SkeletonNetV2: A Dense Channel Attention Blocks for Skeleton Extraction

Sabari Nathan*
Couger Inc
Tokyo,Japan
sabari@couger.co.jp

Priya Kansal*
Couger Inc
Tokyo,Japan
priya@couger.co.jp

Abstract

Geometrical analysis of a shape through skeletonization has some of very important high- and low-level application which includes tracking, manipulation, retrieval, representation, registration, recognition, and compression. The task of skeletonization is defined as the generation of the medial axis of the shape while preserving its original topology and geometry. While the earlier approaches are mainly based on extracting the skeleton and then pruning the unwanted branches, the present study proposes a novel convolutional neural network based method to perform this task. The proposed architecture is an encoder-decoder network that leverages the benefits of the coordinated convolutional layer and multi-level supervision to prevent the loss of information between the extracted skeleton and the ground truth. The dense attention block is used as the backbone block in the encoder and decoder block. This architecture is performing better than the state of art on not only skeletonization of image tasks but also skeletonization from the point cloud. This method achieved an F1 score of 0.7961 on the Pixel Skeleton dataset and a Chamfer Distance (CD) score of 1.9561 on the Point skeleton dataset.

1. Introduction

Skeletonization is a process to reduce the object dimensions to extract its medial axis or skeleton while preserving its essential topological and geometrical information that can be used for the complete recovery of the original object [29]. Skeletonization has a wide range of applications for many decades which includes object description, object recognition, object matching, object retrieval, tracking [16], optical character recognition [25], fingerprint recognition [36], motion detection [20], object tracking [16], etc. Skeletons are also widely used in life sciences for plant morphology [1], [26] and medical image analysis. A large number of approaches are used to extract the skeleton. However, the impressive success of the convolutional neural network is now appealing to the researcher to leverage this approach in skeletonization as well. The robustness and generalization capacity of the CNNs makes these model more lucrative. Except this, most of the time, these approaches provide an end-to-end solution which makes the process of skeletonization easier when compared to traditional approaches which are based on applying some pruning of the branches after extracting the skeleton or comparing the pixel values with a threshold. The present method is a convolutional neural network that contains an encoder-decoder structure. The features in latent space are extracted using the dense attention blocks in the encoder and then the pixel-wise segmentation map is generated using the same dense attention blocks in the decoder. Before routing the input into the residual blocks, the input images are initially fed into the coordinate convolutional layer. The weights are optimized using an average of dice loss and cross-entropy loss. Along with this, multilevel immediate supervisions are used for efficient back-propagation and to make the model robust and generalized. This model also used two different attention modules predicting more precise values of pixels in the skeleton map. As an image prior, the distance-based medial axis is concatenated with the decoder's output for higher recall. The main contributions of the proposed approach are as follows:

1. An end-to-end automatic deep neural network is introduced for easy and effective skeletonization.
2. An image prior and multi-level supervision-based network to eliminate all prepossessing and post-processing steps.
3. An efficient Dense channel attention block is proposed for robust delineation of the medial axis.
4. Attention blocks are introduced before every supervision for higher precision of the predicted pixels which means that all the medial points are connected to form a single medial axis.

*These authors contributed equally to this work

5. The proposed model is generalized enough to perform on two different types of input *viz.* pixel image and point image.

Our model has the ability to learn rich hierarchical and contextual features. The state of art results on two different types of inputs shows the effectiveness, robustness, and generalization capacity of the proposed architecture. The rest of the paper discusses the literature, the details of the proposed architecture experiments, training details, results, and conclusion.

2. Literature Review

Skeleton extraction is a widely investigated area since the last decade. However, the most recent works are mainly focused on the extracting skeleton from the RGB images [31], [14], [10], [33]], [23], [22], which involves segmentation or detection of the objects and extract the skeleton at the same time. Also, an extensive research is done either on edge detection [9], [6], [35], [30] or segmentation [35], [13] individually. These kinds of works are not fully suitable for the present task. Some initial works are done on the extracting skeleton from the binary mask images [4], [3], [5], [12] which is similar to our task. However, most of these works are focused on skeleton pruning to remove the unwanted branches rather than skeleton extraction. In the work done by [11], the authors introduced the boundary noise to avoid the uninformative branch creations. [21] used skeleton strength maps (SSM) which are calculated by the isotropic diffusion of the Euclidean distance transformation of binary images and their gradient. After calculating the SSM, they connected all the local maxima points of SSM with the shortest possible line to extract the skeletons. [34] has extracted the dense skeleton map followed by grafting the backbone branches. [7][19] approached the task of skeleton extraction as an image generation model and used the generative adversarial network to extract the skeletons. However, the recent works, for example, [26], [17], [8], [27] have introduced convolutional neural networks to extract skeletons from the binary mask images. [2] and [17] used deep convolutional neural network for skeleton extraction from cloud images too. Inspired by the above works, the current model also leverages the benefits of CNN to achieve the state of art results. Similar to [26], we have also fused the side layers into the final output layer. However, to improve, the accuracy of the model, instead of taking the output of convolution layers as side layers, we have introduced CS-SE layers at the end of each up-sampling layer and have considered the output of CS-SE layers as side layers. The detail of our approach is discussed in section 3.

3. Details of Proposed Architecture

The detailed architecture is presented in Fig.1. In the following subsection, we will discuss the details of each component of the proposed architecture.

3.1. Coordinate convolutional layer

For improving the model's generalization capacity, extra two channels are created for input image using coordinate convolution layer as proposed in [24]. Coordinate convolutional layer helps the network to decide on the features related to translation equivariance which helps in improving the generalization capacity of the model.

3.2. Dense Channel attention block (DCAB)

Dense connectivity in the convolutional layers improves the information flow throughout the network [15]. The output of each layer is concatenated to all subsequent layers along the channel axis in the network. Hence the output of a layer i is represented by Eq.(1):

$$x_i = f([x_0 \| x_1 \| \ldots\ldots \| x_{i-1} \|]) \tag{1}$$

Since the number of channels at each layer of the block is growing at a rate of k, where k is the number of previous layers, to make the information flow more precise, we added channel attention to each convolutional layer in the dense block. Detail of the Channel attention block is presented in Fig.2(a). There is a total of 8 DCAB blocks, two for each resolution level are used in the encoder which is concatenated with the output of the decoder after up-sampling the last layer, the skip connections are used to concatenate. This concatenated output is then treated as the input for another DCAB block in the decoder followed by CSCA.

3.3. Attention Modules

To boost the performance of the proposed architecture, three attention modules are used.

Channel Attention Inspired by [32, 18], to create the channel attention, we first aggregated the spatial information by creating two pooled feature maps using average pooling and max pooling, thereafter two single-layer perceptrons are used to create the channel attention maps. The output feature maps are then merged, and a sigmoid activation is applied to get the final channel attention. The mathematical representation is given in Eq.(2). A typical channel attention block proposed in [32] is presented in 3(a).

$$C_A = xXf_a[w_1(w_0\frac{\sum_{i=1}^{n} x_i}{n}) + w_1(w_0(max(x_i)))] \tag{2}$$

This channel attention is used in Dense channel Attention block.

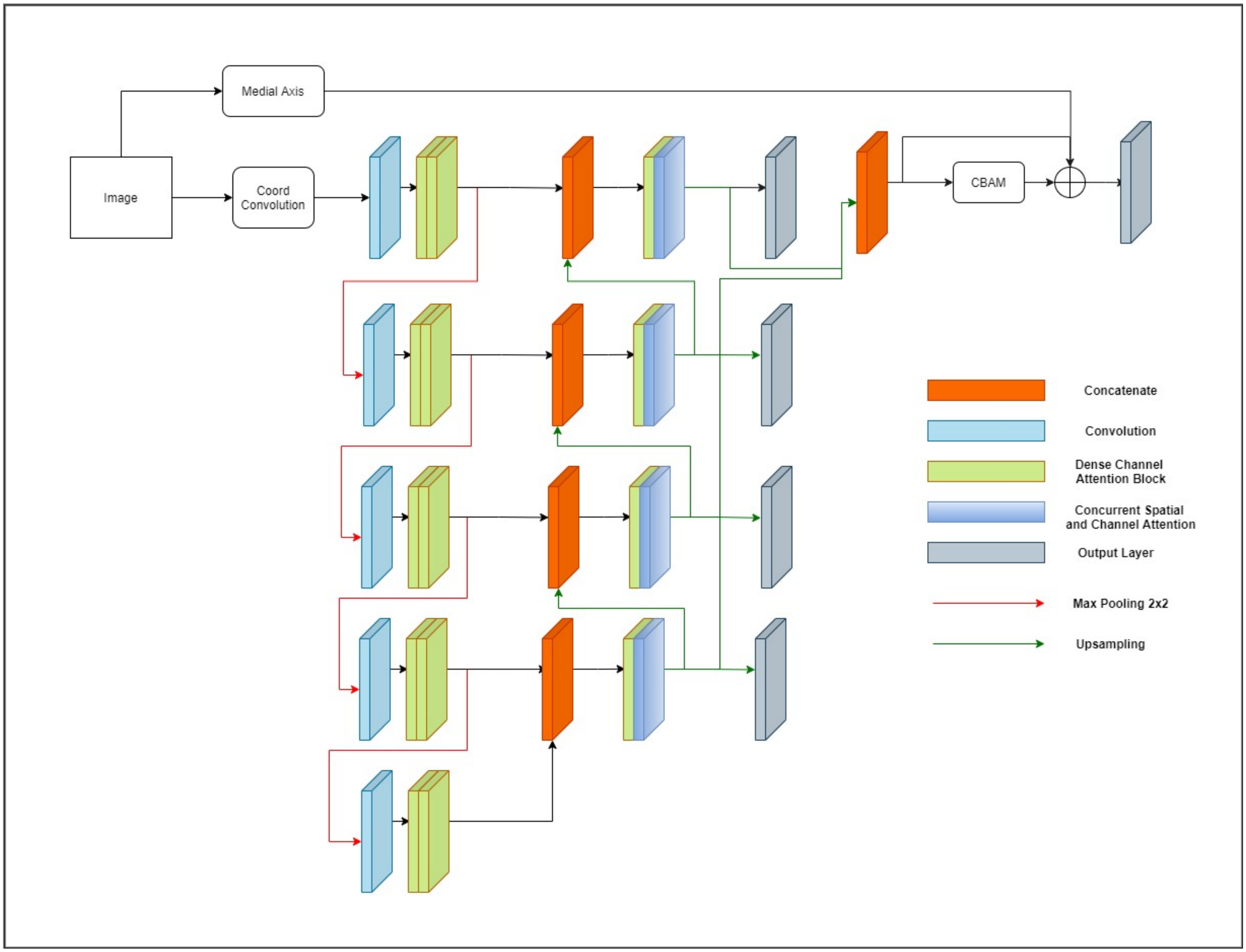


Figure 1. Proposed Model: SkeletonNetV2: A DenseNet Channel Attention Blocks for skeleton extraction

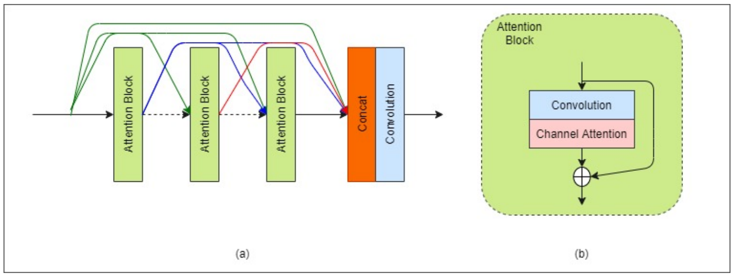


Figure 2. (a) Proposed dense Channel Attention block; (b) Attention Block Connection

Spatial Attention Similar to [32, 18], to apply, spatial attention, the pooling operations are done along the channel axis. Then these two max-pooled and average pooled operations are concatenated and then a convolutional operation is applied with 7x7 filter. Similar to channel attention, the spatial attention is then multiplied by the input feature map.

Eq.(3) shows the mathematical operation involve in spatial attention. A typical spatial attention block proposed in [32] is presented in 3(b).

$$S_A = xXf_{conv}[\frac{\sum_{i=1}^{n} x_i}{n} \| max(x_i))] \tag{3}$$

Convolution Block Attention block In the proposed work, both the channel and spatial attention are used in a sequential manner, the channel attention is created for the input features, and then it is added back to the input. The spatial attention is applied to the output then. Hence the combined attention is the spatial attention on the channel attention as shown in Eq.(4). Fig. 3(c) shows the CBAM block proposed in [32]

$$attention = S_A(C_A(x)) \tag{4}$$

where, x represents the input feature map, f_a and f_{conv} represent the sigmoid activation function and 7x7 convolutional operation respectively. w_0 and w_1 are the shared single layer perceptron.

Concurrent Spatial and Channel Attention Inspired by [26],[28] an additional attention is applied on the output of each dense channel attention block in decoder. This recalibration encourages the network to learn more meaningful feature maps, that are relevant both spatially and channel-wise. The key difference in both the attention is that, unlike CBAM, the attentions are concurrent that means both the channel and spatial attention are applied to the input features and then the two outputs are added.

3.4. Multi-level Supervision

To train the model, multi-level supervision has been used. Since the receptive field increases as the network get deeper, feature maps at different layers utilize the spatial information at different levels. Supervision at different levels helps these layers to learn quickly and efficiently as the gradient is populated in these layers also and thus helps in updating the weights more efficiently. The output of each level is concatenated along the channel axis followed by the CBAM attention block before the final supervision.

4. Experimental Setup

In this section, the details of the dataset, training setup, loss function, and metrics are provided.

4.1. Dataset

The model is trained on the Pixel SkeletonNet dataset [7] and Point SkeletonNet dataset [7]. The Pixel Skeleton dataset contains 1219 training and 242 validation images. The input images of this dataset contain the pixel-wise binary mask for 89 objects. The sample images of the pixel image are presented in Fig. 4. The ground truth images are the skeleton of the object. Similarly, the Point Skeleton dataset contains 1219 training images. The input of the dataset is the shape point cloud given in the basic point cloud export format .pts. Sample shape point clouds and their corresponding skeleton point clouds are shown in Figure 2. For the purpose of training, the data is augmented using +45 and -45 degree spatial rotation. All the images are normalized between 0 and 1.

4.2. Training Details

The model is trained using tensorflow/keras framework for 5 outputs. Adam optimizer is used to update the weights while training. The learning rate is initialized with 0.001 and reduced after 10 epochs to 10 percent if validation loss does not improve. The batch size is set to 4 to be accommodated in the available hardware resources. The total epochs are set to 500. However, training is stopped early when the network started overfitting. The dataset is trained using Nvidia 1080 GTX GPU.

4.3. Loss Function

Similar to [26], to optimize the model weights, a combined loss is used. This combined loss is the sum of the binary cross-entropy and Dice Loss as defined in Eq.(7) The network is trained to minimize this combined loss with sigmoid activation function. Dice Loss is defined in equation (2) and L is cross-entropy loss defined in Eq.(5)

$$DiceLoss = 1 - \frac{2\sum_{i=0}^{k} y_i * p_i + \epsilon}{\sum_{i=0}^{k} y_i + \sum_{i=0}^{k} p_i + \epsilon} \tag{5}$$

Binary cross entropy is represented in Eq.(6)

$$BCE = -\sum_{i=0}^{k}[y_i * \log p_i + (1 - y_i) * \log(1 - p_i)] \tag{6}$$

$$Loss = DiceLoss + BCE \tag{7}$$

where, *yi* and *pi* are the ground truth and the predicted skeleton images respectively. The coefficient is used to ensure the loss function stability by avoiding the zero value in the denominator of dice loss.

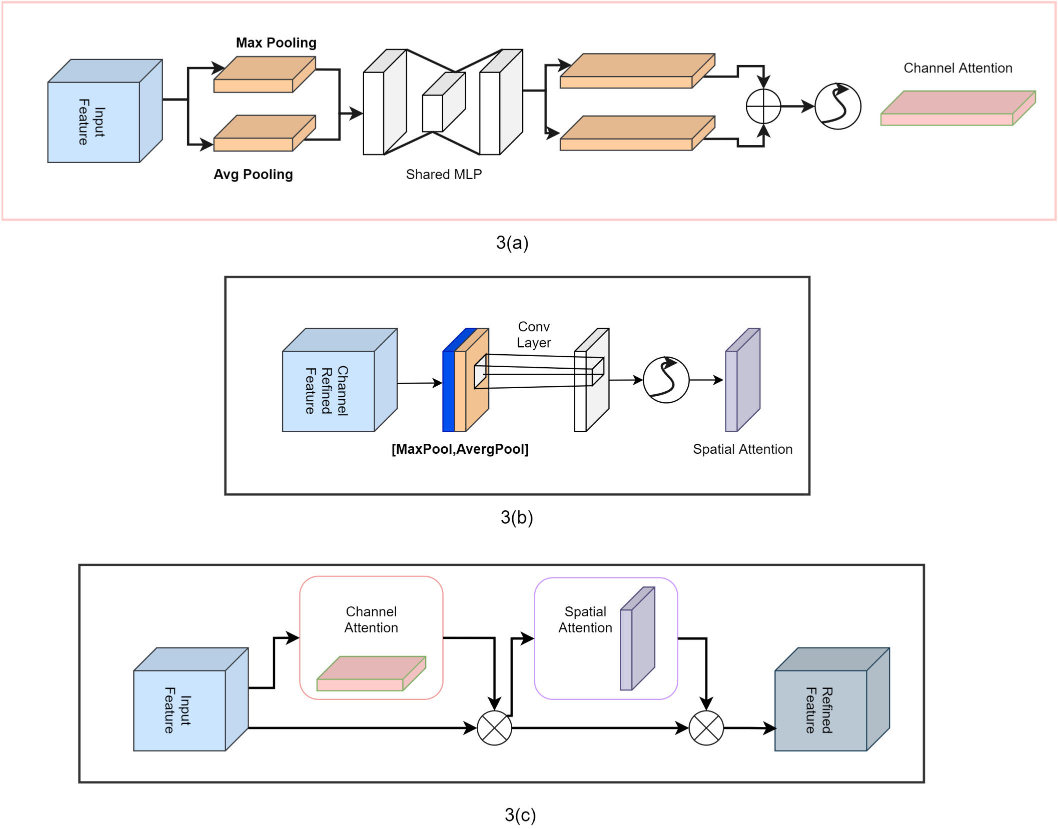
4.4. Evaluation Metrics

For pixel skeleton, the metrics used for evaluation is given as follows

$$F1score = 2 * \frac{precision * recall}{precision + recall} \tag{8}$$

whereas, precision and recalls are defined as

$$precision = \frac{TP}{TP + FP} \tag{9}$$



3(a)

3(b)

3(c)

Figure 3. (a) Channel Attention Module; (b) Spatial Attention Block; (c) Convolutional Block Attention Module

$$recall = \frac{TP}{TP + FN} \tag{10}$$

TP, FP and FN represents the true positive, false positive and false negative respectively.

For evaluating point skeleton, Chamfer Distance (CD) is used. Equ. (11) represents CD.

$$\begin{aligned} CD = \frac{1}{|S1|}\sum_{p_1 \epsilon S_1} min_{p_2 \epsilon S_2} \|p_1 - p_2\|_2 \\ + \frac{1}{|S_2|}\sum_{p_2 \epsilon S_2} min_{p_1 \epsilon S_1} \|p_1 - p_2\|_2 \end{aligned} \tag{11}$$

5. Results

The proposed model achieved the state of art results with a very low computational complexity of the model. The Pixel net model, which is trained with pixel-wise binary mask achieved an F1 score of 0.7961 on the validation dataset. The results of all side layers are calculated for the purpose of ablation. Table 1 shows the results of the ablation study.

Table 2 shows the results of the proposed model on the PixelNet dataset. The results clearly shows the significant improvement in the state-of-art approaches.

Output Layer	**F1 Score on Pixel Dataset**	**CD on shape Point Cloud dataset**
Output Layer 1	0.7098	2.0618
Output Layer 2	0.7211	2.0112
Output Layer 3	0.7466	2.0032
Output Layer 4	0.7681	1.9582
Fused Output	**0.7961**	**1.9561**

Table 1. Results of the output layers and the Fused layer on validation Dataset of Pixel and Point dataset.

Table 3 shows the results of the proposed model on the shape cloud point dataset. The results clearly shows the significant improvement in the state-of-art approaches.

Some images from the training data along with the predicted output and ground truth are presented in Fig 5.

6. Conclusion

The present architecture leverages the DCAB, attention module and customized loss for end to end extraction of the medial axis from the pixel-wise binary mask and shape

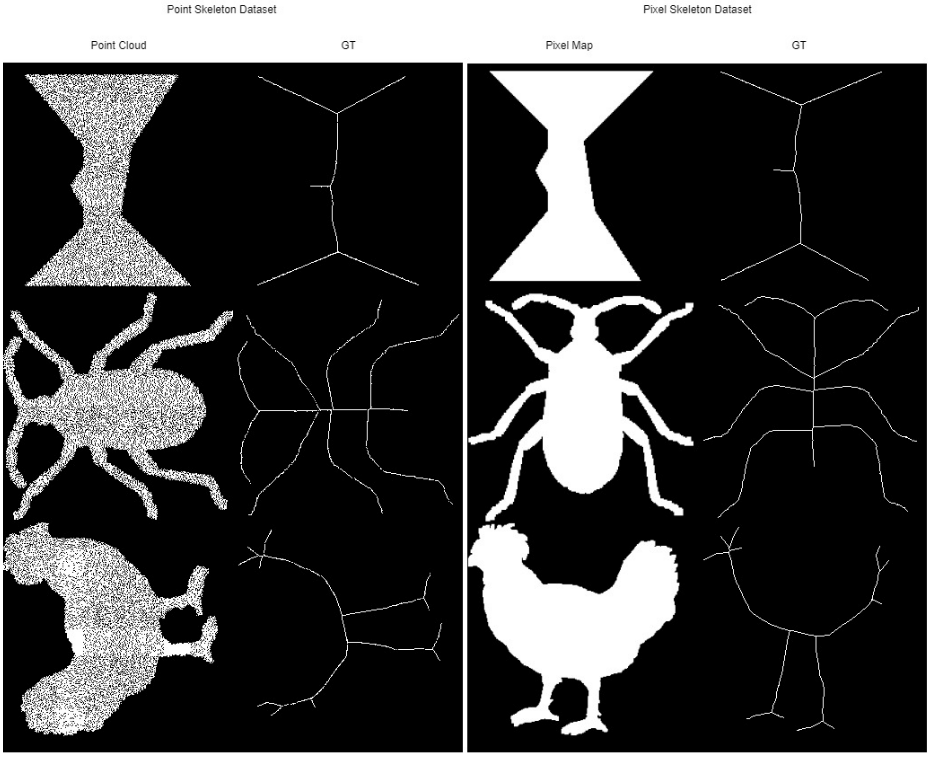


Figure 4. Sample Images for point Skeleton and Pixel Skeleton Dataset.

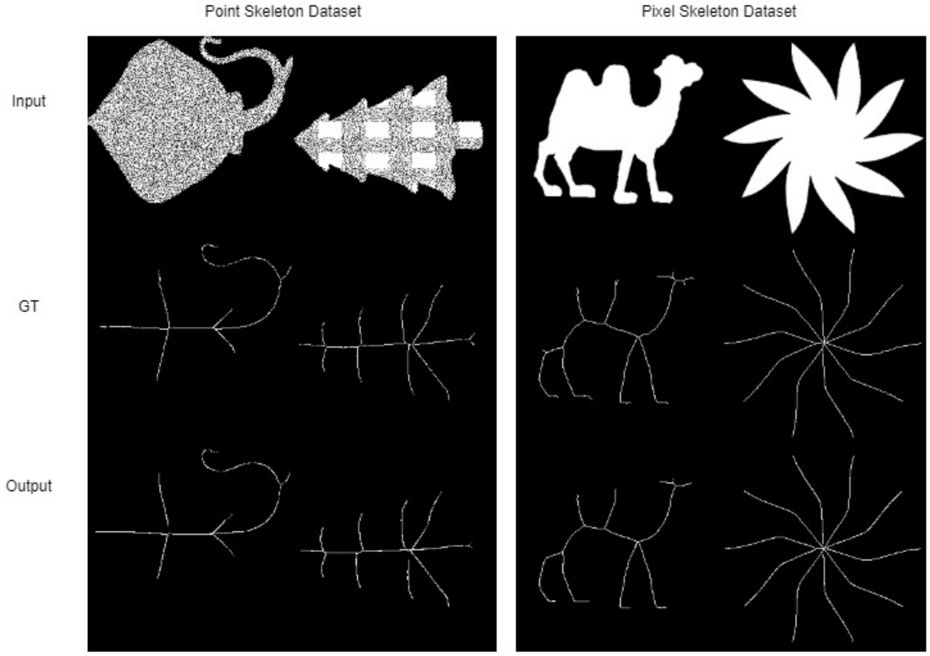


Figure 5. Outputs from Point and pixel datasets compared to ground truth.

cloud point. Overall the present paper is able to handle two task with higher evaluation metrics which seems to be a promising approach for future tasks of medial axis retrieval.

Method	**F1 Score**
Baseline [7]	0.6244
Jiang et. al.[17]	0.6325
Nathan and Kansal [26]	0.7480
Panichev et. al.[27]	0.7500
Dey [8]	0.7780
Ours	**0.7961**

Table 2. Comparison of Results on Pixel Skeleton validation data with existing state-of-art. Higher score represents the better result

Method	**CD Score**
Rowel [2]	2.9105
Jiang et. al. [17]	2.40
Ours	**1.9561**

Table 3. Comparison of Results on Point Skeleton validation data with existing state-of-art. Lower the score, the better the results

7. Acknowledgement

The present work is supported by Couger Inc. Tokyo.

References

[1] Bucksch A. A practical introduction to skeletons for the plant sciences. *Applications in plant sciences*, 2(8), 2014. 1

[2] Rowel Atienza. Pyramid u-network for skeleton extraction from shape points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 7

[3] Dominique Attali and Annick Montanvert. Computing and simplifying 2d and 3d continuous skeletons. *Computer vision and image understanding*, 67(3):261–273, 1997. 2

[4] Liu WY Bai X, Latecki LJ. Skeleton pruning by contour partitioning with discrete curve evolution. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):449–462, 1997. 2

[5] Liu WY Bai X, Latecki LJ. The -medial axis. *Graphical Models*, 67(4):304–331, 2005. 2

[6] Torresani L. Deepedge Bertasius G, Shi J. A multi-scale bifurcated deep network for top-down contour detection. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4380–4389, 2015. 2

[7] Ilke Demir, Camilla Hahn, Kathryn Leonard, Geraldine Morin, Dana Rahbani, Athina Panotopoulou, Amelie Fondevilla, Elena Balashova, Bastien Durix, and Adam Kortylewski. Skelneton 2019: Dataset and challenge on deep learning for geometric shape understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 4, 7

[8] Sohom Dey. Subpixel dense refinement network for skeletonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 258–259, 2020. 2, 7

[9] Zitnick CL Dollar P. Fast edge detection using structured forests. ieee transactions on pattern analysis and machine intelligence. *IEEE transactions on pattern analysis and machine intelligence*, pages 1558–70, 2015. 2

[10] Charles-Olivier Dufresne Camaro, Morteza Rezanejad, Stavros Tsogkas, Kaleem Siddiqi, and Sven Dickinson. Appearance shock grammar for fast medial axis extraction from real images. *arXiv e-prints*, pages arXiv–2004, 2020. 2

[11] Leonard K Mari JL Morin G Durix B, Chambon S. The propagated skeleton: A robust detail-preserving approach. *In International Conference on Discrete Geometry for Computer Imagery*, pages 343–354, 2005. 2

[12] Pauly M Wormser C Giesen J, Miklos B. The scale axis transform. in proceedings of the twenty-fifth annual symposium on computational geometry. *In Proceedings of the twenty-fifth annual symposium on Computational geometry*, pages 106–115, 2009. 2

[13] Girshick R Malik J Hariharan B, Arbeláez P. Hypercolumns for object segmentation and fine-grained localization. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015. 2

[14] Cheng MM Borji A Torr PH Hou Q, Liu J. Three birds one stone: a unified framework for salient object segmentation, edge detection and skeleton extraction. *arXiv preprint*, 2018. 2

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[16] Kim D Jalal A, Kamal S. Depth map-based human activity tracking and recognition using body joints features and self-organized map. *In Fifth International Conference on Computing, Communications and Networking Technologies*, pages 1–6, 2014. 1

[17] Nan Jiang, Yifei Zhang, Dezhao Luo, Chang Liu, Yu Zhou, and Zhenjun Han. Feature hourglass network for skeleton detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 7

[18] Priya Kansal and Sabari Nathan. Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation. 2019. 2, 3

[19] Debbie Honghee Ko, Ammar Ul Hassan, Saima Majeed, and Jaeyoung Choi. Skelgan: A font image skeletonization method. *Journal of Information Processing Systems*, 17(1):1–13, 2021. 2

[20] Dastidar JG Kundu M, Sengupta D. Tracking direction of human movement-an efficient implementation using skeleton. *arXiv preprint*, 2015. 1

[21] Bai X Liu WY Latecki LJ, Li QN. Skeletonization using ssm of the distance transform. *In Fifth International Conference on Computing, Communications and Networking Technologies*, pages 1–6, 2014. 2

[22] K Li, Y Tian, B Wang, Z Qi, and Q Wang. Bi-directional pyramid network for edge detection. electronics 2021, 10, 329, 2021. 2

[23] Xiaolong Liu, Pengyuan Lyu, Xiang Bai, and Ming-Ming Cheng. Fusing image and segmentation cues for skeleton

extraction in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1744–1748, 2017. 2

[24] Molino P Such FP Frank E Sergeev A Yosinski J Liu R, Lehman J. An intriguing failing of convolutional neural networks and the coordconv solution. *In Advances in Neural Information Processing Systems*, pages 9605–9616, 2018. 2

[25] Li N. An implementation of ocr system based on skeleton matching. 1993. 1

[26] Sabari Nathan and Priya Kansal. Skeletonnet: Shape pixel to skeleton pixel. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 4, 7

[27] Oleg Panichev and Alona Voloshyna. U-net based convolutional neural network for skeleton extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 7

[28] Wachinger C Roy AG, Navab N. Concurrent spatial and channel 'squeeze excitation' in fully convolutional networks. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 421–429, 2018. 4

[29] Punam K Saha, Gunilla Borgefors, and Gabriella Sanniti di Baja. Skeletonization and its applications–a review. In *Skeletonization*, pages 3–42. Elsevier, 2017. 1

[30] Jiang Y Wang Y Zhang Z Bai X Shen W, Zhao K. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3982–3991, 2015. 2

[31] Jiang Y Wang Y Zhang Z Bai X Shen W, Zhao K. Object skeleton extraction in natural images by fusing scale-associated deep side outputs. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 222–230, 2016. 2

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 2, 3, 4

[33] Yongchao Xu, Yukang Wang, Stavros Tsogkas, Jianqiang Wan, Xiang Bai, Sven Dickinson, and Kaleem Siddiqi. Deepflux for skeleton detection in the wild. *International Journal of Computer Vision*, 129(4):1323–1339, 2021. 2

[34] Cong Yang, Bipin Indurkhya, John See, and Marcin Grzegorzek. Towards automatic skeleton extraction with skeleton grafting. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 2

[35] Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming-Hsuan Yang. Object contour detection with a fully convolutional encoder-decoder network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 193–202, 2016. 2

[36] Feng Zhao and Xiaoou Tang. Preprocessing and postprocessing for skeleton-based fingerprint minutiae extraction. *Pattern Recognition*, 40(4):1270–1281, 2007. 1