SGDC: STRUCTURALLY-GUIDED DYNAMIC CONVOLUTION FOR MEDICAL IMAGE SEGMENTATION

Bo Shi   BO.SHI@MAIL.CONCORDIA.CA

Wei-ping Zhu (ID)   WEIPING@ECE.CONCORDIA.CA

M.N.S Swamy (ID)   SWAMY@ENCS.CONCORDIA.CA

Department of Electrical and Computer Engineering, Concordia University

Abstract

Spatially variant dynamic convolution provides a principled way of integrating spatial adaptivity into deep neural networks. However, mainstream designs in medical segmentation commonly generate dynamic kernels through average pooling, which implicitly collapses high-frequency spatial details into a coarse, spatially-compressed representation, leading to over-smoothed predictions that degrade the fidelity of fine-grained clinical structures. To address this limitation, we propose a novel Structure-Guided Dynamic Convolution (SGDC) mechanism, which leverages an explicitly supervised structure-extraction branch to guide the generation of dynamic kernels and gating signals for structure-aware feature modulation. Specifically, the high-fidelity boundary information from this auxiliary branch is fused with semantic features to enable spatially-precise feature modulation. This design replaces context aggregation with pixel-wise structural guidance, effectively preventing the information loss in average pooling. Experimental results show that SGDC achieves state-of-the-art performance on ISIC 2016, PH2, ISIC 2018, and CoNIC datasets, delivering superior boundary fidelity by reducing the Hausdorff Distance (HD95) by 2.05, and providing consistent IoU gains of 0.99%-1.49% over pooling-based baselines. Moreover, the mechanism exhibits strong potential for extension to other fine-grained, structure-sensitive vision tasks, such as small-object detection, offering a principled solution for maintaining structural integrity in medical image analysis. To facilitate reproducibility and encourage further research, the implementation for both our SGE and SGDC methods is publicly available at https://github.com/solstice0621/SGDC.

Keywords: Medical Image Segmentation, Dynamic Convolution, Structure Preservation, Feature Refinement, Boundary Awareness

1. Introduction

A core challenge in medical image segmentation is the trade-off between expanding the network's receptive field (Chen et al., 2017) for semantic understanding and preserving high spatial resolution for precise boundary delineation. Operations designed for the former often compromise the latter, leading to fidelity loss of intricate structures critical for clinical diagnosis.

One prominent line of research introduces explicit edge supervision (Lin et al., 2023, 2025), typically employing an auxiliary branch to learn boundary information for guiding the main segmentation network. While widely adopted, the utility of this guidance is

fundamentally constrained by the fusion mechanisms employed. Strategies that compress boundary features into a single-channel attention map impose an inherent information bottleneck, preventing the transmission of rich structural cues. Conversely, methods relying on simple concatenation or addition risk having these high-frequency details smoothed out by subsequent convolutional operations. Therefore, preserving structural fidelity requires a deeper and more interactive fusion mechanism capable of effectively modulating semantic features with structural priors, rather than simple aggregation.

A more advanced approach is spatially variant dynamic convolution (Chen et al., 2020; Lou and Yu, 2025), which employs a kernel prediction network to generate unique convolutional parameters for each spatial location. This enables highly adaptive, location-aware feature extraction and enhances local structural modeling. Existing methods in this paradigm focus on semantic awareness, generating guidance either from the local feature neighborhood itself (Li et al., 2021) or from features at different network stages (Lou and Yu, 2025). Despite their sophistication, these methods share a critical flaw: reliance on adaptive average pooling to generate dynamic kernels. This operation discards pixel-level positional information essential for boundary handling, causing the generated kernels to depend on coarse spatially ambiguous signal, which ultimately leads to over-smoothing of fine details.

Beyond the pooling operation, a more fundamental limitation lies in the intrinsic nature of the conditioning signal itself. In existing paradigms, the guidance is implicitly derived from the network's own backbone features (Wang et al., 2025). These features are inherently trained to maximize intra-class consistency for semantic recognition, resulting in spatially smooth representations dominated by low-frequency information. Using such semantically homogeneous signals to generate dynamic kernels creates a paradox: the kernels are tasked with refining boundaries, yet they are conditioned on features that have already suppressed structural details. Consequently, the generated kernels remain insensitive to fine geometric variations, such as thin boundaries or irregular contours, regardless of the spatial resolution.

These limitations indicate that semantic features alone are insufficient for driving spatially adaptive kernels, especially when boundary precision is required. Instead of relying on implicit conditioning signals (Li et al., 2021) derived from pooled or fused semantics, the dynamic mechanism benefits from a separate source that explicitly encodes structural cues. Incorporating a supervised boundary-guidance stream provides high-frequency information that the semantic pathway cannot supply, allowing the kernel generation process to respond to local geometric variations that are essential for accurate medical segmentation.

We propose a novel structure-guided dynamic convolution (SGDC) network (SGD-Net), a framework designed to directly address this structural information loss through two core innovations. First, we introduce an average-pooling-free SGDC mechanism, which leverages explicit, high-fidelity structural guidance, rather than pooled semantic information, to generate spatially-aware dynamic kernels and gating signals, thereby enabling adaptive feature refinement sensitive to boundary details. Second, we propose a structure guidance extractor (SGE), an independently supervised auxiliary branch that produces multi-channel guidance maps from multi-scale features, delivering the high-frequency structural cues needed by our SGDC modules.

2. Related Works

2.1. Dynamic Convolution

Dynamic convolution has evolved significantly as a powerful mechanism to enhance the representational capacity of neural networks by adapting model parameters to the input. The early stage of this evolution, characterized by sample-variant methods such as Cond-Conv (Yang et al., 2019) and DynamicConv (Chen et al., 2020), introduced the concept of generating a unique set of kernels for each input image. However, these methods primarily rely on Global Average Pooling (GAP) to aggregate a global context vector for kernel generation. While effective for image classification, this design is inherently spatially invariant; it applies the same dynamic kernel across the entire image, treating background and fine-grained foreground structures identically, which is suboptimal for dense prediction tasks like segmentation that require spatial precision.

To address the localization bottlenecks of sample-variant methods, the paradigm has shifted towards spatially variant dynamic convolutions, which can be categorized into two streams based on their guidance source. The first stream, represented by methods, such as Involution (Li et al., 2021), Condconv(Yang et al., 2019) and (Wang et al., 2025) utilizes implicit self-guidance, where kernel generation is conditioned solely on the local neighborhood of the target pixel, or even derived strictly from the individual pixel instance to avoid spatial aggregation. The second stream employs implicit heterogeneous-source guidance, implemented via the Contmix dynamic convolution module (Lou and Yu, 2025) in the OverLock network. Contmix constructs conditioning signals by aggregating feature combinations across different hierarchical levels of the network. However, despite its improved spatial adaptivity, this design shares a critical limitation when applied to medical segmentation: the prevalent reliance on adaptive average pooling for context compression, which implicitly collapses high-frequency spatial details into a coarse representation. The conditioning signals are derived from the network's semantic backbone, optimized for intra-class consistency. This inherent homogeneity suppresses the high-frequency variations necessary for precise boundary delineation. Consequently, driving dynamic kernels with these semantically pooled signals creates a fundamental paradox: the refinement mechanism is effectively blind to the fine structures it is intended to preserve.

2.2. Explicit Structural Guidance in Segmentation

Parallel to dynamic convolution, explicit edge supervision methods (Lin et al., 2023, 2025) adopt a multi-task framework to learn boundary information explicitly, effectively providing valuable high-frequency cues for structural fidelity. However, their effectiveness is fundamentally constrained by the inefficient fusion mechanism. Common strategies—such as compressing boundary features into a single-channel attention map—create a severe information bottleneck. Furthermore, simple element-wise addition or concatenation often fails to effectively modulate semantic features, leading to the washing out of sharp edge details by subsequent layers. Therefore, despite successfully extracting edges, these approaches lack a potent mechanism to fully leverage this structural information for effective feature refinement.

3. Methodology

3.1. Overall Architecture

Our proposed SGD-Net, illustrated in Fig. 1 (a), is built upon a hierarchical encoder–decoder architecture. To achieve robust feature extraction, we employ a Res2Net-50 backbone pretrained on ImageNet to generate multi-scale feature maps. The shallowest features are further processed by a Transformer encoder to capture long-range dependencies. The core innovations of our network involve the SGE and SGDC modules, shown in Fig. 1 (b) and (c), respectively, which are utilized to refine these multi-scale features before passing them to a lightweight decoder.
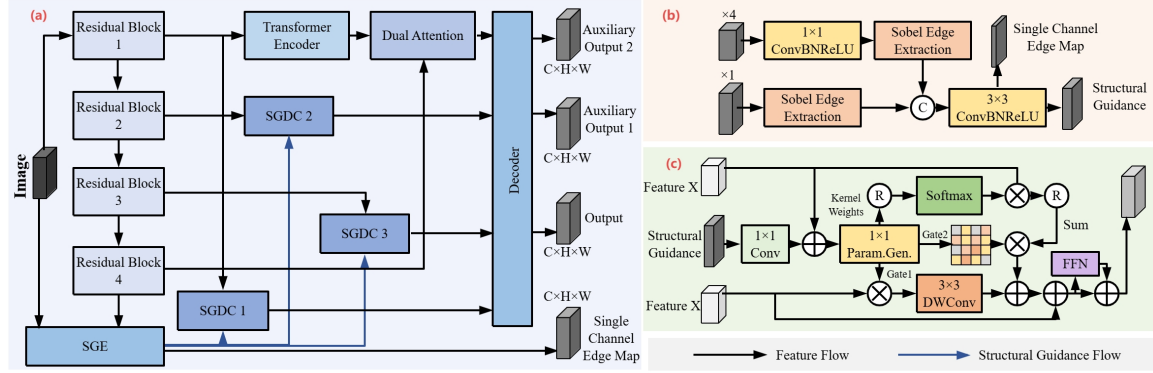

Figure 1: (a) The illustration of our proposed SGD-Net. (b) Structures of SGE Module. (c) structures of SGDC Module

To ensure a comprehensive and effective training, our network is optimized via a multi-output supervision strategy. The decoder produces segmentation predictions at three distinct scales. Outputs from coarser scales, referred to as auxiliary outputs, provide intermediate supervision signals. This deep supervision technique ensures that feature representations across all network depths are effectively optimized.

Furthermore, our SGE generates a dedicated edge map. This output is directly supervised by ground-truth boundary maps, compelling the SGE to explicitly learn high-fidelity structural guidance. This ensures that the guidance provided to our core module SGDC is both accurate and rich in essential boundary details.

3.2. Structure Guidance Extractor

The SGE module, detailed in Fig. 1 (b), is designed to generate the high-fidelity structural guidance essential for our SGDC mechanism. Its core innovation lies in the decoupling of two distinct tasks: generating a single channel edge map for explicit supervision and a multi-channel structural guidance feature for the SGDC modules. This design is crucial because the ideal representations for these tasks differ; supervision requires sharp, precise contours, whereas guidance benefits from richer, multi-dimensional information. The SGE workflow begins by processing features from two distinct network stages: the deep, semantic-rich

features from the final encoder block and the shallow, high-resolution features from the first encoder. To align their channel dimensions and ensure computational efficiency, the deep features are first projected into a lower-dimensional space via a 1×1 convolution. Both feature streams are then processed to extract structural details. In contrast to methods using learnable convolutions for this task, we employ a classic, non-trainable Sobel operator whose fixed kernels explicitly inject a stable and efficient structural guidance into the network, obviating the need to learn basic gradient extraction from scratch. We consciously employ a deterministic Sobel operator over learnable convolutions for this stage. While learnable layers offer flexibility, they are prone to overfitting to specific semantic textures in deep networks, potentially diluting the purity of boundary cues. The fixed Sobel operator acts as a 'structural anchor', ensuring that the extracted guidance remains strictly faithful to geometric gradients regardless of the varying semantic contexts. This injects a stable, domain-independent structural prior into the dynamic mechanism. This process is enhanced through an "Edge Modulation" operation, formulated as:

$$F_{mod} = F_{in} \odot \sigma \left(\sqrt{(F_{in} * K_x)^2 + (F_{in} * K_y)^2} \right) \tag{1}$$

where F_{in} is the input feature, $*$ denotes convolution with the Sobel kernels (K_x, K_y), $\odot$ the element-wise multiplication, and σ is the Sigmoid function. This operation selectively amplifies feature activations corresponding to edges. During the fusion stage, the processed deep and shallow features are concatenated. This combined feature map is then passed through a 3×3 convolution to promote effective interaction and integration. Finally, this refined feature is fed into two independent convolutional heads to generate the decoupled outputs. The single channel edge map is directly supervised by ground-truth boundary maps to ensure localization accuracy, while the Structural Guidance is passed to the SGDC modules, providing the essential high-frequency information that enables effective boundary-aware feature refinement.

3.3. Structure-Guided Dynamic Convolution Module

Standard dynamic convolutions, despite their spatial adaptability, are fundamentally limited by their reliance on pooling operations that degrade structural details. To overcome this, we propose the SGDC, a novel feature refinement unit that replaces the conventional pooling-driven method with a guidance-driven one. As illustrated in Fig. 1 (c), the SGDC is a feature processing block centered around a lightweight parameterization head that generates adaptive weights for a parallel dual-branch execution body.

The SGDC module takes two inputs: the main feature map from the encoder, Feature X (F_X), and the multi-channel structural guidance $(F_{guidance})$ from the SGE. These features are fused via element-wise addition to form the input to the parameter generation head, a single 1x1 convolution. It dynamically generates three outputs for each spatial location: the dynamic kernel weights (W_{dyn}), and two gating signals, Gate1 (g_1) and Gate2 (g_2).

This dual-branch body processes the input feature F_X through two complementary parallel branches. Before branching, it is crucial to normalize F_X by a LayerNorm layer. The main dynamic branch applies an 'unfold' operation to decompose the normalized feature map into overlapping local patches. The generated kernel weights (W_{dyn}) are passed through

a 'Softmax' function to act as spatially-varying weighting coefficients. A weighted sum of the patches is then computed—a process equivalent to a large-kernel, spatially variant convolution—and the result is recomposed via a 'fold' operation and modulated by Gate2 (g_2). The second branch, a local refinement path, modulates the normalized feature with Gate1 (g_1) and processes it with a 3x3 depthwise convolution.

The outputs of the two branches are then merged by element-wise summation, and this branch-level fusion is added back to the original F_X through the primary residual connection. This dual-branch design is motivated by the principle of complementary feature processing. The dynamic branch, driven by the SGE, excels at content-adaptive modulation and long-range structural modeling. However, purely dynamic operators can sometimes introduce spatial instability or over-react in homogenous regions. The local refinement branch, employing a static depthwise convolution, provides a deterministic high-frequency pass. By aggregating these two streams, the SGDC ensures that the feature map benefits from adaptive structural shaping while retaining the fundamental textural integrity and training stability provided by standard convolution. The feature map passes through a second LayerNorm layer and into a Feed-Forward Network (FFN), which consists of two 1x1 convolutions with an intermediate GELU activation. A final residual connection integrates the FFN output. The entire operation can be concisely formulated as:

$$F'_X = \text{LayerNorm}(F_X) \tag{2}$$

$$F_{mix} = \text{dynamic}(F'_X, W_{dyn}, g_2) + \text{local}(F'_X, g_1) \tag{3}$$

$$F_{out} = \text{FFN}(\text{LayerNorm}(F_X + F_{mix})) + (F_X + F_{mix}) \tag{4}$$

3.4. Loss Function

To ensure comprehensive supervision for both semantic representation and structural fidelity, we optimize SGD-Net end-to-end using a composite multi-task objective function. This objective integrates deep supervision on the multi-scale segmentation predictions with explicit boundary supervision on the structural priors. The total loss $\mathcal{L}_{\text{total}}$ is formulated as:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{3} \mathcal{L}_{\text{seg}}(o_i, M_{\text{gt}}) + \lambda \cdot \mathcal{L}_{\text{edge}}(o_e, E_{\text{gt}}) \tag{5}$$

where the first term represents the segmentation loss. To facilitate robust gradient flow and accelerate convergence, we employ a deep supervision strategy, applying the loss to predictions $\{o_1, o_2, o_3\}$ from three different decoder scales against the ground truth mask M_{gt}. The term $\mathcal{L}_{\text{seg}}$ itself is a hybrid loss combining Binary Cross-Entropy (BCE) and Dice loss, designed to handle pixel-level classification accuracy while mitigating class imbalance issues inherent in medical segmentation.

The second term, $\mathcal{L}_{\text{edge}}$, provides direct supervision to the single-channel edge map o_e generated by the SGE module. We employ a Dice loss for this term to enforce high-fidelity alignment between the predicted structural priors and the ground truth boundaries E_{gt}. The

hyperparameter λ acts as a trade-off factor to balance the semantic and structural learning objectives. In our experiments, we empirically set $\lambda = 3$ to emphasize the importance of boundary preservation.

4. Experiments and Results

4.1. Implementation Details

We evaluate SGD-Net on several public datasets: ISIC 2016 (Gutman et al., 2016), ISIC 2018 (Codella et al., 2019), and PH2 (Mendonça et al., 2013) for skin lesion segmentation, and the CoNIC dataset (Graham et al., 2021b) for nuclei segmentation. For ISIC 2018, we report 5-fold cross-validation results. For ISIC 2016 and PH2, to evaluate cross-dataset generalization, the model is trained on ISIC 2016 and tested on PH2. The CoNIC dataset is split into training, validation, and test sets with a ratio of 7:1:2. We use the Dice Coefficient and Intersection over Union (IoU) as primary evaluation metrics. Model efficiency is assessed using FLOPs and the number of parameters.

The model is optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} and a polynomial learning rate decay. We use a batch size of 16 and the input image size is 256×256. The encoder is initialized with ImageNet-pre-trained Res2Net-50 (Gao et al., 2019) weights and fine-tuned for 50 epochs. To ensure a fair comparison, we adopt the experimental setup from (Lin et al., 2023) as our primary baseline.
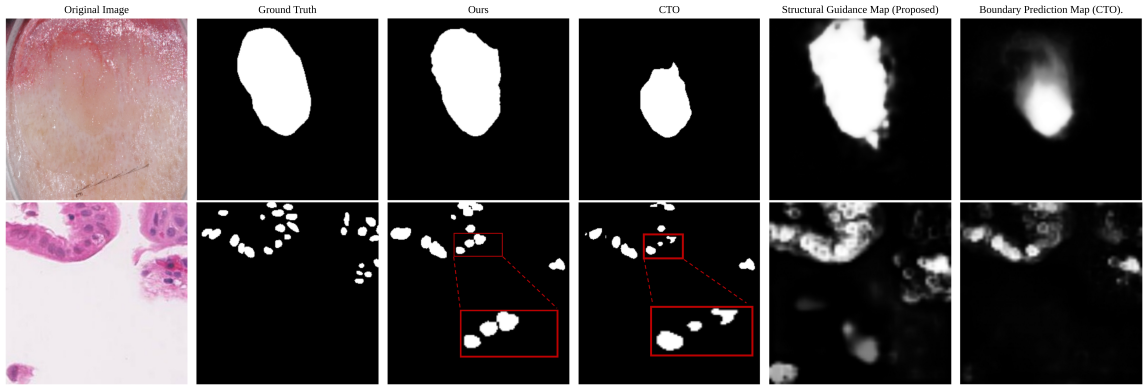

Figure 2: Comparison of qualitative results and structural guidance maps on ISIC 2018(Codella et al., 2019) and CoNIC(Graham et al., 2021b)

4.2. Experiment Results

Comparisons with State-of-the-Art Methods: As shown in Table 1 and Table 2, SGD-Net achieves state-of-the-art performance on skin lesion segmentation. On the ISIC 2016→PH2 test, it attains 92.93% Dice and 87.29% IoU, outperforming the best previous method by 0.37% and 0.69%, respectively. On ISIC 2018 with 5-fold cross-validation, our method reaches 91.41% Dice and 84.96% IoU, surpassing other methods by 0.78% and 0.99%.

Table 1: Dice and IoU comparisons with other methods on ISIC 2016(Gutman et al., 2016) and PH2 (Mendonça et al., 2013) dataset.

ISIC 2016 & PH2		
Methods	**Dice ↑**	**IoU ↑**
MSCA (Bi et al., 2016)	81.57	72.33
Bi *et al.* (Bi et al., 2017)	90.66	83.99
Lee *et al.* (Lee et al., 2020)	91.84	84.30
CTO(Lin et al., 2025)	92.56	86.60
SGD-Net (Proposed)	**92.93**	**87.29**

Table 2: Dice and IoU comparisons with other state-of-the-art methods on ISIC 2018 dataset (Codella et al., 2019), along with model complexity metrics.

Methods	Params (M)	GFLOPs	Dice ↑	IoU ↑
U-Net(Ronneberger et al., 2015)	7.78	14.59	82.92	77.56
LeViT(Graham et al., 2021a)	64.23	23.57	86.98	79.68
TransUNet (Chen et al., 2024)	114.65	37.02	89.97	83.32
CTO(Lin et al., 2025)	62.22	22.70	90.63	83.97
SGD-Net (Proposed)	**75.88**	**36.23**	**91.41**	**84.96**

On the CoNIC challenge (Table 3), in comparison with other methods, SGD-Net achieves 81.61% Dice, 69.46% IoU, and 68.79% PQ, ranking first across all three metrics. In terms of model efficiency, our network achieves these improvements with comparable FLOPs and number of parameters.

Figure 2 provides a qualitative comparison on the ISIC 2018 and CoNIC datasets, highlighting the superior performance and interpretability of our method. In the top row, where the skin lesion exhibits extremely low contrast against the surrounding tissue, the baseline (CTO) fails to discern the weak transition, resulting in significant under-segmentation. In contrast, SGD-Net accurately delineates the complete lesion structure. Similarly, for the dense nuclei in the bottom row, as highlighted in the red zoomed-in boxes, our method preserves fine-grained geometric topology, whereas the baseline produces blurred and fragmented masks. This performance gap can be explained using the intermediate maps: our Structural Guidance Map (Col. 5) generates sharp, high-contrast responses that precisely highlight structural gradients, whereas the baseline's boundary prediction (Col. 6) appears spatially diffuse and coarse. This visual evidence confirms that our SGE successfully extracts high-fidelity structural priors, which the SGDC module then leverages to prevent the over-smoothing characteristic of pooling-based paradigms.

Table 3: Performance comparison on CoNIC dataset (Graham et al., 2021b).

Methods	Dice ↑	IoU ↑	PQ ↑
U-Net (Ronneberger et al., 2015)	78.36	66.39	64.41
Att-UNet (Schlemper et al., 2019)	79.44	66.11	65.28
TransUNet (Chen et al., 2024)	77.03	64.35	64.02
CTO (Lin et al., 2025)	<u>80.67</u>	<u>67.97</u>	<u>67.17</u>
SGD-Net (Proposed)	**81.61**	**69.46**	**68.79**

Table 4: Ablation study of different SGD-Net variants on ISIC 2018 (Codella et al., 2019).

Method	Dice ↑	IoU ↑	HD95 ↓
(1) Baseline (Res2Net + ViT)	89.57	82.91	32.63
(2) Baseline+ SGE	90.39	83.43	18.58
(3) Baseline+ SGE+ Contmix	90.95	84.22	24.21
(4) SGD-Net (Proposed)	**91.41**	**84.96**	**16.09**

4.3. Ablation Study

We conduct a comprehensive ablation study on the ISIC 2018 dataset to validate both the incremental contributions of SGD-Net modules (Table 4) and the internal mechanisms of our SGDC module (Table 5). As shown in Table 4, the standard baseline (1) consisting of Res2Net + ViT exhibits a high HD95 of 32.63, indicating that semantic features alone lack sufficient spatial precision for accurate lesion delineation. In variant (2), we incorporate the Structural Guidance Encoder (SGE) into the baseline by **adding the SGE-generated guidance feature to the baseline feature map via element-wise addition before feeding it to the decoder**, which reduces the HD95 by 14.05 points to 18.58. This empirically confirms that explicit structural priors provide a crucial inductive bias that semantic features alone cannot capture.

To rigorously assess the effectiveness of the fusion strategy, we replace the simple addition used in variant (2) with Contmix (Lou and Yu, 2025), a recent dynamic convolution module that also leverages the SGE-generated guidance features. Despite its complexity, Contmix underperforms relative to the simple addition, yielding a higher HD95 of 24.21. This outcome illustrates the "pooling trap": its adaptive pooling aggregates spatial context, inadvertently smoothing out high-frequency structural details provided by the SGE. In contrast, our SGDC module avoids pooling and employs a dual-branch design, enabling more precise integration of guidance features and leading to superior segmentation performance.

To further isolate the source of these gains, we analyze the internal components of SGDC in Table 5. First, we implemented a "Self-Guidance" variant where kernels are generated from the input feature F_X itself, simulating standard implicit dynamic convolutions. This degradation leads to a sharp increase in HD95 ($16.09 \rightarrow 21.34$), confirming that semantic features alone are insufficient for precise boundary delineation, and the performance gain stems specifically from the explicit structural prior. Second, ablating the dual branches

Table 5: Internal mechanism ablation of the SGDC module on ISIC 2018. We analyze the impact of the guidance source and the dual-branch architecture. 'Self-Guidance' implies using the input feature itself for attention generation, simulating standard dynamic convolutions.

Method	Dice ↑	IoU ↑	HD95 ↓
SGD-Net (Ours)	**91.41**	**84.96**	**16.09**
w/ Self-Guidance (F_X)	90.52	83.65	21.34
w/o Local Branch	90.88	84.10	18.95
w/o Dynamic Branch	89.92	82.55	25.12

reveals a clear division of labor: removing the Local Branch causes a spike in HD95 (18.95), indicating that the static depthwise convolution acts as a deterministic "safety net" for high-frequency details, whereas removing the Dynamic Branch drops the Dice score (89.92), highlighting its role in capturing long-range semantic context. This confirms that SGDC succeeds through the complementary integration of stable local processing and adaptive dynamic modulation.

5. Conclusion

In this work, we identified and addressed a critical limitation in spatially variant dynamic convolution methods, where the reliance on adaptive average pooling for kernel generation leads to the degradation of fine-grained structural details. To overcome this bottleneck, our proposed SGD-Net incorporates the SGDC module, an average-pooling-free unit specifically designed to preserve high-frequency information during feature modulation.

Rather than compressing spatial contexts through aggregation, SGDC generates spatially precise kernels by leveraging structural priors provided by the auxiliary SGE branch, allowing dynamic modulation to adapt to local structure without sacrificing detail. Our experiments demonstrate that this design effectively mitigates the over-smoothing problem inherent in pooling-based approaches, yielding notable improvements in boundary fidelity (HD95).

Further internal analysis shows that the performance gain arises not merely from the additional structural cues, but from the synergistic interaction between SGDC's deterministic local refinement and its adaptive dynamic modulation. We believe this strategy holds significant potential for broader vision tasks requiring high fidelity in fine-grained structures, such as small object detection.

References

Lei Bi, Jinman Kim, Euijoon Ahn, Dagan Feng, and Michael Fulham. Automated skin lesion segmentation via image-wise supervised learning and multi-scale superpixel based cellular automata. In *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*, pages 1059–1062. IEEE, 2016.

Lei Bi, Jinman Kim, Euijoon Ahn, Ashnil Kumar, Michael Fulham, and Dagan Feng. Dermoscopic image segmentation via multistage fully convolutional networks. *IEEE transactions on biomedical engineering*, 64(9):2065–2074, 2017.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11030–11039, 2020.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael 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.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021a.

Simon Graham, Mostafa Jahanifar, Quoc Dang Vu, Giorgos Hadjigeorghiou, Thomas Leech, David Snead, Shan E Ahmed Raza, Fayyaz Minhas, and Nasir Rajpoot. Conic: Colon nuclei identification and counting challenge 2022. *arXiv preprint arXiv:2111.14485*, 2021b.

David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

Hong Joo Lee, Jung Uk Kim, Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Structure boundary preserving segmentation for medical image with ambiguous boundary. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4817–4826, 2020.

Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12321–12330, 2021.

Yi Lin, Dong Zhang, Xiao Fang, Yufan Chen, Kwang-Ting Cheng, and Hao Chen. Rethinking boundary detection in deep learning models for medical image segmentation. In *International conference on information processing in medical imaging*, pages 730–742. Springer, 2023.

Yi Lin, Dong Zhang, Xiao Fang, Yufan Chen, Kwang-Ting Cheng, and Hao Chen. Rethinking boundary detection in deep learning-based medical image segmentation. *Medical Image Analysis*, page 103615, 2025.

Meng Lou and Yizhou Yu. Overlock: An overview-first-look-closely-next convnet with context-mixing dynamic kernels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 128–138, 2025.

Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Lsnet: See large, focus small. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9718–9729, 2025.

Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in neural information processing systems*, 32, 2019.