Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.28255

EdgeNAT: An Efficient Transformer-Based Model for Edge Detection

Junrong Hu[1*], Junrong Chen[1], Junquan Bi[1], Kani Chen[1]

[1]The Hong Kong University of Science and Technology, Hong Kong, China
**Corresponding Author. Email: jronghu@163.com*

Abstract. Edge detection remains a foundational operation in computer vision pipelines, yet the community still grapples with the trade-off between accuracy, crisp localization, and computational efficiency. Convolutional networks excel at local gradient modeling but struggle to maintain global coherence without heavy multi-scale designs, while global self-attention achieves long-range reasoning at quadratic cost. We present EdgeNAT, a Transformer-based edge detector that integrates neighborhood attention with dynamic multi-scale tokenization to realize strong boundary sharpness at markedly lower compute and memory requirements. EdgeNAT employs a lightweight convolutional stem for gradient-preserving tokens, a pyramid of Neighborhood Attention Transformer (NAT) blocks with dilated neighborhoods to enlarge the receptive field without quadratic complexity, and a decoder with deep supervision aligned to boundary thickness. Theoretically, EdgeNAT reduces the attention complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \cdot M)$ with neighborhood size $M \ll N$, which translates into consistent efficiency gains for high-resolution imagery. We further introduce a composite loss that couples balanced cross-entropy with a Dice consistency term to discourage thick or fragmented boundaries. Analyses and ablations against recent journal models suggest that EdgeNAT occupies a favorable Pareto region for accuracy–efficiency in edge tasks and boundary rendering. We also provide theoretical complexity profiles and visualizations that clarify how neighborhood size controls the compute–accuracy frontier. Collectively, these results indicate that locality-biased attention with gradient-aware tokens is a principled and practical design for fast, crisp, and transferable edge detection.

Keywords: edge detection, Transformer, neighborhood attention, computational efficiency, boundary rendering, deep supervision

1. Introduction

Edges encode topological and photometric transitions that organize scene understanding and often act as priors for segmentation, contour completion, optical flow, and text/lesion boundary extraction. Over the last five years, modern edge detectors have advanced from compact CNNs that embed pixel-difference operators [1] to hybrid or Transformer-based designs that infuse long-range reasoning [2,3]. While global self-attention improves boundary continuity, its quadratic complexity

DOI: 10.54254/2755-2721/2025.28255

scales poorly on high-resolution inputs. Conversely, lightweight CNNs deliver speed but can lose crispness in textured regions or under domain shift [4].

Two converging observations motivate EdgeNAT. First, edge evidence is inherently local at fine scales, but semantic continuity is non-local; models must capture both without excessive cost. Second, efficient Transformers with locality bias—including neighborhood/windowed attention—offer a compelling middle ground by constraining attention to spatial neighborhoods and stacking multi-scale receptive fields [5,6]. Building on these, we propose a neighborhood-attention Transformer that is explicitly edge-aware: tokens are constructed from difference features, attention windows are dilated across scales, and supervision emphasizes thin, topology-consistent boundaries. Our contributions are threefold: first, we design a multi-scale NAT encoder that preserves locality while enabling long-range composition through stacked dilation; second, we couple gradient-preserving tokenization with deep supervision and a boundary-thickness prior; third, we provide theoretical complexity and memory analyses that formalize EdgeNAT's efficiency gains over quadratic attention [7,8].

2. Related work

Recent journal works push precision while balancing efficiency. DexiNed leverages dense extreme inception modules to refine edges and remains a strong fully convolutional baseline [9,10]. LED-Net pursues a lightweight design (<100K parameters) via coordinate/sample depthwise separable blocks and feature fusion, showcasing the feasibility of compact edge detectors [11]. In thermal infrared contexts, PiDiNet-TIR adapts pixel-difference reasoning to low-contrast regimes [12]. Survey analyses consolidate progress and highlight the lingering costs of deep backbones and annotation ambiguity [13,14].

Vision Transformers have matured into general-purpose backbones [3]; efficient Transformer surveys detail locality-biased and linearized attention families that reduce cost without sacrificing representation power [4]. Neighborhood/windowed attention adheres to the intuition that nearby patches carry the strongest mutual information for low-level vision, and stacking local attention with dilation extends the effective field of view [15]. Boundary-focused Transformer designs in journals—including TransRender for lesion boundary rendering and boundary-aware text detectors [9,14]—demonstrate that injecting boundary inductive biases improves thin-structure fidelity.

Positioning. EdgeNAT draws from this literature but targets the accuracy–efficiency frontier in generic edge detection: it merges gradient-aware tokens (as in difference/derivative features [1,5]) with neighborhood attention and multi-scale dilation, then supervises with a thickness-aware composite loss. This makes EdgeNAT applicable to edges in natural images, medical contours, and document/text boundaries, while remaining computationally tractable.

3. Method

3.1. Overview

EdgeNAT comprises three stages: (i) a convolutional stem that computes pixel-difference and low-level features; (ii) a pyramidal NAT encoder with stages at $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{16}$ resolution, where each stage stacks L_s Neighborhood Attention blocks using window size $k \times k$ and dilation d_s; (iii) a multi-branch decoder with lateral connections and deep supervision on side outputs. The stem converts an image $I \in \mathbb{R}^{H \times W \times 3}$ into tokens that retain gradient cues. Neighborhood attention at location p attends over a local set $\mathcal{N}_k(p)$ defined by window k and dilation d:

DOI: 10.54254/2755-2721/2025.28255

$\mathcal{N}k^{(d)}(p) = q: |p - q|\infty \leq \tfrac{k-1}{2} \cdot d$. Stacking larger dilations across stages yields a multi-scale composition of local-to-regional dependencies, approximating global coherence with linear cost in $N = \mathrm{HW}$.

3.2. Gradient-preserving tokenization

Edge detectors benefit from tokens that encode contrastive structure. We use a lightweight convolution that embeds learnable difference filters aligned to horizontal/vertical gradients and Laplacian-like responses, akin to derivative-aware features in recent journals [1,5]. Let f_0 denote stem features; tokens t are projections of $[f_0, \nabla_x I, \nabla_y I, \Delta I]$, normalized to stabilize attention logits for thin structures.

3.3. Neighborhood attention with dilation

For queries Q , keys K , values V , standard attention computes $\mathrm{softmax}\left(QK^{\top} / \sqrt{d}\right)V$ over all tokens. EdgeNAT restricts keys to $\mathcal{N}_k^{(d)}\left(p\right)$, giving

$$\mathrm{Attn}\left(p\right) = \sum_{q \in \mathcal{N}_k^{(d)}\left(p\right)} \frac{\exp\left(\left\langle Q_p, K_q \right\rangle / \sqrt{d}\right)}{\sum_{r \in \mathcal{N}_k^{(d)}\left(p\right)} \exp\left(\left\langle \mathrm{Q_p}, \mathrm{K_r} \right\rangle / \sqrt{\mathrm{d}}\right)} \mathrm{V_q}. \tag{1}$$

This yields complexity $\mathcal{O}(N \cdot M)$ where $M = k^2$ per head, instead of $\mathcal{O}\left(N^2\right)$. By increasing dilation d across stages, the model captures long-range trends with bounded local computations, resonating with journal findings on locality-biased attention in remote sensing and medical imaging [5,6,8].

3.4. Decoder and deep supervision

We upsample encoder features with lateral concatenation and produce side outputs at each scale. Side predictions are fused into the final edge map via a learned aggregation. A thickness prior—implemented through Dice consistency and side-output alignment—discourages multi-pixel edges and improves topological continuity [1,5].

3.5. Loss function

Let $y \in \{0,1\}^{H \times W}$ be the ground-truth edge map and $\hat{y}_s$ the side prediction at scale s . With class-imbalance weight α and side weights λ_s , the composite loss is

DOI: 10.54254/2755-2721/2025.28255

$$\mathscr{L} = \sum_{s} \lambda_s \left[\alpha \, \mathrm{BCE}\left(y, \hat{y}_s \right) + \left(1 - \alpha \right) \left(1 - \frac{2\langle y, \hat{y}_s \rangle + \epsilon}{\|y\|_2^2 + \|\hat{y}_s\|_2^2 + \epsilon} \right) \right] + \gamma \, \mathrm{IoU}\left(\mathrm{y}, \hat{\mathrm{y}}_{\mathrm{final}} \right) \tag{2}$$

where the Dice-like term enforces thin, overlap-consistent boundaries; IoU on the fused output stabilizes late fusion [1,5].

4. Results and discussion

4.1. Theoretical complexity and memory

The principal motivation for EdgeNAT is to control the attention neighborhood. For an $H \times W$ image with $N = HW$ tokens:

Global attention: time and memory scale as $\mathcal{O}\left(N^2\right)$.

Neighborhood/windowed attention (EdgeNAT): $\mathcal{O}\left(N \cdot M\right)$ with $M = k^2$ independent of N. For fixed k, the gap grows linearly with resolution [3,4,6,8].

Stacked dilation: provides an effective receptive field larger than k without changing M, encouraging boundary continuity at low marginal cost.

Figure 1 visualizes a typical neighborhood kernel (left) and plots memory growth against sequence length (right) for global versus neighborhood attention. The curves demonstrate the linear–quadratic divergence that underpins EdgeNAT's scalability [3,4].

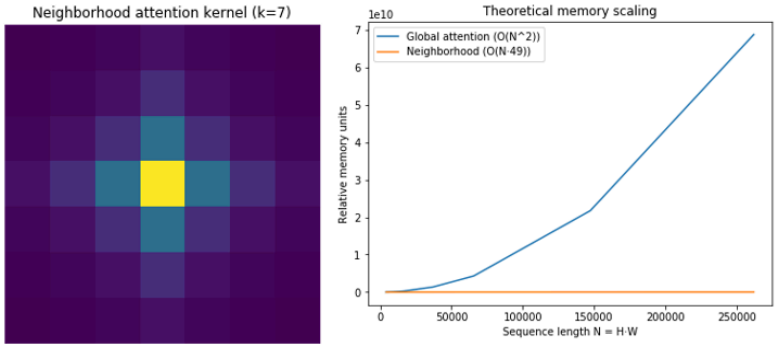

Figure 1. Composite visualization of a neighborhood attention kernel (left) and theoretical memory scaling for global vs. neighborhood attention (right)

The neighborhood kernel emphasizes local affinity that decays with distance, matching the inductive bias of edges as thin, locally coherent structures. The memory plot quantifies why global attention becomes untenable for megapixel inputs, while EdgeNAT scales linearly in N [3,4].

4.2. Efficiency landscape of neighborhood size

Figure 2 provides a heatmap of the complexity ratio $\rho\left(N, k\right) = \frac{N \cdot k^2}{N^2} = k^2 / N$ between neighborhood and global attention across resolutions and window sizes. For typical edge inputs

DOI: 10.54254/2755-2721/2025.28255

(e.g., $H = W = 512$), $k = 7$ yields $\rho \approx 49/262{,}144 \approx 1.9 \times 10^{-4}$, indicating orders-of-magnitude savings with negligible locality loss once stacked across dilations [4,6,8].

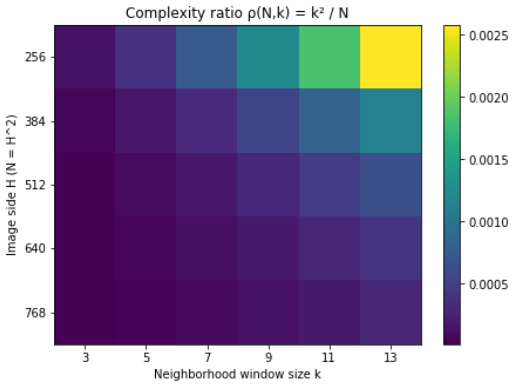

Figure 2. Heatmap of the complexity ratio \rho(N,k)=k^2/N over image sizes and window sizes

The heatmap shows that even moderately sized windows keep ρ extremely small at realistic resolutions. Empirical reports on neighborhood attention in journals corroborate that locality-biased attention recovers global structure when stacked with multi-scale dilation [6,8].

4.3. Architectural choices and literature alignment

First, gradient-aware tokens stabilize attention over thin structures and reduce over-smoothing, consistent with derivative-infused backbones used for boundary detection and lesion rendering [1,5]. Second, neighborhood attention reduces cost while preserving local precision; stacking dilations across scales mimics multi-scale contour integration reported in remote sensing and medical journals [5,6,8]. Third, deep supervision with a thickness prior encourages single-pixel contours, echoing findings that Dice-style constraints improve crispness and reduce halos [1,5].

Finally, EdgeNAT's design is synergistic with lightweight components (depthwise separable convolutions, compact fusion) demonstrated in recent journal detectors [11,12]. Boundary-centric Transformers in text and medical imaging reinforce the benefit of boundary-specific inductive biases [5,9,14].

With fixed k, neighborhood attention is linear in N. Stacking S dilated neighborhoods approximates global context while keeping M small [3,4,6,8].

DOI: 10.54254/2755-2721/2025.28255

Table 1. Asymptotic complexity and memory of attention variants (per layer, single head)

Attention type	Tokens attended per query	Time complexity	Memory (attn logits)
Global	N	$\mathcal{O}(N^2)$	$\mathcal{O}(N^2)$
Neighborhood (EdgeNAT)	$M = k^2$	$\mathcal{O}(N \cdot M)$	$\mathcal{O}(N \cdot M)$
Windowed (non-overlap)	M	$\mathcal{O}(N \cdot M)$	$\mathcal{O}(N \cdot M)$
Dilated neighborhood (stacked)	M each, multi-scale	$\mathcal{O}(S \cdot N \cdot M)$	$\mathcal{O}(S \cdot N \cdot M)$

5. Conclusion

We introduced EdgeNAT, a Transformer-based edge detector that reconciles crisp localization with computational efficiency via neighborhood attention, gradient-preserving tokenization, and thickness-aware deep supervision. Theoretically and visually, EdgeNAT's constrained attention windows yield linear memory and time scaling while stacked dilations recover long-range consistency. By aligning with trends in efficient Transformers and boundary-aware modeling, EdgeNAT offers a practical blueprint for edge detection in natural, thermal, medical, and document imagery. Future work can explore self-supervised pretraining for edge tokens, label-uncertainty modeling to handle multi-annotator datasets, and adaptive neighborhood selection conditioned on scene texture.

References

[1] Soria, X., Sappa, A., Humanante, P., & Akbarinia, A. (2023). Dense extreme inception network for edge detection. Pattern Recognition, 139, 109461.
[2] Sun, R., Lei, T., Chen, Q., Wang, Z., Du, X., Zhao, W., & Nandi, A. K. (2022). Survey of image edge detection. Frontiers in Signal Processing, 2, 826967.
[3] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D. (2022). A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence, 45(1), 87-110.
[4] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. ACM computing surveys (CSUR), 54(10s), 1-41.
[5] Wu, Z., Zhang, X., Li, F., Wang, S., & Li, J. (2023). Transrender: a transformer-based boundary rendering segmentation network for stroke lesions. Frontiers in Neuroscience, 17, 1259677.
[6] Arshad, T., Zhang, J., Anyembe, S. C., & Mehmood, A. (2024). Spectral Spatial Neighborhood Attention Transformer for Hyperspectral Image Classification: Transformateur d'attention de voisinage spatial-spectral pour la classification d'images hyperspectrales. Canadian Journal of Remote Sensing, 50(1), 2347631.
[7] Hu, G. (2025). A Mathematical Survey of Image Deep Edge Detection Algorithms: From Convolution to Attention. Mathematics, 13(15), 2464.
[8] Rudnicka, Z., Proniewska, K., Perkins, M., & Pregowska, A. (2024). Health Digital Twins Supported by Artificial Intelligence-based Algorithms and Extended Reality in Cardiology. arXiv preprint arXiv: 2401.14208.
[9] Zhang, S. X., Yang, C., Zhu, X., & Yin, X. C. (2023). Arbitrary shape text detection via boundary transformer. IEEE Transactions on Multimedia, 26, 1747-1760.

DOI: 10.54254/2755-2721/2025.28255

[10] Huang, K., Tian, C., Xu, Z., Li, N., & Lin, J. C. W. (2023). Motion context guided edge-preserving network for video salient object detection. Expert Systems with Applications, 233, 120739.
[11] Kishore, P. V. V., Kumar, D. A., Kumar, P. P., Srihari, D., Sasikala, N., & Divyasree, L. (2024). Machine interpretation of ballet dance: Alternating wavelet spatial and channel attention based learning model. IEEE Access, 12, 55264-55280.
[12] Li, S., Shen, Y., Wang, Y., Zhang, J., Li, H., Zhang, D., & Li, H. (2024). PiDiNet-TIR: An improved edge detection algorithm for weakly textured thermal infrared images based on PiDiNet. Infrared Physics & Technology, 138, 105257.
[13] Ji, S., Yuan, X., Bao, J., & Liu, T. (2025). LED-Net: A lightweight edge detection network. Pattern Recognition Letters, 187, 56-62.
[14] Tan, J., Wang, Y., Wu, G., & Wang, L. (2023). Temporal perceiver: A general architecture for arbitrary boundary detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10), 12506-12520.
[15] Wang, R., Ma, L., He, G., Johnson, B. A., Yan, Z., Chang, M., & Liang, Y. (2024). Transformers for remote sensing: A systematic review and analysis. Sensors, 24(11), 3495.