Tong, "Htc-net: A hybrid cnn-transformer framework for medical image segmentation," *Biomedical Signal Processing and Control*, vol. 88, p. 105605, 2024.

[13] Z. Li, Y. Zheng, D. Shan, S. Yang, Q. Li, B. Wang, Y. Zhang, Q. Hong, and D. Shen, "Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 6, pp. 2254–2265, 2024.

[14] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[15] H. Li, D. Liu, Y. Zeng, S. Liu, T. Gan, N. Rao, J. Yang, and B. Zeng, "Single-image-based deep learning for segmentation of early esophageal cancer lesions," *IEEE Transactions on Image Processing*, vol. 33, pp. 2676–2688, 2024.

[16] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2441–2449.

[17] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[18] A. Gupta, S. Gehlot, S. Goswami, S. Motwani, R. Gupta, Á. G. Faura, D. Štepec, T. Martinčič, R. Azad, D. Merhof *et al.*, "Segpc-2021: A challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images," *Medical Image Analysis*, vol. 83, p. 102677, 2023.

[19] E. Calabrese, J. E. Villanueva-Meyer, J. D. Rudie, A. M. Rauschecker, U. Baid, S. Bakas, S. Cha, J. T. Mongan, and C. P. Hess, "The university of california san francisco preoperative diffuse glioma mri dataset," *Radiology: Artificial Intelligence*, vol. 4, no. 6, p. e220058, 2022.