Context Attention Network for Skeleton Extraction

Zixuan Huang[1], Yunfeng Wang[1], Zhiwen Chen[1], Xin Gao[1], Ruili Feng[2], Xiaobo Li[1]

[1]Alibaba Group [2]University of Science and Technology of China

{zixuan.huangzixuan, weishan.wyf, zhiwen.czw, zimu.gx}@alibaba-inc.com, ruilifengustc@gmail.com, xiaobo.lixb@alibaba-inc.com

Abstract

Skeleton extraction is a task focused on providing a simple representation of an object by extracting the skeleton from the given binary or RGB image. In recent years many attractive works in skeleton extraction have been made. But as far as we know, there is little research on how to utilize the context information in the binary shape of objects. In this paper, we propose an attention-based model called Context Attention Network (CANet), which integrates the context extraction module in a UNet architecture and can effectively improve the network's ability to extract the skeleton pixels. Meanwhile, we also use some novel techniques including distance transform, weight focal loss to achieve good results on the given dataset. Finally, without model ensemble and with only 80% of the training images, our method achieves 0.822 F1 score during the development phase and 0.8507 F1 score during the final phase of the Pixel SkelNetOn Competition, ranking 1st place on the leaderboard.

1. Introduction

Skeleton extraction, also known as skeletonization, is a task focused on providing a simple representation of an object by extracting the skeleton pixels from the given binary or RGB image [6]. Nowadays, skeleton extraction is widely used in many fields, including object recognition [18], pose estimation [17] and motion forecasting [14], etc. Traditional methods are usually divided into three categories: morphological thinning methods [19], geometric methods, and distance transform-based methods [10]. These classical methods usually provide low accuracy results and are sensitive to noise at the edge of the shape. In recent years, with the development of artificial intelligence, some deep learning-based skeleton extraction approaches have been proposed, which usually treat extracting skeleton pixels as a classification problem. These methods take the binary or RGB image as input and directly predict the skeleton pixels, which avoids complex post-processing. But how to improve the performance of deep learning-based methods is still a challenge.

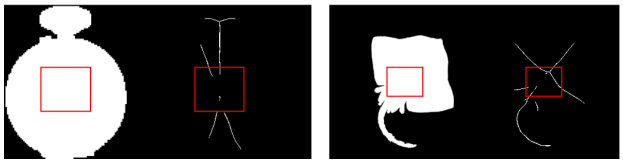

Figure 1. Badcase Analysis. For each subfigure, left is the binary shape input and right is the predicted skeleton of a basic UNet. Since there is no texture information in this scenario, structure information is crucial to final predictions. However, the center area (the red boxes) of binary shape image is lacking structure information, prone to broken and finely skeletons.

As shown in Figure 1, a typical skeleton extraction network fails to predict skeleton pixels in the plain area of the input shape image. On the one hand, the kernel size of convolutional layers is limited, thus the pixels in the plain area can't adopt the guidance from object shape. On the other hand, purely shape inputs lose most structure information in the object (except the boundaries), making it hard to determine the label of pixels in the center of objects.

In this paper, we propose Context Attention Network (CANet) to alleviate the aforementioned problems. In order to extract contextual cues efficiently, we modify the vanilla UNet with several vital updates, which is proved to be useful to improve the accuracy of skeleton extraction. Besides, we find that using the distance transform image that contains more information about object structure as input can ease the difficulty of skeleton extraction, further improving accuracy. With these novel optimizations, CANet outperforms existing methods and other participants on the Pixel SkelNetOn benchmark.

2. Related Work

Traditional methods of skeleton extraction are usually divided into three parts. For example, Zhang et al. [19] pro-

pose a fast parallel thinning algorithm, which iteratively removes the boundary and corner pixels of the object. After several iterations, only a skeleton of the object remains. Lu et al. [7] improve this method by preserving necessary and essential structures which should not be deleted. However, these classical methods usually provide low accuracy results and are sensitive to noise at the edge.

Recently, many deep learning-based approaches have been proposed. Shen et al. [12] propose a fully convolutional network, which is designed to extract the skeleton in different scales from multi stages. In [2], the authors use a vanilla pix2pix model with distance transformation preprocessing to extract the skeleton pixels. This preprocessing can reduce the learning difficulty of the network. Panichev et. al [10] introduce a U-Net based approach for direct skeleton extraction and get high performance. In [13], the authors continue to study preprocessing methods and propose to use Smooth Distance Estimation (SDE) and Edge Transformation to preprocess the input data, which wins the 1st place in the Pixel SkelNetOn 2021 Challenge. Nguyen [9] makes improvements to the original U-Net architecture using the attention mechanism and exploiting the auxiliary tasks, which also got excellent performances.

3. Method

In this section, we will introduce our main method and the tricks used to improve model performance, which contains network design, data processing, and loss strategies.

3.1. Network Design

Following [2], we treat the skeleton extraction as a pixel-wise binary classification problem. The input is binary images without texture, which contains less information and can be regarded as a low-level task. Therefore, we use UNet network [11] as the baseline model. Based on this vanilla architecture, we make several optimizations to improve its learning capabilities. The model structure is shown in Figure 2.

Residual Block. In order to improve the effectiveness of the model, we replace the dual convolutional block (DualConvBlock) in UNet with the residual block (ResBlock) [3] since the residual structure can integrate semantic features of different convolutional layers while preventing the model from over-fitting. Specifically, we replace each DualConvBlock with 3 ResBlocks in all stages of encoder and decoder.

Context Attention Block. In the neural network design, the attention module has become a powerful architecture to improve the model effect. SENet [4] adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. Non-local [15] adopt spatial attention module to acquire the correlation between every position of the feature map. Inspired by these ideas, the Context Attention block contains spatial and channel attention simultaneously. As shown in Figure 3, the long-range space dependence is captured by matrix multiplication [15]. Then we use channel attention to reassign channel weight. ReLU function is added to increase the non-linear of the model [4]. At the same time, in order to reduce the calculation of the model, the number of channels in the middle convolutional layer is reduced to 1/16. Applying Context Attention Block at the beginning of the network hurts the performance, thus we only deploy it in the higher stages(3-5).

3.2. Data Processing

Initially, the binary shape image of the object is used as the input. However, as shown in Figure 4, we find that the output of the distance transform applied to the shape image contains more intuitive information closer to the skeleton of the object and can reduce the learning difficulty of the network. Meanwhile, we find that there are some holes in the output of the distance transform. The prediction has been significantly improved after filling the holes in the image. Then we try to concatenate the original image, the image with distance transform and the outline extracted by canny algorithm [1], but unfortunately, the score is lower than expected. Following [13], We also try to replace distance transform with soft distance transform (SDE) but do not have a good result.

3.3. Loss Design

Following [9], Auxiliary loss is adopted to boost the performance of the model. Feature maps with different resolutions extracted from each stage of the decoder are fed into a 1×1 convolutional layer to get low-resolution prediction results. The losses calculated from low-resolution skeletons with down-sample ground truth are added to the final loss.

In terms of the loss function, we use the combination of Dice Loss [8] and Focal Loss [5] in our experiments to optimize the network because skeleton extraction is a data-imbalanced task.

Dice Loss is commonly used in semantic segmentation tasks, which is defined as below:

$$L_{dice} = 1 - 2\frac{\sum_i y_i p_i + \epsilon}{\sum_i y_i + \sum_i p_i + \epsilon}, \quad (1)$$

where y_i is the target label, p_i is the predict result and ϵ is a small constant to avoid division by zero (We set $\epsilon = 1.0$ in all experiments).

Besides, we use weighted focal loss, which is defined as below:

$$L_{focal} = W_{pos} p^{\gamma} log(p) + W_{neg}(1-p)^{\gamma} log(1-p), \quad (2)$$

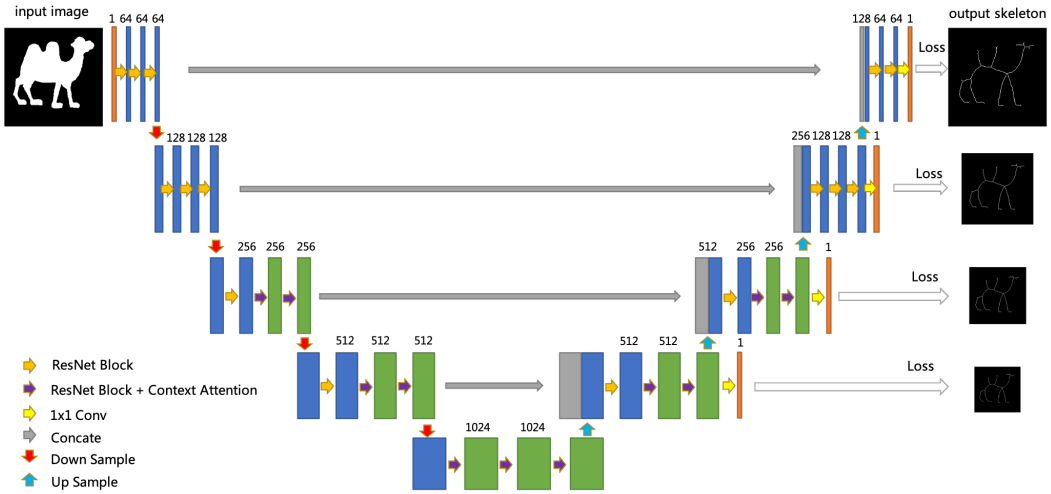

Figure 2. Context Attention Network: Our network is based on an encoder-decoder structure. Each stage of the encoder and decoder has a sequence of 3 ResBlocks. We add context attention block at stages 3-5 of the network. At the end of each encoder stage, max pooling is applied. The decoder stage contains a transposed convolutional layer, concatenated with a corresponding output of the encoder stage. At the end of the encoder stage, a 1×1 convolution is applied to generate skeleton predictions.

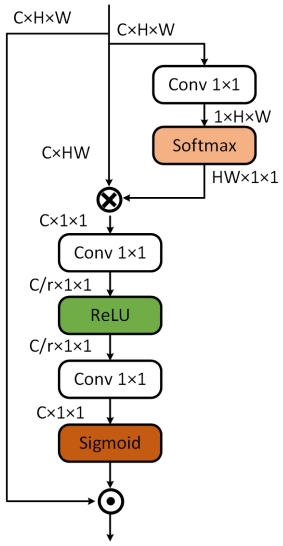

Figure 3. Context Attention Block: This Block is combined with spatial attention and channel attention. Firstly, matrix multiplication is adopted to acquire the correlation between every position of the feature map. Then we use channel attention to reassign channel weight.

where W_{pos} and W_{neg} are the weights of positive and negative class respectively, p is the probability that the sample belongs to positive class and γ is the focusing parameter.

The final loss formula is:

$$L_{total} = \lambda_{dice} L_{dice} + \lambda_{focal} L_{focal}, \tag{3}$$

Where λ_{dice} and λ_{focal} are hyperparameters to balance the value of losses. We set $\lambda_{dice} = 1.0$ and $\lambda_{focal} = 100.0$.

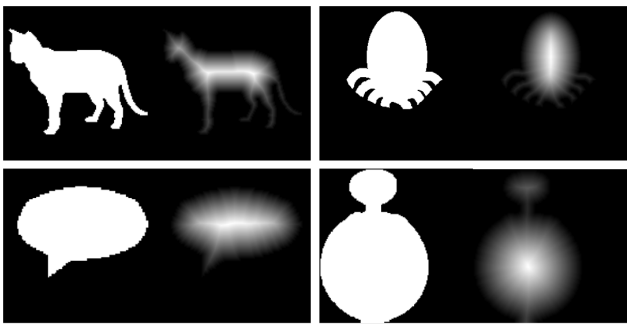

Figure 4. Comparison of the binary shape image and the output of the distance transform. Pixels in the latter are more distinguishable, making the skeleton extraction easier.

4. Experiments

In this section, we first detail our experimental settings, followed by a comparison between our method and other methods. Then we conduct the ablation studies with different parameters. Finally, we visualize the results of our approach.

4.1. Experimental Settings

Datasets. Our model is trained on the Pixel SkelNetOn Dataset provided by the SkelNetOn 2022 Challenge [2]. Pixel SkelNetOn Dataset contains 1,725 binary images with resolution 256×256, which splits into 1,218 training images, 241 validation images and 266 test images. We divide the training dataset into the training set and the split-test set

as a ratio of 80%:20%.

Implementation Details. The SGD optimizer with a learning rate of 0.02 and cosine annealing algorithm is used. We also use the F1 score to evaluate the model performance. Following [9], adaptive threshold selection is adopted to select the best threshold on the split-test set.

4.2. Main Results

Our method is compared with participants shown on the Pixel SkelNeton leaderboard. As shown in Table 1 and in Table 2, whether in development or final phase, our method outperforms current methods with a large margin.

Table 1. Leaderboard of Pixel SkelNetOn Challenge (Development phase).

Rank	Team name	F1 score
1	**huangzixuan0508(Ours)**	**0.8220**
2	neptuneai	0.7972
3	lv.zf	0.7935
4	_likyoo	0.7846
5	Young_Ji	0.7400
6	kyriemelon	0.6745
-	1st Place of Pixel SkelNetOn, 2021	0.8129

Table 2. Leaderboard of Pixel SkelNetOn Challenge (Final phase).

Rank	Team name	F1 score
1	**huangzixuan0508(Ours)**	**0.8507**
2	Young_Ji	0.8359
3	lv.zf	0.8333
4	jiliushi	0.8299
5	_likyoo	0.8289

4.3. Ablation Study

In this part, we record our attempts to improve model performance in different ways, including network design, input format, and loss design.

4.3.1 Network design

Table 3 shows the results of different network architectures. After adding ResBlock, the representation ability of the model is improved, thus the F1 score is raised from 0.788 to 0.801. When Context Attention Block is added to the deep stages of the model, high-level semantic features can be weighted in spatial and channel dimensions, which improves the skeleton extraction score. At the same time, we find that the results get worse when adding Context Attention Block to the shallow stages of the model. Usually, the shallow layers are used to extract low-level features without high-level semantic information, and this ability may be adversely affected by attention modules.

Table 3. Ablation experiments on network architectures. "RB" means "ResBlock", "CA" means "Context Attention". "1-5" means stage1 to stage5, "3-5" means stage3 to stage5.

Network Architecture	F1 score
UNet	0.788
UNet + RB	0.801
UNet + RB + CA 1-5	0.801
UNet + RB + CA 3-5	**0.806**

4.3.2 Data Processing

Table 4 shows the results of different input formats. Initially, we use 80% training data of binary shape image and get a score of 0.806. We try to use the instance segmentation network [16] to predict the shapes of RGB images in Image SkelNetOn Challenge. 2630 images were selected and added to the training set. Unexpectedly, the score is decreased, due to the inconsistent data distribution from different datasets. In terms of the input data format, images processed by distance transform are used as the input with a score of 0.803. After the hole areas are repaired, the score is further improved to 0.822. Furthermore, we also try to concatenate binary images and the outputs of distance transform as input data, but the score is not improved. We also implement the Soft Distance Transform used in [13], but the results get worse.

Table 4. Ablation experiments on the input data format of the network.

Input data	F1 score
Shape	0.806
Shape + RGB	0.777
Distance	0.803
Shape + Distance	0.794
Repaired distance	**0.822**
Soft distance	0.799

4.3.3 Loss

In this part, we show the influence of different loss functions. From Table 5 we can find that combining Focal Loss

and Dice Loss improves networks' accuracy. Besides, we find that the weights of positive and negative samples are important for skeleton extraction. Specifically, adjust W_{pos} from default 50 to 0.01, W_{neg} from 0.1 to 0.99 can further improve F1 score from 0.782 to 0.788.

Table 5. Ablation experiments on loss strategies.

Loss	F1 score
Focal Loss (W_{pos}=50.0, W_{neg}=0.1)	0.778
Dice Loss	0.770
Dice Loss + Focal Loss (W_{pos}=50.0, W_{neg}=0.1)	0.782
Dice Loss + Focal Loss (W_{pos}=0.01, W_{neg}=0.99)	**0.788**

4.4. Visualization

We visualize the predicted skeletons of our method and the baseline method in Figure 5. With the help of techniques proposed in CANet, the skeletons of plain area can be recovered successfully.

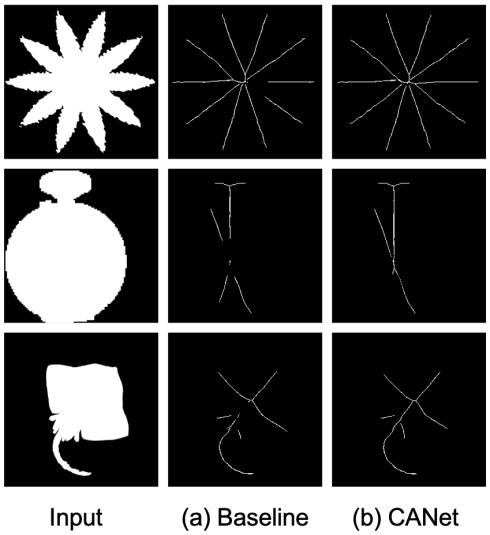

Figure 5. Qualitative comparison. Here, we compare CANet with the baseline on the validation dataset.

5. Conclusion

In this paper, we propose the Context Attention Network for skeletonization. With only 80% of the training data and a single model, our method achieves 0.822 and 0.8507 on the F1 score metric during the development and final phase of Pixel SkelNetOn Challenge, ranked as top-1 on the leaderboard. But in some cases, the problem of line breaks has not been well solved. In the next step, we will further explore the model structure and data processing method to reduce line breaks in the prediction process.

References

[1] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 2

[2] 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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[6] Chang Liu, Yunjie Tian, Zhiwen Chen, Jianbin Jiao, and Qixiang Ye. Adaptive linear span network for object skeleton detection. *IEEE Transactions on Image Processing*, 30:5096–5108, 2021. 1

[7] H. E. Lü and P. S. P. Wang. A comment on a fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 29(3):239–242, 1986. 2

[8] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 2

[9] Nam Hoang Nguyen. U-net based skeletonization and bag of tricks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2105–2109, 2021. 2, 4

[10] 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. 1, 2

[11] 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. 2

[12] W. Shen, K. Zhao, Y. Jiang, Y. Wang, X. Bai, and A. Yuille. Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images. *IEEE Transactions on Image Processing*, 26(11):5298–5311, 2017. 2

[13] Xiaojun Tang, Rui Zheng, and Yinghao Wang. Distance and edge transform for skeleton extraction. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 2136–2141, 2021. 2, 4

[14] Chenxi Wang, Yunfeng Wang, Zixuan Huang, and Zhiwen Chen. Simple baseline for single human motion forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2260–2265, 2021. 1

[15] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[16] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020. 4

[17] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 1

[18] Cong Yang, Oliver Tiebe, Kimiaki Shirahama, and Marcin Grzegorzek. Object matching with hierarchical skeletons. *Pattern Recognition*, 55:183–197, 2016. 1

[19] Tongjie Y Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984. 1