Automatic Tooth and Background Segmentation in Dental X-ray Using U-Net Convolution Network

Conference Paper · October 2020

DOI: 10.1109/ICSITech49800.2020.9392039

3 authors:

Arna Fariza
Electronics Engineering Polytechnic Institute of Surabaya
86 PUBLICATIONS **152** CITATIONS
SEE PROFILE

Agus Zainal Arifin
Institut Teknologi Sepuluh Nopember
168 PUBLICATIONS **788** CITATIONS
SEE PROFILE

Eha Renwi Astuti
Airlangga University
28 PUBLICATIONS **97** CITATIONS
SEE PROFILE

Automatic Tooth and Background Segmentation in Dental X-ray Using U-Net Convolution Network

Arna Fariza
Informatics and Computer Eng. Dept.
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
arna@pens.ac.id

Agus Zainal Arifin
Informatics Engineering Dept.
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
agusza@cs.its.ac.id

Eha Renwi Astuti
Dentistry Dept.
Universitas Airlangga
Surabaya, Indonesia
eharenwi@gmail.com

***Abstract*— Tooth and background segmentation in dental X-ray is used to produce an area of a tooth by removing areas of tissue and other neighboring teeth. This presents challenges due to the large number of superimposed (overlapping) images of teeth between the adjacent and adjacent teeth and the difficulty of determining the area of the tooth with other tissues. This study proposes a new approach for automatic segmentation of dental X-ray images using the U-Net convolution network. The stages used in the training process consist of data augmentation, pre-processing with Contrast Limited Adequate Histogram Equalization (CLAHE) and gamma adjustment, and training with the U-Net architecture. While the testing process consists of pre-processing, prediction, and removing small areas in the background. The average accuracy of the U-Net convolutional network segmentation accuracy achieved excellent results, 97.60%.**

Keywords—dental X-ray, tooth and background segmentation, U-Net

I. Introduction

Deep learning currently provides excellent performance for image classification [1], segmentation [2], detection and tracking [3], and text writing [4]. Efficient automated processing without human involvement to reduce human error and also reduce overall time and cost. Due to the slow and tedious process of manual segmentation approaches, there is a significant demand for computer algorithms that can segment quickly and accurately without human interaction [5].

There were several traditional machine learning and image processing techniques available for medical segmentation before the deep learning revolution, for example segmentation based on histogram features [6], region-based segmentation methods [7], and graph-cut approaches [8]. However semantic segmentation approaches that utilize deep learning have become very popular in recent years in the fields of medical image segmentation, lesion detection, and localization [9]. In addition, a deep learning-based approach is known as a universal learning approach, where a single model can be used efficiently in various medical imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), and X-ray.

Panoramic radiography is one of the extraoral and non-invasive radiographs that has been used generally in dentistry to obtain a picture of the entire maxillofacial. Panoramic radiographic images have problems, such as uneven lighting due to superimposition and other positioning errors that can interfere with the object segmentation process [10]. The quality of X-ray images from panoramic radiography has low resolution which contributes to the noise in the image, so the first step to processing dental X-ray images is to distinguish between region of interest (ROI) and background [11]. The main function in image segmentation is to divide the image into constituent regions or objects, which helps the feature extraction stage to extract more accurate and distinctive features.

Tooth and background segmentation in dental X-ray is used to produce an area of a tooth by removing areas of tissue and other neighboring teeth. This presents challenges due to the large number of superimposed (overlapping) images of teeth between the adjacent and adjacent teeth and the difficulty of determining the area of the tooth with other tissues.

Tooth and background segmentation using conditional spatial Fuzzy C-means with gaussian kernel-based (csFCM-GK) [12] is able to segment well dental X-ray images which have clear margins between the tooth and its neighbors, but cannot be performed on superimposed teeth.

This study proposes a new approach for automatic segmentation of dental X-ray images using the U-Net convolution network. The stages used in the training process consist of data augmentation, pre-processing with Contrast Limited Adequate Histogram Equalization (CLAHE) and gamma adjustment, and training with the U-Net architecture. While the testing process consists of pre-processing, prediction, and removing small areas in the background.

II. U-Net Convolution Network

The convolution neural network (CNN) model architecture for classification requires a unit of coding and provides a class probability as output. In contrast to classification, the segmentation architecture requires encoding and decoding of convolutional units. An encoding unit is used to encode the input image onto a large number of maps with lower dimensions. The decoding unit is used to perform up-convolution (de-convolution) operations to produce a segmentation map with the same dimensions as the original input image.

One of the earliest and most popular approaches to semantic medical image segmentation is called the "U-Net" [13]. The U-Net basic model diagram is shown in Figure 1. The U-Net network structure consists of two main parts: the encoding unit and the convolutional decoding. Basic convolutional operation is followed by ReLU activation in both parts of the network. For bottom sampling in the encoding unit, a max-pooling operation of 2 × 2 is performed.

In the decoding phase, a transpose convolution operation (representing up-convolution, or de-convolution) is performed to sample the feature map. The first version of the U-Net was used to cut and copy the feature map from the encoding unit to the decoding unit. The U-Net model provides several advantages for segmentation: first, it allows for the simultaneous use of global locations and contexts. Second, this model works with very few training samples and provides better performance for segmentation [13]. Third, end-to-end processing of the entire image in the forward pass and directly generate the segmentation map. This ensures that the U-Net maintains the full context of the input image, which is a major advantage.

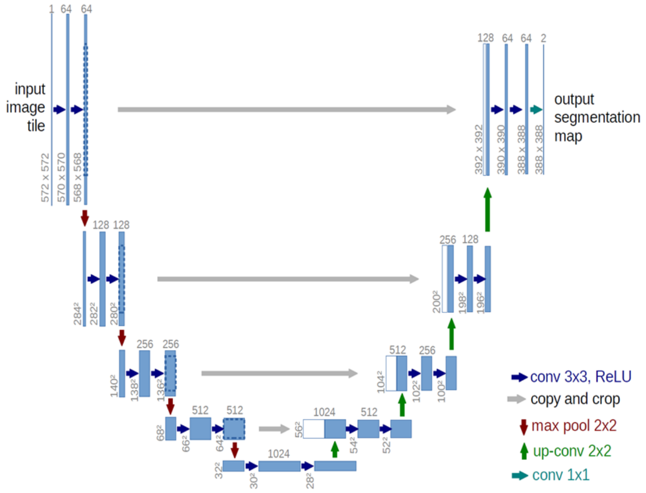


Fig. 1. U-Net convolution architecture of Ronneberger et al., 2015 [13].

The U-Net architecture consists of three parts: contraction, bottleneck, and expansion section. The contraction section is made up of many contraction blocks. Each block requires input using two 3 × 3 convolution layers followed by a maximum pooling of 2 × 2. The number of kernels or feature maps after each block is multiplied so that the architecture can study complex structures effectively. The lower layer mediates between the contraction layer and the expansion layer. It uses two CNN 3 × 3 layers followed by a 2 × 2 convolution layer. The essence of this architecture lies in the expansion section. Similar to the contraction layer, it also consists of several expansion blocks. Each block passes input to two CNN 3 × 3 layers followed by 2 × 2 layer up-sampling. Also after each the number of map feature blocks used by the convolutional layer get half to maintain symmetry. However, each time the input is also added by the feature map of the corresponding contraction layer. This will ensure that the features learned when contracting the image will be used to reconstruct it. The number of expansion blocks is equal to the number of contraction blocks. After that, the resulting mapping passes through another CNN 3 × 3 layer with the number of feature maps equal to the number of desired segments.

III. METHODOLOGY

This study aims to separate the part of the teeth from a dental X-ray image with the background, the parts of the right and left teeth, and other tissues as shown in Figure 2 (a). The results of the tooth and background segmentation can be seen in Figure 2 (b) where the teeth are marked in white color and the rest (background and other tissues) in black color.

The dental and background segmentation methodology with the U-Net convolution network on the dental X-ray image can be seen in Figure 3. The ROI image is augmented by flip and rotation to expand the training dataset. The training dataset is pre-processed consisting of Contrast Limited Adequate Histogram Equalization (CLAHE) and gamma adjustment. Next, the process of splitting the training dataset and validation is carried out. The data set is carried out by a training process using U-Net architecture with certain training parameters to produce the best model. The best model is used to predict the segmentation results on the test dataset.

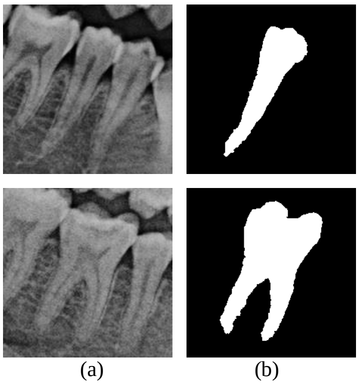


Fig. 2. Tooth segmentation and background (a) ROI image; (b) segmentation results.

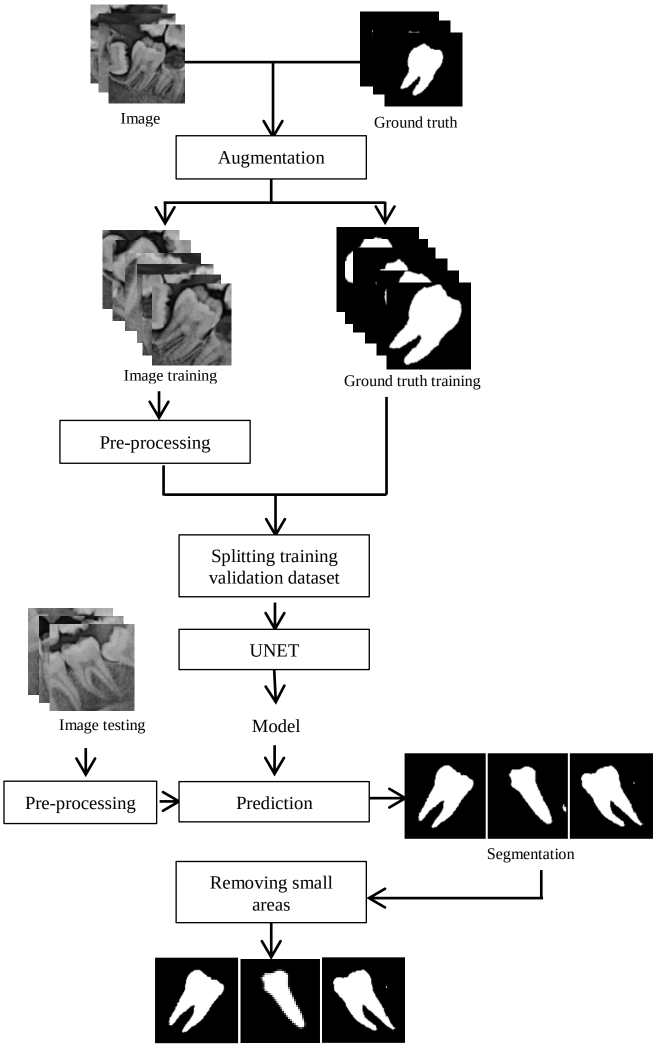


Fig. 3. Metodhology tooth and background segmentation using U-Net.

A. Image Dental X-ray

The input data is in the form of ROI which is cut from panoramic radiographs manually with a size of 224 × 224. Panoramic radiographs come from the results of shooting at the Parahita Laboratory, Sidoarjo, Indonesia. ROI is taken from the molar and premolar area of the mandible provided that the crown area to the root of the tooth is visible, there are no fillings, caries, cut teeth, overlapping teeth and other tooth decay. The examples of ROI results can be seen in Figure 2 (a). For the training process, 119 image data sizes of 224 × 224 were formed. Each data image was paired with a ground truth that was formed with validation from a dental forensic expert.

B. *Augmentation*

The original image and ground truth dataset measuring 224 × 224 as many as 119 data were first carried out by the augmentation process. Data augmentation is very important to train the network to be invariant and have the desired resistance, when only a few training samples are available [13]. In the dental X-ray image, it uses horizontal mirroring augmentation and random rotation in order to be resistant to deformation and gray value variations, which are the concept of training segmentation in very little annotated images. The augmentation process uses the Augmentor library to form 5000 data used as training dataset. Horizontal mirroring is formed with probability 0.5 and random rotation between -25° to + 25°.

C. *Pre-processing*

The original image dataset was pre-processed using Contrast Limited Adequate Histogram Equalization (CLAHE) and gamma adjustment. Both of these pre-processing processes are widely used in medical grayscale image data. The histogram calculation for each region is carried out directly. This function is generally obtained using the CDF (cumulative distribution function) calculation. If the number of pixels and grayscale in the respective regions, respectively *X* and *Y*, and $h_{i,j}(k)$ for $k = 0,1,2,. . . , X - 1$ is the region histogram (*i*, *j*), then the corresponding CDF calculation, scaled with (*X* - 1) for the grayscale mapping, is:

$$f_{i,j}(k) = \frac{(X-1)}{Y} \cdot \sum_{k=0}^{X-1} h_{i,j}(k) \tag{1}$$

Equation (1) changes the density function in the gray scale image. In order to handle the increase in the contrast area to the maximum, we limiting the contrast value to the desired level by limiting the slope of the maximum value using the boundary value β for the intersection of all histograms. This limit value (clip limit) can be related to the so-called clip factor, α (in percent), as follows:

$$\beta = \frac{Y}{X}\left\{1 + \frac{\alpha}{100}(s_{max} - 1)\right\} \tag{2}$$

In the context of exploratory contrast enhancement of an image, gamma applications are used to adjust the threshold in image processing [14]. The gamma correction function is used to correct the exposure of the image to handle the exposure that is incorrectly captured. The gamma correction function is used to map luminance levels to compensate for the effects of non-linear luminance on display devices. Gamma can be any value between 0 and infinity. If gamma is 1 (default), the mapping is linear. If gamma is less than 1, the mapping is weighted towards a higher (brighter) output value. If gamma is greater than 1, mapping is weighted towards a lower (darker) output value. After pre-processing with CLAHE and gamma adjustment, the next process is normalizing the original image and ground truth dataset to be between 0 and 1.

D. *Splitting*

The original image and ground truth dataset are stored in a different file directory but have the same index. Before the training process is carried out, the data is stored in the tensor data structure. The input training dataset of 5000 image augmentation results are randomly divided into training and validation dataset. Validation dataset is needed to obtain the best model during training. The training and validation dataset are set with a composition of 0.9 and 0.1.

E. *Training*

The U-Net convolution network architecture used as shown in Figure 4 is set to accept image input = (1, 224, 224) and output = (1, 224, 224). The image input is grayscale (1 channel) measuring 224 × 224, while the output becomes 1 target class with a size of 224 × 224 which states white as the target. The U-Net architecture consists of 4 block of 2D convolution and 3 block of 2D deconvolution. Input size 1 × 224 × 224 is convoluted with a 3 × 3 kernel and produces 64 filters followed by ReLu and maximum pooling. Furthermore, the same convolution process is carried out and produces 128, 256 and 512 filters. After that, a deconvolution process was carried out using a bilinier upsample to produce 256, 128 and 64 filters respectively. The final step is to return the size of the modeled image to the original size of 1 × 224 × 224.

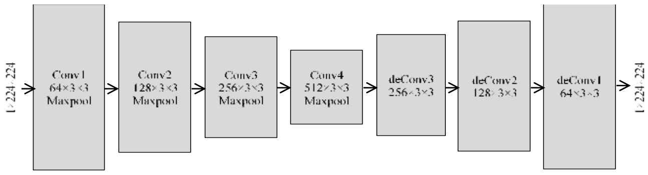


Fig. 4. U-Net structure architecture for image input 1×224×224 and output 1×224×224.

Model parameters, loss function, optimizer and learning rate scheduling used in the training process can be seen in Table I. The loss function used is binary cross-entropy with logits as measured by Equation 3.

$$\ell(x,y) = L = \{l_1, \dots, l_N\}^T, l_n$$

$$= -w_n\left[y_n \cdot \log\sigma(x_n) + (1-y_n)\cdot\log\left(1-\sigma(x_n)\right)\right] \tag{3}$$

where *N* is the batch size and

$$\ell(x,y) = \begin{cases} mean(L), jika\ reduction = 'mean' \\ sum(L), jika\ reduction = 'sum' \end{cases} \tag{4}$$

TABLE I. TRAINING PARAMETER FOR SEGMENTATION TRAININIG USING U-NET CONVOLUTION NETWORK

Model	U-Net
Fungsi loss	Binary cross-entropy with logits
Optimizer	Adam with learning rate = 1e-4
Scheduler and learning rate	StepLR with step = 30 and gamma = 0.1

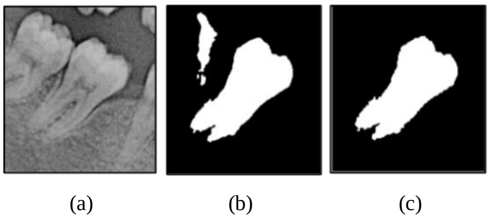
Binary cross-entropy loss (BCELoss) called sigmoid cross-entropy loss which combines sigmoid and cross-entropy loss is used for binary classification. Whereas binary cross-entropy with logits loss combines sigmoid layers and BCELoss in a single class which is numerically more stable than using ordinary sigmoid followed by BCELoss. Optimizer uses Adam with initial learning rate = 1e-4 and decreases learning rate using every 30 steps (epoch) with gamma = 0.1.

The outcome of the training process is the best model for prediction. The training process is carried out by initializing the model parameters, loss function, optimizer and learning rate scheduling. At each epoch, training and validation data are fed into the U-Net architecture. After that, the calculation of the loss value is carried out between the U-Net model output and the ground truth. In the training process, the loss value is carried out backward processes to produce smaller losses and optimization based on the specified learning rate. Training on the validation dataset will produce the optimal model with the smallest loss. This model is stored for use in the dental and background prediction process.

F. Prediction

The prediction made on the test image dataset, in this case is the ROI data that will be carried out by the tooth and background segmentation process. Before testing the testing dataset, pre-processing CLAHE and gamma adjustment, the same as the training dataset.

The result of the predictive pixel is in the form of a value between -1.0 to +1.0, if> 0.0 then the pixel is a tooth pixel (value = 1.0) and if ⩽ 0.0 then the pixel is a background pixel. An example of the prediction results can be seen in Figure 5. Figure 5 (a) is a testing image and Figure 5 (b) is the prediction result. However, the predicted image in Figure 5 (b) still has small areas on the background that are undesirable so that it must be removed as shown.



(a) (b) (c)

Fig. 5. Fig. 5. Image segmentation result with U-Net convolution network (a) image testing; (b) segmentation result; (c) after removing small areas.

G. Removing Small Areas

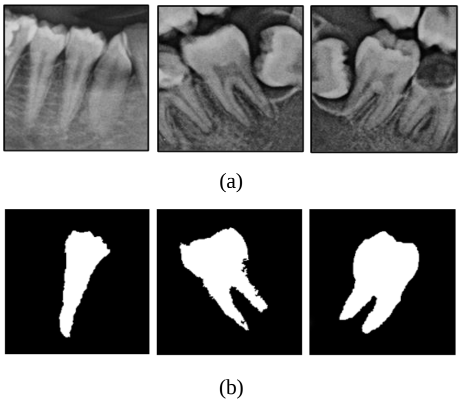
After the cluster formation stage, a process is added to remove unwanted small areas in the background using the region property function. The algorithm for eliminating these small areas is as follows

1. Determine the 4 L component relationships from the black-and-white image
2. Create an object that contains the actual number of pixels in the 4 component relationship area L
3. Sort all area objects in descending manner and generate sorted area objects and area indexes
4. Calculate the largest area from L based on the sorted objects
5. In the largest area, remove any holes in the area.

The results of the prediction process using the U-Net model and the results of removing small areas in the background with the region property function can be seen in Figure 5 (c).

IV. Result and Discussion

The segmentation output has a size of 224 × 224. The training data consisted of 119 image data and ground truth respectively. The training data is stored in a folder that stores the original image data and ground truth in a different folder. An example of training data and ground truth can be seen in Figure 6. As testing data, 10 data were randomly selected with various contrasts and different inhomogeneities as an evaluation of the results of segmentation of the U-Net network convolution.



(a)

(b)

Fig. 6. Image training dataset (a) original image; (b) ground truth image.

Evaluation is done by looking at the performance of the training process with the proposed U-Net convolution network and evaluating the results of the segmentation. The evaluation of segmentation results is calculated based on confusion matrix to measure accuracy by comparing the classification results with the references they have (ground truth) [15]. By using confusion matrix, it can be calculated from several performance measurement matrices, such as accuracy, sensitivity, and specificity. The image resulting from the U-Net segmentation and ground truth consists of 2 intensity values, namely 255 which shows the tooth image and 0 which indicates the background.

A. Training Evaluation

Evaluation of convolutional network training performance is carried out to see a decrease in training loss and validation loss during the training process as shown in Figure 7. The training process is run on an NVIDIA RTX2070 GPU engine and requires execution time per epoch between 343 - 436 seconds. The training and validation processes show good performance because the loss training and validation values decrease exponentially and reach convergence after the 40th epoch of 100 epochs. The best validation loss is 0.038590 which is stored as a prediction model.

B. Segmentation Testing Result

Evaluation of the results of the proposed U-Net convolutional network segmentation of 10 dental X-ray image data measuring 224 × 224 was evaluated based on ground truth images. The accuracy evaluation is based on the convolution matrix of the U-Net convolution network segmentation results. The accuracy, sensitivity and specificity values have the same value because the results of the U-Net convolutional network segmentation and ground truth only consist of 2 grayscale intensity values, 255 and 0. Table II shows the average accuracy of the U-Net convolutional network segmentation accuracy achieves very good results 97.60%.

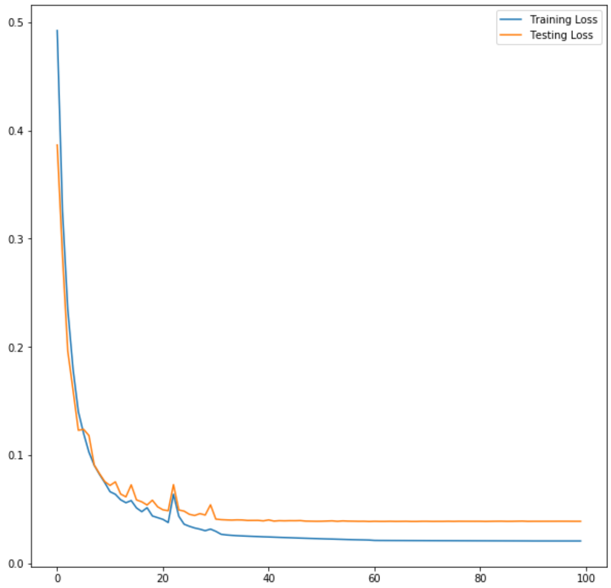


Fig. 7. Graph of decreasing training loss and validation loss by 100 epochs

TABLE II. ACCURACY OF U-NET SEGMENTATION TESTING RESULT

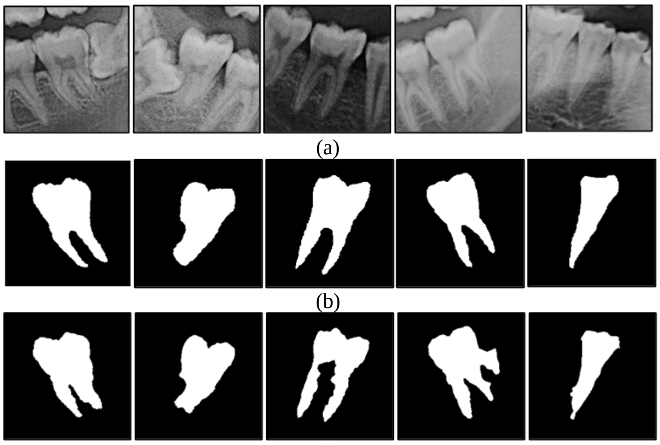
Data	Accuracy
1	0.9712
2	0.9768
3	0.9836
4	0.9865
5	0.9542
6	0.9792
7	0,9603
8	0.9885
9	0.9746
10	0.9859
Average	**0.9760**

The visualization of several images resulting from the U-Net convolution network segmentation can be seen in Figure 8. Figure 8 represents some image data with different contrasts and inhomogeneities. Figures 8 (a) and 8 (b) are the original images and ground truths, while the results of the segmentation of the U-Net convolutional network can be seen in Figure 8 (c). Visually, the image of the proposed U-Net convolution network segmentation shows a smooth image and is closer to the ground truth image.

C. Discussion

Several segmentation analyzes with the U-Net convolutional network on the dental X-ray image test dataset are as follows:

- Segmentation with the U-Net convolution network which results in fast segmentation and smooth image edges after obtaining the optimal model from the training process.
- Segmentation with the U-Net convolutional network requires a training dataset that is not too many, namely 119 image data and predicting large amounts of image data. The addition of augmentation processes (data expansion) and image pre-processing is effective in producing an optimal model that is used as a prediction of segmentation results.
- Segmentation with the U-Net convolutional network requires sufficient ground truth to produce an optimal training model. This takes a relatively long time to establish and validate ground truths.


(a)

(b)

(c)

Fig. 8. Visualization of teeth and background segmentation results (a) original image; (b) ground truth; (c) segmentation result using U-Net convolution network.

In general, the segmentation results using the U-Net convolutional network deep learning method on the dental segmentation and background dental X-ray images show superior results. Data varies widely, especially for image data that has biased boundaries in the tissue at the root of the tooth and overlapping images on the enamel portion.

Evaluation of the results of segmentation was also carried out on the 1907 testing image data. There were 278 image segmentation results that were not suitable for processing as classification data or 14.58%. An example of an image that is not successfully segmented properly using the U-Net convolution network can be seen in Figure 9. Figure 9 is the result of multiplying the original image with the segmented image to show the difference more clearly

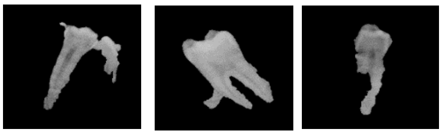


Fig. 9. Example of poorly segmented tooth image using the U-Net convolution network.

This shows that although the U-Net segmentation gets superior results, it does not have the ability to avoid segmentation errors. In terms of the segmentation process time, U-Net segmentation is able to segment collectively in datasets stored in a folder quickly.

V. CONCLUSION

The dental and background segmentation methods using the U-Net convolution network deep learning method showed superior results and approached the ground truth. The average accuracy of the U-Net convolutional network segmentation accuracy achieved excellent results, 97.60%. However, because the difference in data intensity values varies greatly, especially for image data that has biased boundaries on the tissue at the root of the tooth and overlapping images on the enamel, from the data it causes 14.58% to produce segmentation errors in 1907 image testing.

ACKNOWLEDGMENT

We would like to thank Politeknik Elektronika Negeri Surabaya who has supported this research.

REFERENCES

[1] A. Krizhevsky, I. Sutskever and E. Geoffrey, "ImageNet classification with deep convolutional neural networks," Advances in neural information processing systems, pp. 1097-1105, 2012.

[2] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in The IEEE conference on computer vision and pattern recognition, 2015.

[3] N. Wang, S. Li, A. Gupta and D. Y. Yeung, "Transferring rich feature hierarchies for robust visual tracking," arXiv preprint arXiv:1501.04587, 2015.

[4] J. Mao, W. Xu, Y. Yang, J. Wang , Z. Huang and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," arXiv preprint arXiv:1412.6632, 2014.

[5] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (R2U-net) for medical image segmentation," in arXiv preprint arXiv:1802.06955., 2018.

[6] N. Ramesh, J. H. Yoo and I. K. Sethi, "Thresholding based on histogram approximation. IEE Proceedings-Vision, Image and Signal Processing, 142(5), 271-279.," 1995.

[7] N. Sharma and A. K. Ray, "Computer aided segmentation of medical images based on hybridized approach of edge and region based techniques," in International Conference on Mathematical Biology', Mathematical Biology Recent Trends by Anamaya Publishers, 2006.

[8] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in eighth IEEE international conference on computer vision. ICCV, 2001.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi and C. I. Sánchez, "A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.," 2017.

[10] Z. Z. Akarslan, H. Erten, K. Güngör and I. Celik, "Common errors on panoramic radiographs taken in a dental school," Journal of Contemporary Dental Practice, vol. 4, no. 2, pp. 24-34, 2003.

[11] S. M. Kahaki, M. J. Nordin , N. S. Ahmad, M. Arzoky and W. Ismail, "Deep convolutional neural network designed for age assessment based on orthopantomography data," Neural Computing and Applications, pp. 1-12, 2019.

[12] A. Fariza, A. Z. Arifin, and E. R. Astuti, “Dental X-ray Image Segmentation using Gaussian Kernel-Based in Conditional Spatial Fuzzy C-means”, International Journal on Advanced Science, Engineering and Information Technology, Vol. 7 No. 6, pp. 2159-2167, 2017.

[13] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, Springer, Champ, 2015.

[14] A. Karthikeyan, P. Kala and A. Ramachandran, "Image Quality Improvement in Kidney Stone Detection on Computed Tomography Images," International Journal of Scientific Research in Science, Engineering and Technology, vol. 3, no. 3, p. 484–488, 2017.

[15] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," Remote sensing of Environment, vol. 62, no. 1, pp. 77-89, 1997.