Supplementary Materials

Clinical electrocardiogram digitization based on U^{2}Net

Fei Deng[1,2], Hao Xu[1,2], Zhexin Li[1,2], Xiaokun Qin[3], Zheng Lou[1,2], Xiantao Song[3], Lili Wang[1,2], Dongsheng Kong[4], Yanlong Ren[5]

[1]State Key Laboratory of Semiconductor Physics and Chip Technologies, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China.
[2]Center of Materials Science and Optoelectronic Engineering, University of Chinese Academy of Sciences, Beijing 100049, China.
[3]Department of Cardiology, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China.
[4]Department of Neurosurgery, First Medical Center, People's Liberation Army (PLA) General Hospital, Beijing 100853, China.
[5]Department of Cardiology, Beijing Anzhen Hospital, Capital Medical University, Beijing 100029, China.

Correspondence to: Prof. Xiantao Song, Department of Cardiology, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China. E-mail: song0929@mail.ccmu.edu.cn; Prof. Lili Wang, State Key Laboratory of Semiconductor Physics and Chip Technologies, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China; Center of Materials Science and Optoelectronic Engineering, University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: liliwang@semi.ac.cn; Dr. Dongsheng Kong, Department of Neurosurgery, First Medical Center, People's Liberation Army (PLA) General Hospital, Beijing 100853, China. E-mail: kdsdoc301@163.com; Yanlong Ren, Department of Cardiology, Beijing Anzhen Hospital, Capital Medical University, Beijing 100029, China. E-mail: renyanlong@yeah.net

METHODS

Baseline architectures

In this study, four different baseline models U-Net[1], Res-UNet[2], Attention U-Net[3], and DeepLabV3[4]—were trained under a supervised learning framework to perform ECG signal segmentation. The following section elaborates on the specific architecture and implementation details of each baseline model.

U-Net is a symmetrically structured supervised learning framework consisting of an encoder and a decoder. The encoder comprises multiple repeating modules, each containing two 3×3 convolutional layers followed by a ReLU activation function, and a 2×2 max pooling layer for downsampling. This process doubles the number of feature map channels while halving the spatial dimensions, thereby extracting multi-scale contextual information. The decoder employs transposed convolutions to upsample the feature maps, doubling the spatial size and halving the number of channels. The upsampled features are then concatenated with the corresponding feature maps from the encoder pathway via skip connections. The concatenated feature maps are further processed through two 3×3 convolutional layers with ReLU activation for feature fusion. These skip connections directly transmit high-resolution, low-semantic features from each block of the encoder to the decoder at the same scale, enabling the decoder to recover fine-grained details and compensate for spatial information lost during downsampling.

Res-UNet is a framework built upon the U-Net backbone, in which the standard convolutional blocks are replaced with residual blocks. The skip connections within the residual blocks allow gradients to propagate directly to shallower layers, significantly alleviating the vanishing gradient problem. This facilitates the construction of deeper models and enhances the network's feature extraction capacity.

Attention U-Net incorporates an attention gate mechanism into the skip connections of the U-Net architecture. In this mechanism, the features g from the encoder and the upsampled features x from deeper decoder layers are linearly transformed, summed, and then passed through a ReLU activation function followed by a 1×1 convolution. A Sigmoid function is applied to generate an attention coefficient matrix with values between 0 and 1. This matrix is then multiplied element-wise with the original encoder features g, producing weighted feature maps. The attention gate leverages high-level semantic information from the decoder to generate attention maps that highlight salient regions and suppress irrelevant information, which are subsequently passed to the decoder.

DeepLabV3 typically employs a pre-trained classification network as its feature extractor and integrates multi-scale contextual information using dilated convolutions and pyramid pooling. Its core component is the Atrous Spatial Pyramid Pooling (ASPP) module, which applies multiple parallel dilated convolutional layers with different dilation rates alongside a global average pooling branch. The dilated convolutions capture context at various scales, while the global average pooling branch extracts image-level global contextual information. The outputs from all branches are upsampled to a consistent size and then concatenated. Finally, a 1×1 convolution is applied to fuse the multi-scale information. The output of the ASPP module is upsampled to the original image resolution to produce the final segmentation map.

Digitalization algorithm

In this work, we applied a signal extraction tool to perform digitization on the segmentation results produced by the CU^2-NeXt model. A novel dynamic programming approach was employed, which preserves both the high resolution of the ECG signal and the high consistency of its waveform.

The Viterbi dynamic programming algorithm[5] was utilized for signal extraction[6]. Based on dynamic programming, it aims to find the most probable sequence of hidden states, i.e., the most likely path of the ECG signal. First, each column (or horizontal slice) of the input binary ECG image is traversed to identify all contiguous regions of "true" pixels, and their center points are marked. By scanning the image from left to right, a topologically ordered list of points is generated for each column. Subsequently, a dynamic programming (DP) table is constructed to store the optimal path and its cumulative score for each node. During initialization, all points in the first column are entered into the DP table as base cases with zero initial cost. Then, the points in each subsequent column, from the second to the last, are processed sequentially. For each point in the current column, the algorithm examines all possible connections to candidate points from the previous column (or the nearest non-empty column). The total cost of each connection consists of two components: the cumulative cost to reach the candidate point, and the transition cost from that candidate to the current point. The transition cost is computed by a custom cost function C, defined as follows:

$$C(p_1, p_2, a) = D(p_1, p_2) \times \alpha + S(a, \angle(p_1, p_2)) \times (1-\alpha),$$
$$S(a_1, a_2) = \frac{(180 - |a_1 - a_2|)}{180} \tag{S1}$$

where p_1 denotes the candidate point, p_2 denotes the current point, D represents the Euclidean distance between the two points, a is the angle of the connection reaching the candidate point, $\angle(\cdot)$ is the angle between the two points, and α is a weighting parameter that balances the relative importance of distance and angular alignment in the total cost.

The algorithm selects the candidate point with the minimum total cost as the optimal predecessor of the current point, and records both the connection and the cost in the DP table. If no valid connection can be established with any predecessor, the current point is treated as a new starting point with a cost of zero. The cost function is designed to prioritize spatially proximate points while penalizing distant connections or those with large directional changes, thereby adhering to the continuous, smooth, and non-abrupt nature of typical signals in images. The parameter α was set to 0.5, a value determined through empirical analysis on the development dataset. After the entire DP table is constructed, the endpoint of the globally optimal path is identified by retrieving the node with the smallest cumulative cost within the last 20 columns of the image. The complete path is then reconstructed by backtracking through the stored back pointers in the DP table. Finally, the sequence of points is converted into a corresponding array of y-coordinates, with linear interpolation applied to any missing columns, resulting in a complete signal trajectory.

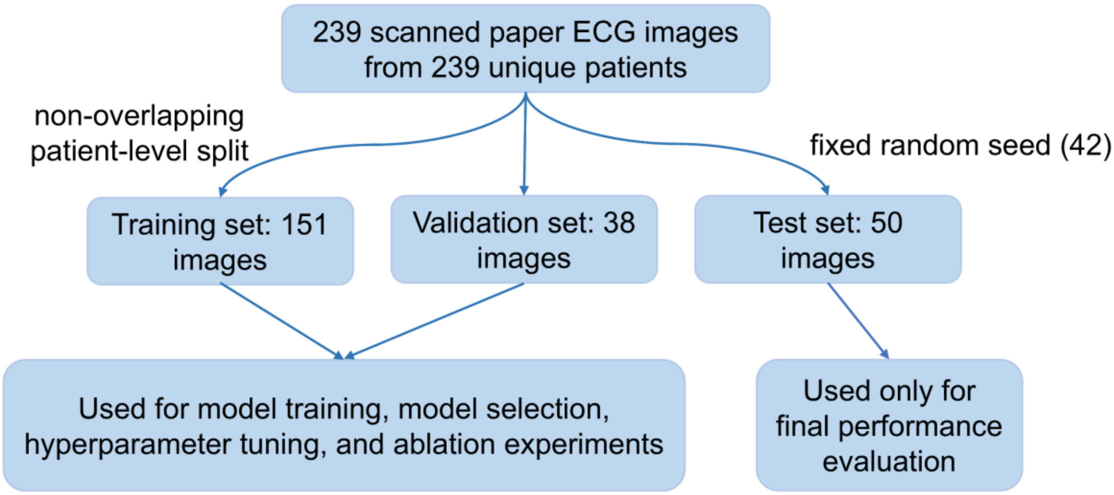

Figure 1. Data-flow diagram of dataset partitioning and model evaluation. The dataset consisted of 239 scanned paper-based ECG images from 239 unique patients. Each image appeared in only one subset. Ablation experiments were performed using the complete 189-image development dataset. The 50-image held-out test set remained separate and was used only for final performance evaluation.

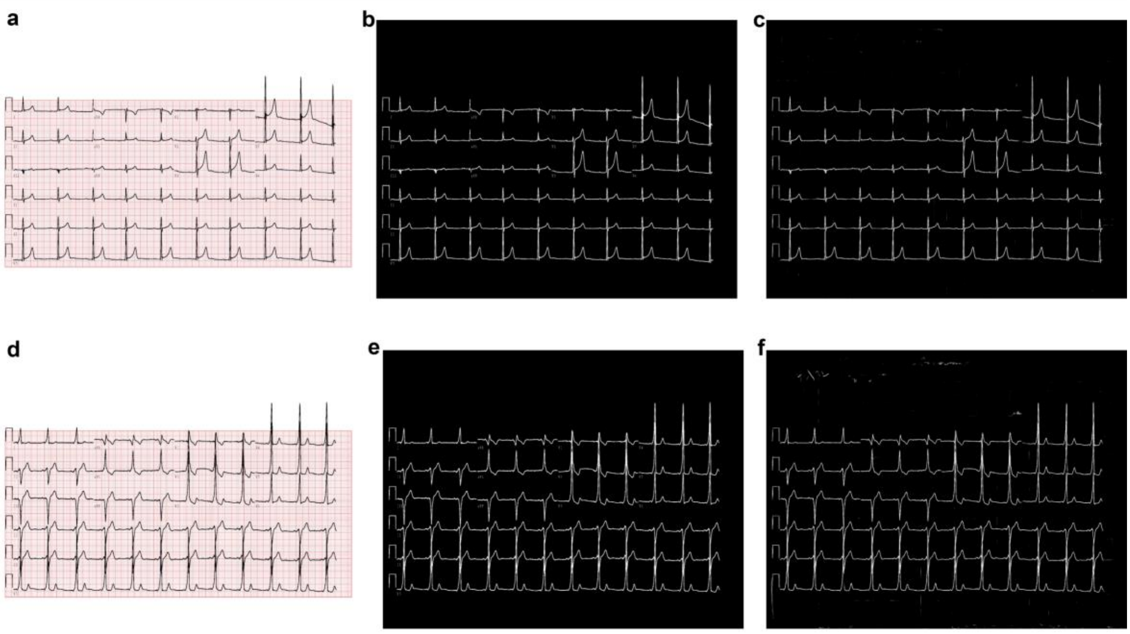

Figure 2. Comparison between preprocessing-generated waveform masks and masks derived from original digital ECG signals for two representative patients. (A) Original ECG image from patient 1; (B) Original digital ECG signal from patient 1; (C) Preprocessing-generated waveform mask for patient 1; (D) Original ECG image from patient 2; (E) Original digital ECG signal from patient 2; (F) Preprocessing-generated waveform mask for patient 2.

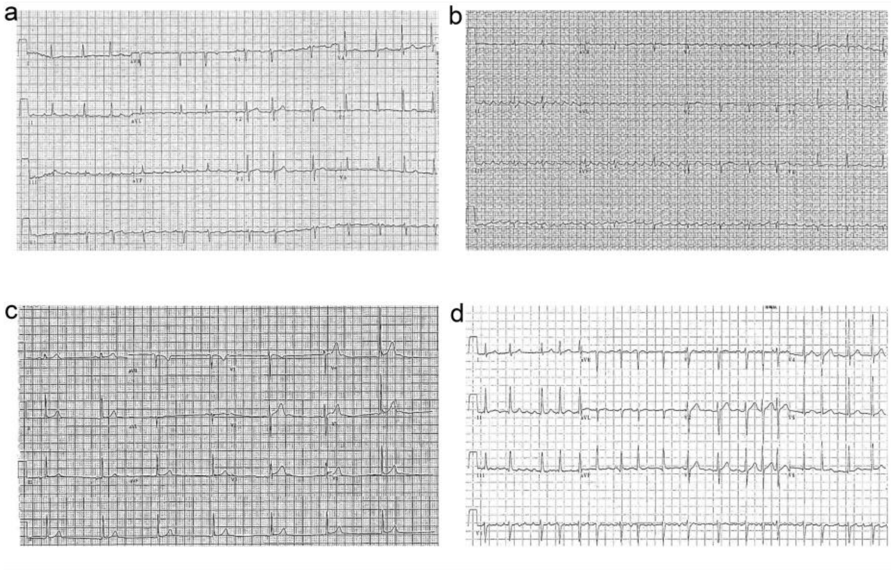

Figure 3. Representative examples of scanned ECG recordings with four different types of noise from Hospital 301. (A) Medium-noise ECG image; (B), (C) High-noise ECG image; (D) Low-noise ECG image. All representative ECG images shown in this figure have been de-identified. Patient names, medical record numbers, barcodes, examination dates, and other textual identifiers were removed or masked before presentation.

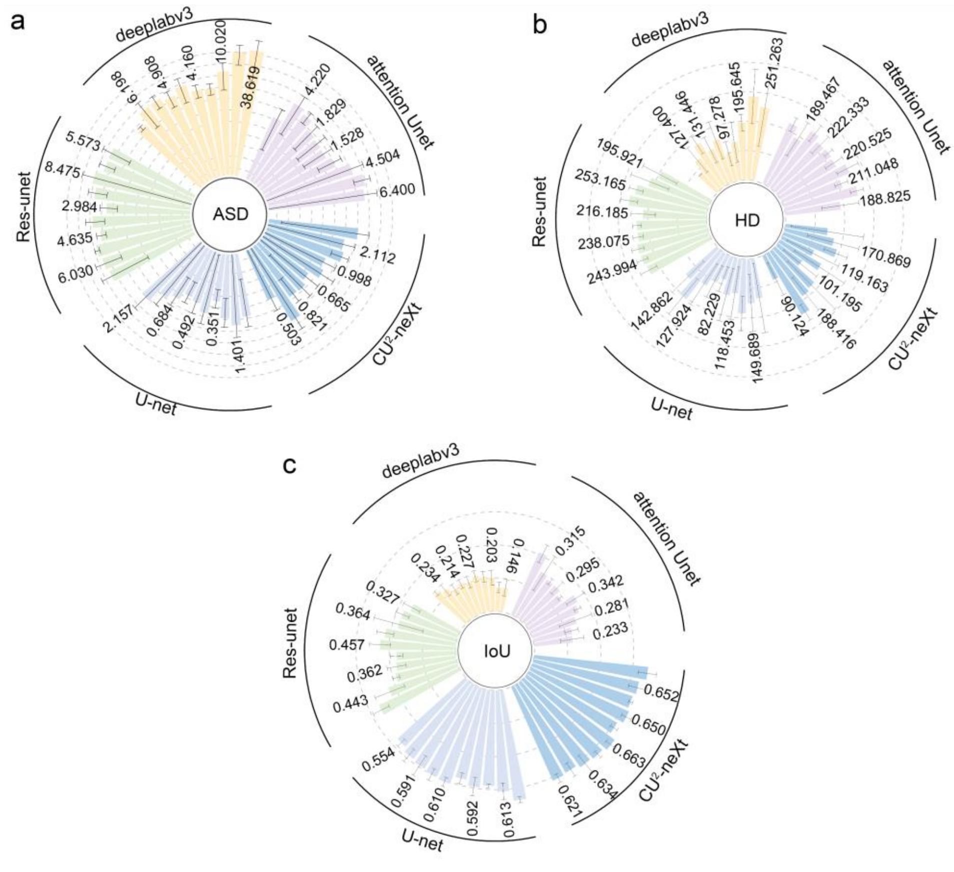

Figure 4. Quantitative ECG segmentation results. (A) Comparison of the average surface distance (ASD) between CU^2-NeXt and the baseline models on scanned ECGs

from Hospital 301. (B) Comparison of the Hausdorff distance (HD) between CU^2-NeXt and the baseline models on scanned ECGs from Hospital 301. (C) Comparison of the Intersection over Union (IoU) between CU^2-NeXt and the baseline models on scanned ECGs from Hospital 301.

Supplementary Table 1. Comparison between preprocessing-generated waveform masks and masks derived from original digital ECG signals

Case	IoU	Dice
1	0.6548	0.7914
2	0.6876	0.8149
3	0.5708	0.7268
4	0.5898	0.7420
5	0.7473	0.8554
6	0.6367	0.7780
Mean ± SD	0.6478 ± 0.0647	0.7847 ± 0.0473

Supplementary Table 2. Values are reported at the ECG image level

Baseline model	Metric	CU^2-NeXt mean	Baseline mean	Mean paired difference	Bootstrap 95% CI	*P* value	Adjusted P value
U-Net	Macro F1	0.9637	0.9333	0.0304	0.0269–0.0344	1.78E-15	5.33E-15
U-Net	Dice	0.7781	0.7379	0.0401	0.0353–0.0451	3.55E-15	7.11E-15
U-Net	IoU	0.6378	0.5862	0.0516	0.0458–0.0576	3.55E-15	7.11E-15
Attention U-Net	Macro F1	0.9637	0.8910	0.0727	0.0544–0.0935	1.78E-15	5.33E-15
Attention U-Net	Dice	0.7781	0.4443	0.3338	0.3151–0.3538	1.78E-15	5.33E-15

Attention U-Net	IoU	0.6378	0.2896	0.3482	0.3343–0.3630	1.78E-15	5.33E-15
Res-UNet	Macro F1	0.9637	0.8132	0.1505	0.1333–0.1691	1.78E-15	5.33E-15
Res-UNet	Dice	0.7781	0.5510	0.2270	0.2066–0.2486	1.78E-15	5.33E-15
Res-UNet	IoU	0.6378	0.3859	0.2519	0.2343–0.2701	1.78E-15	5.33E-15
DeepLab-related baseline	Macro F1	0.9637	0.6643	0.2994	0.2811–0.3197	1.78E-15	5.33E-15
DeepLab-related baseline	Dice	0.7781	0.3265	0.4516	0.4357–0.4688	1.78E-15	5.33E-15
DeepLab-related baseline	IoU	0.6378	0.1972	0.4406	0.4284–0.4536	1.78E-15	5.33E-15

Positive mean paired differences indicate higher CU^2-NeXt values for Macro-F1, Dice coefficient, and IoU. Confidence intervals were estimated using paired bootstrap resampling. P values were obtained using Wilcoxon signed-rank tests and adjusted using the Holm–Bonferroni method. Tests were two-sided, and P values were adjusted within each baseline comparison across Macro-F1, Dice coefficient, and IoU. An adjusted *P* value < 0.05 was considered statistically significant.

Supplementary Table 3. Five-fold cross-validation results of CU^2-NeXt

Fold	Macro F1	Dice	IoU	HD	ASD
1	0.876	0.762	0.617	4.517	0.244
2	0.892	0.792	0.659	1.721	0.209
3	0.887	0.782	0.643	1.869	0.225
4	0.864	0.738	0.586	10.350	0.452

5	0.884	0.778	0.638	1.861	0.219
Mean ± SD	0.881 ± 0.011	0.770 ± 0.021	0.629 ± 0.028	4.064 ± 3.704	0.270 ± 0.103

The analysis was performed at the patient level and was used as a supplementary assessment of internal stability rather than external validation. Values in the final row are reported as mean ± SD across the five folds

Supplementary Table 4. Subgroup analysis of segmentation performance according to ECG format and noise level

Subgroup	n	IoU	Macro F1	Dice
12-lead ECG	1	0.6362	0.9673	0.7777
13-lead ECG	38	0.6560 ± 0.0330	0.9661 ± 0.0059	0.8118 ± 0.0259
15-lead ECG	11	0.5750 ± 0.0510	0.9550 ± 0.0145	0.6616 ± 0.0446
Low-noise images	9	0.5555 ± 0.0427	0.9593 ± 0.0125	0.6979 ± 0.0331
Medium-noise images	14	0.6365 ± 0.0567	0.9638 ± 0.0124	0.7235 ± 0.0460
High-noise images	27	0.6659 ± 0.0302	0.9651 ± 0.0059	0.8331 ± 0.0241

Supplementary Table 5. Paired statistical analysis results comparing CU^2-NeXt and U^2-Net+CBAM on 50 test images

Baseline model	Metric	CU^2-NeXt mean	Baseline mean	Mean paired difference	Bootstrap 95%CI	P value	Adjusted P value
U^2-Net+CBAM	Macro F1	0.9637	0.9486	0.0151	0.0124–0.0176	4.38E-11	1.31E-10

U^2-Net+CBAM	Dice	0.7781	0.7377	0.0403	0.0318–0.0483	9.08E-11	1.61E-10
U^2-Net+CBAM	IoU	0.6378	0.5846	0.0532	0.0420–0.0639	8.06E-11	1.61E-10

Supplementary Table 6. Agreement analysis of clinical interval measurements between original and digitized ECG signals

Parameter	MAE (ms)	RMSE (ms)	Bias (ms)	SD of differences (ms)	Lower 95% LoA (ms)	Upper 95% LoA (ms)
PR interval	16.88	25	-7.57	23.92	-54.46	39.32
QRS duration	12.11	16.46	0.13	16.51	-32.23	32.49
QT interval	21.84	32.53	-7.09	31.84	-69.5	55.31

REFERENCES

1. Olaf R, Philipp F, Thomas B. U-net: Convolutional networks for biomedical image segmentation. *in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* 2015;9351. [DOI: 10.48550/arXiv.1505.04597]

2. Foivos ID, François W, Peter C, Chen W. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm.* 2015;162:94-114. [DOI: 10.1016/j.isprsjprs.2020.01.013]

3. Ozan O, Jo S, Loic LF, et al. Attention u-net: Learning where to look for the pancreas. *ArXiv.* 2018;1804.03999. [DOI: 10.48550/arXiv.1804.03999]

4. Chen LC, Zhu YK, George P, Florian S, Hartwig A. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Computer Vision – ECCV 2018.* 2018;11211:833-851. [DOI: 10.1007/978-3-030-01234-2_49]

5. Huang L. Advanced dynamic programming in semiring and hypergraph frameworks. *Coling 2008 Organizing Committee.* 2008;1-18. [https://aclanthology.org/C08-5001/]

6. Julian DF, Natalie EC, Kazi TH, Hetal P, Larisa GT. Digitizing ECG image: A new method and open-source software code. *Clin. Physiol.* 2015;221:106890. [DOI: 10.1016/j.cmpb.2022.106890]