RESEARCH-ARTICLE

ConvPSNet - A convolution-based single step approach for pruned skeleton from noisy data

BINCY ANTONY, Indian Institute of Technology Madras, Chennai, TN, India

ANSHUMAN MISHRA, Indian Institute of Technology Madras, Chennai, TN, India

ANANTHAKRISHNAN A, Indian Institute of Technology Madras, Chennai, TN, India

DHARANIVENDHAN V, Indian Institute of Technology Madras, Chennai, TN, India

RAMANATHAN MUTHUGANAPATHY, Indian Institute of Technology Madras, Chennai, TN, India

Published: 15 December 2025

SA Technical Communications '25: SIGGRAPH Asia 2025 Technical Communications
December 15 - 18, 2025
Hong Kong, Hong Kong

Conference Sponsors:
SIGGRAPH

https://doi.org/10.1145/3757376.3771414
ISBN: 9798400721366

ConvPSNet - A convolution-based single step approach for pruned skeleton from noisy data

Bincy Antony
Indian Institute of Technology Madras
Chennai, India
bincyndl@gmail.com

Anshuman Mishra
Indian Institute of Technology Madras
Chennai, India
anshumanmishra21345@gmail.com

Ananthakrishnan A
Indian Institute of Technology Madras
Chennai, India
ananthu2014@gmail.com

Dharanivendhan V
Indian Institute of Technology Madras
Chennai, India
dharanivendhanv01@gmail.com

Ramanathan Muthuganapathy
Indian Institute of Technology Madras
Chennai, India
mraman@iitm.ac.in

Abstract

We present a deep learning framework for directly extracting a pruned skeleton from a noisy 2D shape. A key component of our work is a carefully constructed synthetic dataset of diverse shapes with human-in-the-loop validation of pruned skeletons, addressing the lack of suitable training data for this task. Inspired by recent insights into the design differences between convolutional and transformer architectures, we propose a convolution-based minimalist architecture built on a few large kernels emphasizing simplicity and elegance in design, particularly for resource and data constrained environments.

We introduce a novel loss function that penalizes spurious branches without removing essential structures. Our convolution-only model achieves higher F1 scores than existing small-kernel convolution and transformer-based approaches. This work highlights the importance of carefully crafted datasets and minimalist architectures for downstream applications like pruning skeletons.

CCS Concepts

• **Computing methodologies** → **Neural networks**.

ACM Reference Format:
Bincy Antony, Anshuman Mishra, Ananthakrishnan A, Dharanivendhan V, and Ramanathan Muthuganapathy. 2025. ConvPSNet - A convolution-based single step approach for pruned skeleton from noisy data. In *SIGGRAPH Asia 2025 Technical Communications (SA Technical Communications '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3757376.3771414

Authors' Contact Information: Bincy Antony, Indian Institute of Technology Madras, Chennai, India, bincyndl@gmail.com; Anshuman Mishra, Indian Institute of Technology Madras, Chennai, India, anshumanmishra21345@gmail.com; Ananthakrishnan A, Indian Institute of Technology Madras, Chennai, India, ananthu2014@gmail.com; Dharanivendhan V, Indian Institute of Technology Madras, Chennai, India, dharanivendhanv01@gmail.com; Ramanathan Muthuganapathy, Indian Institute of Technology Madras, Chennai, India, mraman@iitm.ac.in.

SA Technical Communications '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2136-6/25/12
https://doi.org/10.1145/3757376.3771414

1 Introduction

A medial axis (referred to as skeleton) of a shape is the locus of centres of maximally inscribed discs [Blum 1967]. It serves as a compact shape representation that preserves the essential geometry and topology of the underlying shape. A major limitation of skeletonization is its sensitivity to small perturbation on the boundary or noise, which leads to the appearance of many spurious branches. Figure 1(a) shows a perturbed boundary whose skeleton has many spurious branches (Figure 1(b)). To remove spurious branches, algorithmic approaches [Rong and Ju 2023] typically require manual tuning, which is an arduous task across different shapes and varying noise levels to achieve the desired output. Figure 1(c) shows the color-coded significance measure [Rong and Ju 2023], and the skeleton is then pruned using a threshold of 50° (Figure 1(d)), determined through empirical tuning. Generating a perceptually satisfactory skeleton automatically, without any manual intervention, is more desirable. For instance, automatic pruning in [Yang et al. 2020] achieves this but requires substantial computational time. Hence, in this paper, a learning-based approach is proposed for obtaining a pruned skeleton (Figure 1(e)) directly from a perturbed boundary (Figure 1(a)).

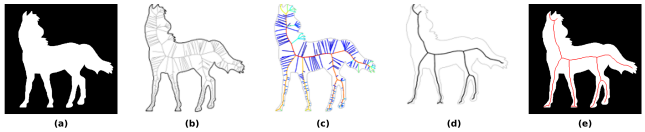

Figure 1: From a noisy shape to pruned skeleton: (a) a noisy shape. (b) medial axis exhibiting many spurious branches. (c), (d) color-coded branches after computing significance measure [Rong and Ju 2023] and pruned skeleton using 50° threshold (e) ours: directly obtained the pruned skeleton for the noisy input.

Recent advances in deep learning have enabled data-driven approaches for skeletonization [Panichev and Voloshyna 2019; Xu et al. 2019], mapping each pixel in the input shape to a binary skeleton representation. Their performance is heavily dependent on the availability of high-quality datasets. To the best of our knowledge, there are no existing datasets for producing pruned skeletons and hence such approaches are restricted to generating unpruned ones.

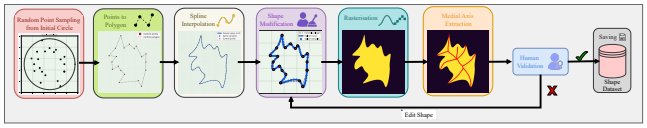

Figure 2: Shape-Skeleton generation pipeline.

Recent advancements such as ConvNeXt [Liu et al. 2022] and VanillaNet [Chen et al. 2023] highlight how modern design choices such as scaling kernel sizes, architectural minimalism, and effective receptive field (ERF) optimization, enable convolutions to achieve competitive or even superior performance compared to transformers in several downstream tasks. These insights motivated us to explore a convolution based strategy for skeleton pruning, that leverages the inherent strengths of convolutional architectures, such as inductive biases, data efficiency, and hierarchical feature reasoning. Our objective is to design models capable of generating perceptually clean and structurally meaningful skeletons even under boundary perturbations, using a newly developed dataset.

Our contributions are threefold. (1) We construct a synthetic dataset of 2,928 diverse shapes with pruned skeletons, through a human-in-the-loop process. (2) We design a convolution only architecture that employs large kernels within a shallow, minimalist framework, well suited for resource limited settings. (3) We propose a novel loss that suppresses spurious branches by enforcing consistency between the predicted and ground truth neighborhoods, thereby promoting better topological connectivity.

2 Methodology

2.1 Dataset Creation and curation

We synthesize a medial–axis dataset with a human-in-the-loop workflow. Base shapes (**1,000**) are generated as in Algorithm 1: sample points on a unit circle, connect to a polygon, fit a periodic spline to obtain a smooth closed contour, rasterize to an image, and compute its medial axis with `scikit-image` (Figure 2). If a shape or its medial axis exhibits any artifacts, we edit the control points and regenerate shape and skeleton. This data generation process enables creation of diverse shapes under human control. To obtain a pruned dataset (Figure 3), we perturb the boundary only within a narrow boundary band (Algorithm 2), generating up to five variants per base by adding a noise field N_T with amplitude A and maximum spatial frequency (in case of sinsoidal noise). Noise amplitude A scales displacements. For each variant, parameters are drawn from preset ranges (Table 1 in supplementary); for sinusoid we cap spatial detail by $f_{\max}$ (minimum wavelength $1/f_{\max}$) to keep perturbations smooth. This process is done with a human in the loop who has knowledge about medial axis and geometry to ensure that modifications do not alter the topology. We then remove near-identical shapes and unstable cases, yielding compact, diverse training **2,928** pairs.

2.2 Architecture and Model Design

Design Insight [Ding et al. 2022]:

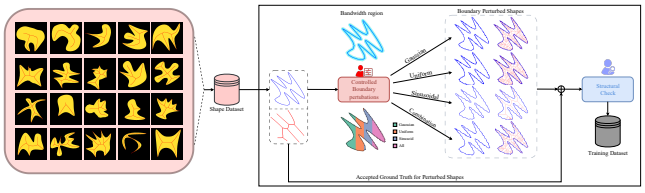

Figure 3: Dataset for pruned skeleton created by adding controlled boundary variations to base shapes.

ALGORITHM 1: Human-in-the-Loop Generation of Shape–Skeleton Pairs

Input: Image size $S \times S$; control points N; spline samples M
Output: Binary mask Y; skeleton S

1. **Seed points.** Sample N jittered points on a unit circle.
2. **Polygon.** Connect points to form a closed polygon.
3. **Spline.** Fit a periodic cubic spline; sample M points.
4. **Rasterize.** Scale to $S \times S$ and rasterize to Y.
5. **Skeleton.** S $\leftarrow$ MedialAxis(Y).
6. **Human editing.** Display the mask and skeleton; *edit control points to generate shape variants*; repeat as needed.
7. **Save.** Store Y and S.

ALGORITHM 2: Boundary-Noise Augmentation with Reference Consistency Check

Input: Ref. mask M_{ref}, skeleton S_{ref}, noise $T \in \{\text{Gaussian, Uniform, Sine, Combination}\}$, band w
Output: Accepted variants $\hat{M}$ paired with S_{ref}

1. **Band.** Compute boundary $B = \partial M_{\text{ref}}$; form a w-wide band R.
2. **Noise.** Choose amplitude A and max freq. $f_{\max}$; sample field N_T on B.
3. **Perturb.** Displace $p \in B$ along its normal by $N_T(p)$, clipped to $[-w, w]$; rasterize to $\hat{M}$.
4. **Consistency.** Visual check: accept iff change is small and salient features are preserved; else retune A, $f_{\max}$ or discard.
5. **Save.** Store $(\hat{M}, S_{\text{ref}})$.

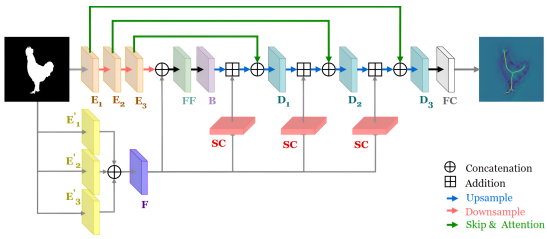

Figure 4: ConvPSNet Model architecture overview. E_1–E_3 and E'_1–E'_3 denote parallel branches with kernel sizes up to 7×7. (SC indicates structural conditioning.)

- **ERF scaling & implication:** By ERF theory, ERF $\in O(K\sqrt{L})$; it grows *linearly* with kernel size K and only *sub-linearly* with depth L.

- **Design choice & optimization:** Prefer *few large kernels* over deep stacks of 3×3; shallow large-kernel models expand ERF efficiently and are easier to train.
- **Shape bias:** Large kernels models are more similar to human *shape* bias.

Overall Design: Our network, **ConvPSNet**, employs larger kernels (up to 7×7) in parallel branches inspired by [Liu et al. 2022].

Multi-Branch Feature Extraction: The input is processed by three parallel branches, each operating with a different kernel size: a 3×3 branch for fine local details, a 5×5 branch for mid-range context, and a 7×7 branch for long-range structural continuity. The outputs of these branches are bilinearly aligned, concatenated, and compressed with a 1×1 convolution to form a fused structural prior F.

Structural Conditioning: During decoding, the fused prior F is injected at each stage (denoted as **SC** in Figure. 4) through a 1×1 projection followed by addition to the upsampled features. SC acts as a global bias *reinforcing topological connectivity.*

Final Prediction: The output of the decoder is passed through a final 1×1 convolution and a sigmoid activation to produce the skeleton likelihood map:

2.3 Loss Formulation

The total loss consists of : a standard binary cross-entropy loss, focal loss to handle class imbalance and the novel spurious endpoint penalty.

***Spurious branch loss (neighborhood MAE).*.** Let $\hat{Y} \in [0,1]^{H \times W}$ be the predicted skeleton map and $Y \in \{0,1\}^{H \times W}$ the ground truth. For each pixel u, consider the $k \times k$ window $\mathcal{N}_k(u)$ centered at u (we use k=3) and define the local skeleton occupancy

$$m_{\hat{Y}}(u) = \frac{1}{k^2 - 1} \sum_{v \in \mathcal{N}_k(u) \setminus \{u\}} \hat{Y}(v), \quad m_Y(u) = \frac{1}{k^2 - 1} \sum_{v \in \mathcal{N}_k(u) \setminus \{u\}} Y(v).$$

The loss is the mean absolute error between these occupancies over pixels with a full window ($\mathcal{V}$ ignores a $\lfloor k/2 \rfloor$-pixel border):

$$\mathcal{L}_{\text{spur}} = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \left| m_{\hat{Y}}(u) - m_Y(u) \right|. \tag{1}$$

This encourages the predicted local *neighborhood* to match the ground truth neighborhood:

The final training loss for our model is a weighted sum of all components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{BCE}} \cdot \mathcal{L}_{\text{BCE}} + \lambda_{\text{Focal}} \cdot \mathcal{L}_{\text{Focal}} + \lambda_{\text{spur}} \cdot \mathcal{L}_{\text{spur}} \tag{2}$$

(Ablation on loss in supplementary)

3 Experimental setup

All experiments were conducted on a machine equipped with NVIDIA GeForce RTX 4070 Ti GPU using PyTorch 2.1.0 with CUDA 12.1.

Table 1: Network performance with F1 score.

Model	Params (M)	Size (MB)	Time (ms)	F1
ConvPSNet (ours)	11.38	43.41	7.73	0.798
U-Net	31.04	118.46	4.84	0.775
ViT	22.26	84.93	7.20	0.765
ConvNeXt	34.10	130.10	7.61	0.744
ResNet-50	68.20	260.39	8.69	0.710

3.1 Training

Our skeleton prediction model was trained on a dataset of 2,928 samples, split into training, validation, and test sets with a 70–20–10 ratio. All inputs were resized to a resolution of 224×224 pixels. We used the AdamW optimizer with an initial learning rate of 0.001, scheduled via cosine annealing. Training was performed for 50 epochs with a batch size of 8.

3.2 Evaluation protocol

We evaluate the accuracy of skeleton prediction using the F1-measure between the prediction and the ground truth , computed with 1 pixel localization tolerance. We also performed qualitative analysis.

4 Results and discussion

Qualitative comparison: Our test set contains complex real-world shapes with noise, thin parts, high curvature, and weak structural cues. At inference, the network outputs a likelihood map $p(x)$; we extract skeletons by binarization of the skeleton map. (Figure 5). We compare against ViT, ConvNeXt, U-Net, ResNet-50, and the `scikit-image` medial axis. All models are trained on the same dataset until convergence. **ConvPSNet** consistently produces centered, connected centerlines and stable junctions with few spurious branches. **ViT** offers strong global context but *overfits under limited data*, leading to unstable generalization. **ConvNeXt** uses large kernels yet often yields *slightly overshot/offset* centerlines, likely due to limited data. **U-Net** localizes well on our small dataset but frequently shows *breaks* in thin or high-curvature regions. **ResNet-50** exhibits similar breakages. The **medial axis** baseline is brittle to boundary noise and produces extra branches.

The noise resistance of our model is compared with the variational pruning approach [Rong and Ju 2023] as shown in Figure 6. Their approach is compute intensive and also requires computation of medial axis geometrically and the object angle.

Quantitative comparison. As shown in Table 1, ConvPSNet attains the best F1 while remaining lightweight (11 M params, 43 MB). Its inference time is slightly slower than U-Net but comparable to ViT/ConvNeXt and faster than ResNet-50.

5 Conclusion

We have presented a deep learning architecture for directly producing pruned skeletons from noisy 2D shapes. For training, we carefully created a dataset of noisy shape–skeleton pairs. We demonstrated a carefully constructed architecture with few large shallow kernels along with a multifeature extraction module performed better than small size kernel deep architectures (both qualitatively and

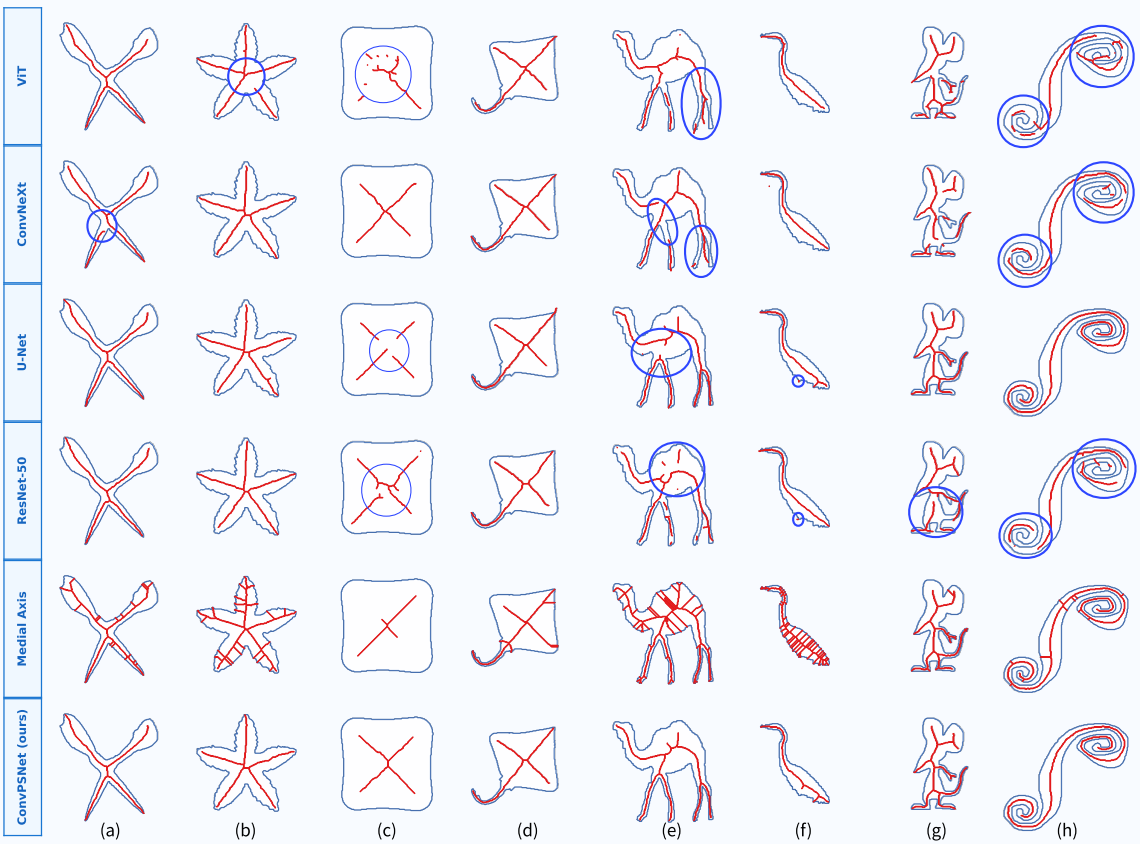

Figure 5: Skeletons generated by different networks and medial axis algorithm on noisy shapes with (a) sharp corners, (b) complex junctions, (c), (d) shapes lacking enough structural information, (e)-(h) high curvatures. All shapes are noisy and hence the medial axis algorithm creates spurious branches. The circled regions highlight areas where prediction is wrong or missing.

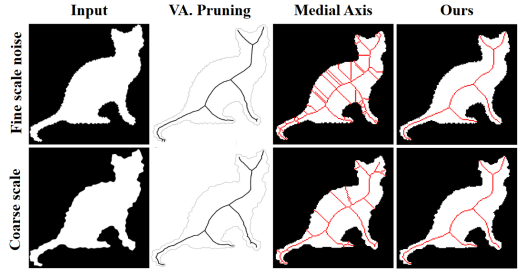

Figure 6: Robustness of pruning in presence of fine noise(row 1) and coarse noise (row 2) [as defined by [Rong and Ju 2023]].

quantitatively), particularly in data and resource constrained settings. The introduced neighborhood-consistency (spurious-branch) loss promoted connectivity and suppressed extraneous branches.

Limitations and future work: Our method may underperform on geometrically degenerate cases, such as junctions where multiple medial branches converge. While the model is highly parameter-efficient, its inference time trails U-Net since current libraries are optimized for 3×3 stacks. Reducing sequential stages like structure conditioning we used is a promising direction to close this latency gap.

References

Harry Blum. 1967. A transformation for extracting new descriptors of shape. *Models for the Perception of Speech and Visual Form* (1967), 362–380.

Hanting Chen, Yunhe Wang, Jianyuan Guo, and Dacheng Tao. 2023. VanillaNet: the power of minimalism in deep learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 308, 15 pages.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11963–11975.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11976–11986.

Oleg Panichev and Alona Voloshyna. 2019. U-net based convolutional neural network for skeleton extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

Peter Rong and Tao Ju. 2023. Variational Pruning of Medial Axes of Planar Shapes. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, e14902.

Mengyang Xu, Song Bai, Zhichao Zhang, Errui Yang, and Xiang Bai. 2019. DeepFlux for Skeletons in the Wild. In *CVPR*.

Cong Yang, Bipin Indurkhya, John See, and Marcin Grzegorzek. 2020. Towards automatic skeleton extraction with skeleton grafting. *IEEE transactions on visualization and computer graphics* 27, 12 (2020), 4520–4532.