creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

보건학 석사학위 논문

Development of Automated Volumetric Analysis of Canine Heart in Thoracic Radiograph Using Deep Learning

개 흉부 방사선 자료의 딥러닝 적용을 통한 심장 면적 자동 분석 방법 개발

2020 년 12 월

서울대학교 보건대학원

보건학과 유전체 & 건강빅데이터 전공

정여진

Development of Automated Volumetric Analysis of Canine Heart in Thoracic Radiograph Using Deep Learning

개 흉부 방사선 자료의 딥러닝 적용을 통한 심장 면적 자동 분석 방법 개발

지도교수 성주헌

이 논문을 보건학 석사 학위논문으로 제출함

2021 년 2 월

서울대학교 보건대학원

보건학과 유전체 & 건강빅데이터 전공

정여진

정여진의 석사 학위논문을 인준함

2021 년 2 월

위 원 장	조성일 (인)
부 위 원 장	황승식 (인)
위 원	성주헌 (인)

Abstract

Introduction : Measurement of canine heart size in thoracic lateral radiograph is crucial in detecting heart enlargement caused by diverse cardiovascular diseases. The purpose of this study was 1) to develop deep learning (DL) model that segments heart and 4th thoracic vertebrae (T4) body, 2) develop new radiographic measurement using calculated 2 dimensional heart area and length of T4 body, and 3) calculate performance of our new measurement to detect heart enlargement using echocardiographic measurement as gold standard.

Methods : Total 1,000 thoracic radiographic images of dog were collected from Seoul National University Veterinary Medicine Teaching Hospital from 2018. 01. 01 to 2020. 08. 31. Given ground truth mask, two Attention U-Nets for segmentation of heart and T4 body were trained using different hyperparameters. Among 1,000 images, model was trained with 800 images, validated with 100 images and tested with 100 images. Performance of DL model was assessed with dice score coefficient, precision and recall. New calculation method was developed to calculate heart volume and adjust by T4 body length, which was named vertebra-adjusted heart volume (VaHV). Correlation of VaHV of 100 test images and reported VHS (vertebral heart score) was assessed. With 188 images with concurrent echocardiographic examination, diagnostic performance of VaHV for detecting cardiomegaly was assessed by student's t-test, receiver operating characteristic (ROC) curve and area under the curve (AUC).

Results : The two trained DL model showed very good similarity with ground truth (dice score coefficient 0.956 for heart segmentation, 0.844 for T4 body segmentation). VaHV of 100 test images showed statistically significant correlation with VHS. VaHV showed better diagnostic performance in detecting left atrial enlargement and left ventricular enlargement than VHS.

Conclusions : DL model can be used to segment heart and vertebrae in veterinary radiographic images. Our new radiographic measurement obtained from DL model can potentially be used to assess and monitor cardiomegaly in dogs.

Keywords : Deep Learning, Semantic Segmentation, Veterinary Radiography, Heart Volume Analysis

Student Number : 2019-24084

Table of Contents

List of Figures

List of Tables

1. Introduction

Thoracic radiographs have been used as a first-line examination for the evaluation of various thoracic diseases including cardiomegaly [1,2]. In humans, cardiothoracic ratio (CTR) measurement on antero-posterior (AP) chest radiogrph is used to diagnose cardiomegaly, a condition that is strongly correlated with both congenital and congestive heart disease [3]. However, in dogs, diverse chest shape among dog breeds have limited the use of CTR [4]. In veterinary medicine, vertebral heart score (VHS) is most widely used to measure heart size in canine thoracic radiography. VHS was developed on the basis that there is a good correlation between heart size and body length regardless of the conformation of the thorax, so it has advantage over CTR in dogs. To calculate VHS in lateral radiographic views, the long axis of the heart is measured from the ventral border of the left main stem bronchus to the most distant ventral contour of the cardiac apex. Then the maximal short axis, which is perpendicular to the long axis, is measured from the cranial border of the heart where the diameter is greatest and mark the caudal border. The long axis measurement includes the left atrium and left ventricle, and the short axis measurement includes right atrium and left heart chambers. The long and short axes are then transposed onto the vertebral column and recorded as the number of vertebrae beginning with the cranial edge of 4^{th} thoracic vertebrae (T4). The sum of these values is the VHS. VHS has been shown to correlate with other means of measuring cardiomegaly (i.e., ECG, echocardiography) in dogs with progressively increasing heart size, and it is considered by some to be the gold standard in determining cardiomegaly in dogs [5]. Although it has diverse reference range for each dog breeds, VHS over 10.5 is commonly used as cutoff for clinical determination of cardiomegaly in adult dogs [4].

Detecting cardiomegaly in dogs is crucial since majority of cardiovascular diseases in dogs is characterised by progressive cardiac enlargement. Myxomatous mitral valve disease (MMVD) is the most prevalent cardiac disease in dogs [6]. It has been estimated that MMVD accounts for 75~80% of cardiac diseases in dogs [7]. It is characterised by a progressive degeneration of the valvular apparatus leading to mitral regurgitation (MR) [6]. VHS values in thoracic radiographs are important in assessing the severity of mitral regurgitation (MR) caused by MMVD by determining the presence of generalised heart enlargement (HE) and left atrial enlargement (LAE) [6, 7]. In order to define cardiac enlargement, guideline for staging MMVD provided by Consensus Statements of the American College of Veterinary Internal Medicine (ACVIM) uses VHS and echocardiographic measurement (Left atrial : Aorta ratio in the right-sided short axis view in early diastole (LA/Ao), Left ventricular internal diameter in diastole, normalized for body weight (LVIDDN)) to assess left atrial enlargement and left ventricular enlargement [8]. Since presence of cardiac enlargement determines whether treatment is needed or not, monitoring cardiac enlargement in thoracic radiography is important component in staging and treatment of cardiac disease.

Although VHS is widely used, it consists of two 1-dimensional lengths of heart. So in this study we aimed to measure 2-dimensional area of heart in lateral thoracic radiographs using semantic segmentation by deep learning and develop new measurement to assess volume of heart. Since it is known in earlier studies that vertebral body length has good correlation with heart size [4], we also segmented 4th thoracic vertebrae body to adjust calculated heart volume.

In humans, deep learning (DL) [9] have been successfully applied to automatically calculate CTR [10] using networks for sementic segmentation. However, little work has been done on segmentation in veterinary images. The purpose of this study was 1) to develop DL model for sementic segmentation of heart and T4 body in lateral thoracic radiographs of dogs, 2) develop new measurement of heart volume using segmented 2-dimensional heart area and T4 length (vertebra-adjusted heart volume, VaHV), 3) compare VaHV with VHS, and 4) evaluate the usefulness of VaHV in predicting cardiac enlargement, using echocardiographic measurement as a gold standard.

2. Materials and Methods

2.1 Data Collection

All thoracic radiographs of dogs used in this study were collected from Seoul National University Veterinary Medicine Teaching Hospital (SNU-VMTH) between January 2018 and August 2020. All radiographs had radiology reports referred from veterinary radiology specialists in SNU-VMTH. We collected right lateral thoracic radiographs, radiology reports, signalments, and study dates from Picture Archiving and Communication Systems (PACS). Total 1,000 lateral thoracic radiographs were randomly selected to train, validate, and test DL model. To train DL model with diverse patients, we did not include serial images of same patient in the dataset. In all radiology reports, veterinary radiology specialists in SNU-VMTH had reported VHS. We extracted VHS from radiology report using Python (Python Software Foundation. Python Language Reference, version 3.6. Available at http://www.python.org) and regular expression library.

For comparison of volumetric measurement obtained from DL model and echocardiographic measurements, we collected 188 studies with lateral thoracic radiographs which had a complete echocardiographic examination within 1 month from radiologic examination. From echocardiography reports, we extracted LA/Ao ratio and LVIDDN for further analysis.

All the images had undergone quality control. Poorly positioned and poorly exposed radiographs were initially excluded. For accurate segmentation of heart, cases with indistinct heart margin (due to pleural effusion, overlying pulmonary/mediastinal nodules, focal/disseminated alveolar pattern superimposed over the heart) were excluded. Cases with unreliable VHS (due to chest conformation, pectus excavatum, hemivertebrae) were also excluded. Entire flow of this study is shows in Figure 1.

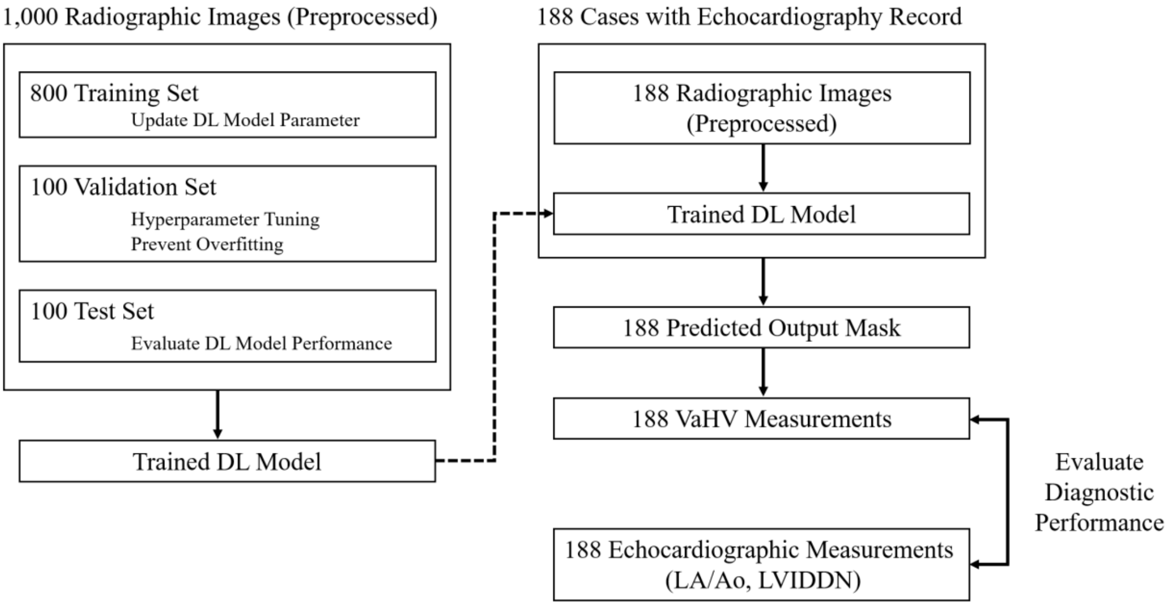

1,000 Radiographic Images (Preprocessed)
800 Training Set
Update DL Model Parameter
100 Validation Set
Hyperparameter Tuning
Prevent Overfitting
100 Test Set
Evaluate DL Model Performance
Trained DL Model
188 Cases with Echocardiography Record
188 Radiographic Images (Preprocessed)
Trained DL Model
188 Predicted Output Mask
188 VaHV Measurements
Evaluate Diagnostic Performance
188 Echocardiographic Measurements (LA/Ao, LVIDDN)

Figure 1. Flow of training DL model and evaluation in 188 cases. DL model for segmentation of heart and T4 body was trained with 800 images, validated with 100 images, and tested with 100 images. With trained DL model, 188 cases with concurrent radiographic images and echocardiographic measurements were used to evaluate diagnostic performance of new volumetric measurement (VaHV) obtained from DL model.

2.2 Development of DL model

2.2.1 Introduction to Semantic Segmentation

In deep learning, convolutional neural network (CNN) is widely used for image analysis tasks, including image classification, object detection, and sementic segmentation [11, 12]. With diverse operation layers including convolution and pooling, CNNs have outperformed the state of the art in many visual recognition tasks. The most typical use of CNN is image classification. DL model for image classification receives image as input, and the output to the image is a single class label.

Recently, many studies to develop CNN for semantic segmentation task had been conducted. Unlike image classification, semantic segmentation network needs ground truth mask which has same shape with input image as label. In ground truth mask, each pixel is assigned with class label. The desired output of semantic segmentation network has same size with input image, with each pixel assigned to predicted probability for the class.

U-Net [13] is most commonly DL network used for semantic segmentation of biomedical images (e.g., radiograph, microscope images). It is based on fully convolutional network [14], with some modifications that it works with very few training images and yields more precise segmentation results. U-Net consists of contracting path and expansive path. Contracting path down-samples image with convolution and max pooling followed by Rectified Linear Units (ReLU), and expansive path up-samples the image with transposed convolution. To compensate decrease in resolution while expanding images, U-Net uses skip connections that combine spatial information from the contracting path with the expansive path [13]. Beside original U-Net, many researchers developed modified U-Net with diverse architecture to improve segmentation performance.

2.2.2 Attention U-Net with Focal Tversky Loss, Surface Loss

2.2.2.1 Attention U-Net

Although skip connection in U-Net makes precise prediction, this brings many redundant low-level feature extractions, as feature representation is poor in the initial layers. To address this problem, Oktay et. al proposed Attention U-Net [15], which added a novel attention gate (AG) mechanism that allows the U-Net to focus on target structures of varying size and shape. AG uses soft attention, which works by weighting different parts of the image. Areas of high relevance are multiplied with a larger weight and areas of low relevance are tagged with smaller weights. As the model is trained, more focus is given to the regions with higher weights [16]. So the soft attention implemented at the skip connections actively suppress activations in irrelevant regions. Structure of AG is depicted in Figure 2.

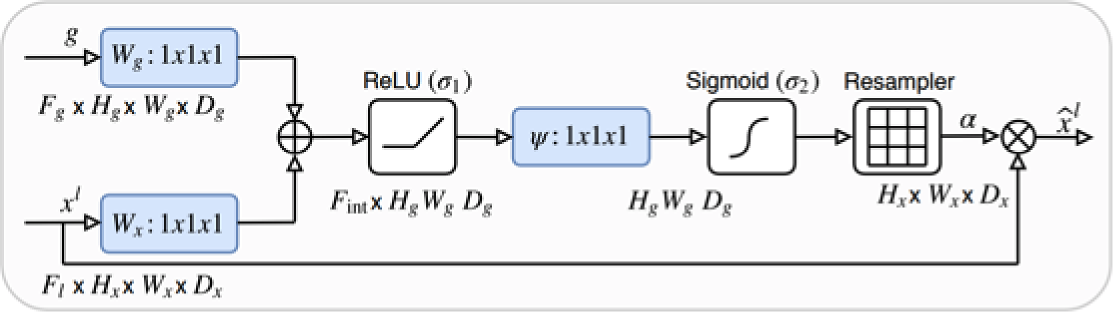


Figure 2. Structure of attention gate (AG) [15, 16]. AG takes in two inputs, vectors x^l and g. g is taken from the next lowest layer of the network. To match dimension, vector x^l goes through a strided convolution and g undergoes 1x1x1 convolution. Then the two vectors are summed element-wise. Resultant vector goes through a ReLU activation, 1x1 convolution and a sigmoid layer, producing the attention coefficients. The attention coefficients are upsampled to the original dimensions of the x^l vector using trilinear interpolation. The attention coefficients are multiplied element-wise to the original x^l vector, scaling the vector according to relevance. This is then passed along in the skip connection as normal [17].

2.2.2.2 Improved Attention U-Net with Focal Tversky Loss

Because areas of heart and T4 body were highly imbalanced in all radiographic images, in this study we used improved Attention U-Net with focal tversky loss proposed by Abraham, Nabila and N. Khan [18]. They improved Attention U-Net by two ways. First, they included a novel focal tversky loss function for highly imbalanced data and small regions of interest (ROI) segmentation. Secondly, they introduced a deep supervision to Attention U-Net, with a multi-scaled input image pyramid for better intermediate feature representations. The underlying architectures for DL model are shown in Figure 3.

In semantic segmentation task, dice score coefficient (DSC) is an overlap index that is widely used to assess segmentation quality. However, the limitation of DSC is that it equally weighs false positive (FP) and false negative (FN) detections. The Tversky similarity index is a generalization of the DSC which allows for flexibility in balancing FP and FNs. In tversky index, hyperparameter α can be tuned to control weights for FP and FNs. Tversky loss (TL) is defined as 1 – Tversky index. α value larger than 0.5 means that the tversky loss focus to minimize FN detections, and α value smaller than 0.5 means that the loss focuses to minimize FP detections. To further improve segmentation of small regions, Nabila and Khan introduced novel focal tversky loss (FTL), which is exponentiated version of TL. FTL includes hyperparameter γ to control between easy background and hard ROI training examples. To combat the over-suppression of the loss function, Nabila and Khan trained intermediate layers with FTL but supervised the last layer with TL to provde a strong error signal and mitigate sub-optimal convergence [18].

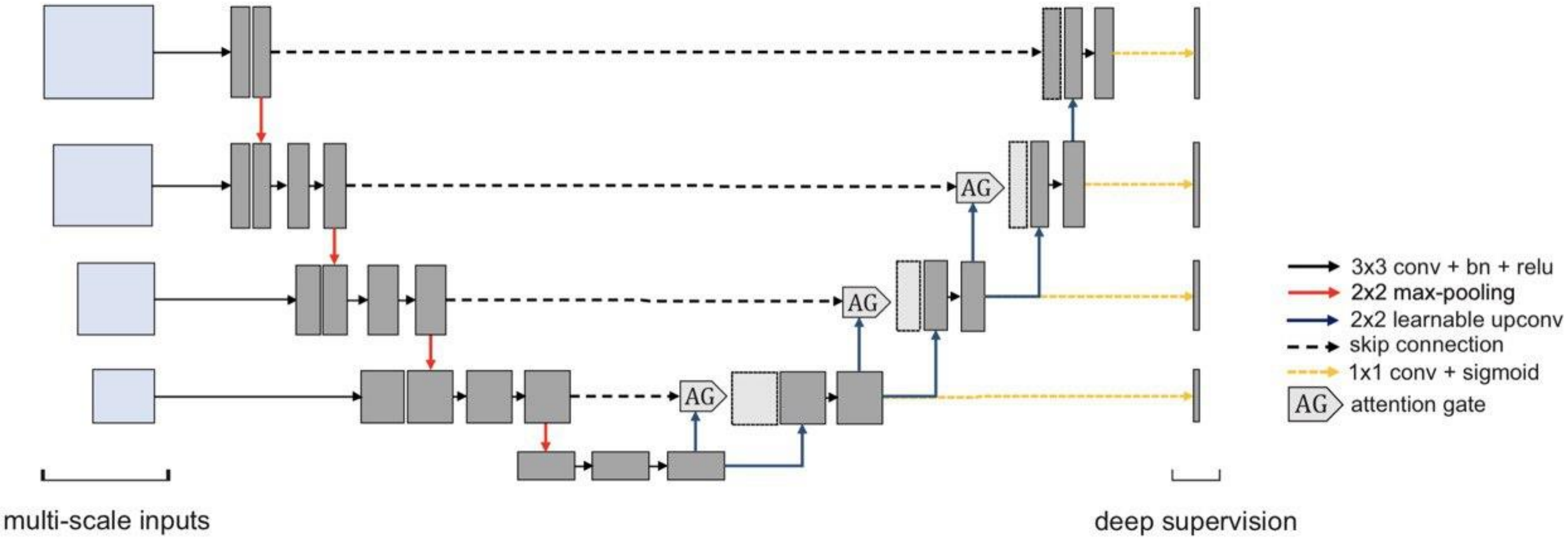


Figure 3. Architecture of improved Attention U-Net with input image pyramid and deep supervised output layers [18]. The network is composed of a contracting path to extract locality features and an expansive path, to resample the image maps with contextual information. Skip connections are used to combine high-resolution local features with low-resolution global features and encourage more semantically meaningful outputs. AGs are used to give larger weights to relevant region, and multi-scale inputs are concatenated to improve segmentation results [18].

2.2.2.3 Surface Loss

Since Tversky loss is region-based loss which aims to minimize the mismatch or maximize the overlap regions between ground truth and prdiected segmentation [19], it does not control the distance between ground truth and predicted segmentation. To reduce geometric distance between ground truth and predicted segmentation of heart, we added boundary loss [20] to heart segmentation DL model. Boundary loss uses integrals over the boundary to mitigate the difficulties of highly unbalanced segmentation.

2.2.3 Image Preprocessing

All radiographic images were exported in Joint Photographic Experts Group (JPEG) format. To make image size compatible with DL model, we center cropped images with Pillow library[1] in Python and resized images to 256 x 256 size. Since all images had different exposure level, we applied the Contrast Limit Adaptive Histogram Equalization (CLAHE) [21] to enhance contrast of images with open-cv library [22] in Python. We set clip limit as 2.0 and tile grid size as (8,8). Example of preprocessing original radiographic image is shown in Figure 4.

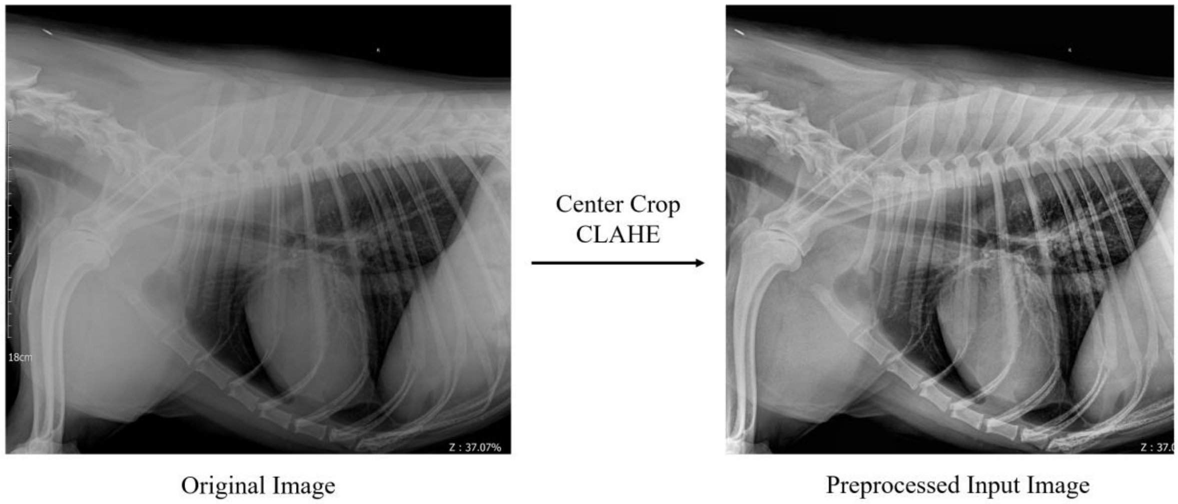
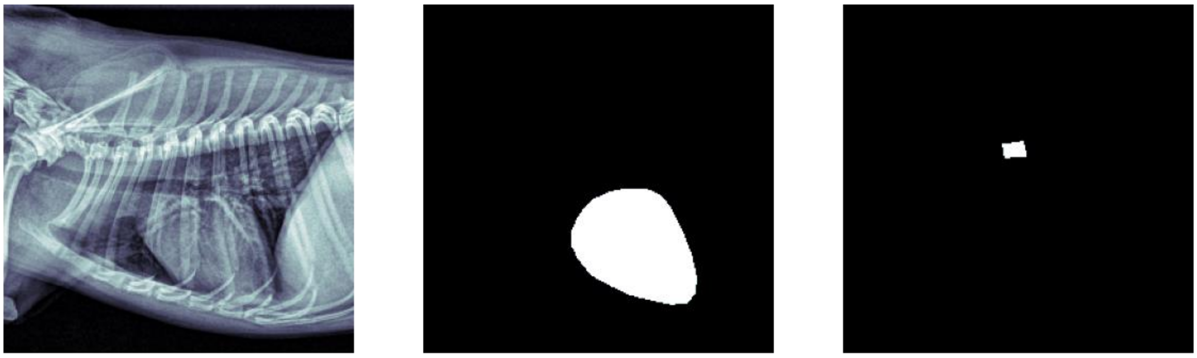


Figure 4. Example of preprocessing original radiographic image. First the center of image is cropped to make squard image, and the Contrast Limit Adaptive Histogram Equalization (CLAHE) [21] is applied to enhance contrast of image.

[1] https://github.com/python-pillow/Pillow

2.2.4 Establishing Ground Truth

CNN for semantic segmentation needs ground truth mask. For 1,000 images of training, validation, and test set, we made corresponding ground truth mask where each pixel value represented whether it belongs to background, heart or T4 body. We used labelme [23] opensource tool for annotation. Example of ground truth mask is shown in Figure 5.



(a) Input Image (b) Ground Truth of Heart (c) Ground Truth of T4 Body

Figure 5. Example of input image (left) and ground truth mask of heart (middle) and T4 body (right)

2.2.5 Training DL Model

With earlier experiments with original U-Net, there were two main problems in segmenting heart and T4 body areas. Because of RoI imbalance between heart and T4 body, predicted T4 masks were much larger than ground truth (high FP), and predicted heart maskes had high FNs. Since we needed to penalize FP and FN differently in each segmentation task, we trained improved Attention U-Net [18] separately to segment heart and T4 body with different hyperparameters. Flow of training and evaluating segmentation DL network is shown in Figure 6.

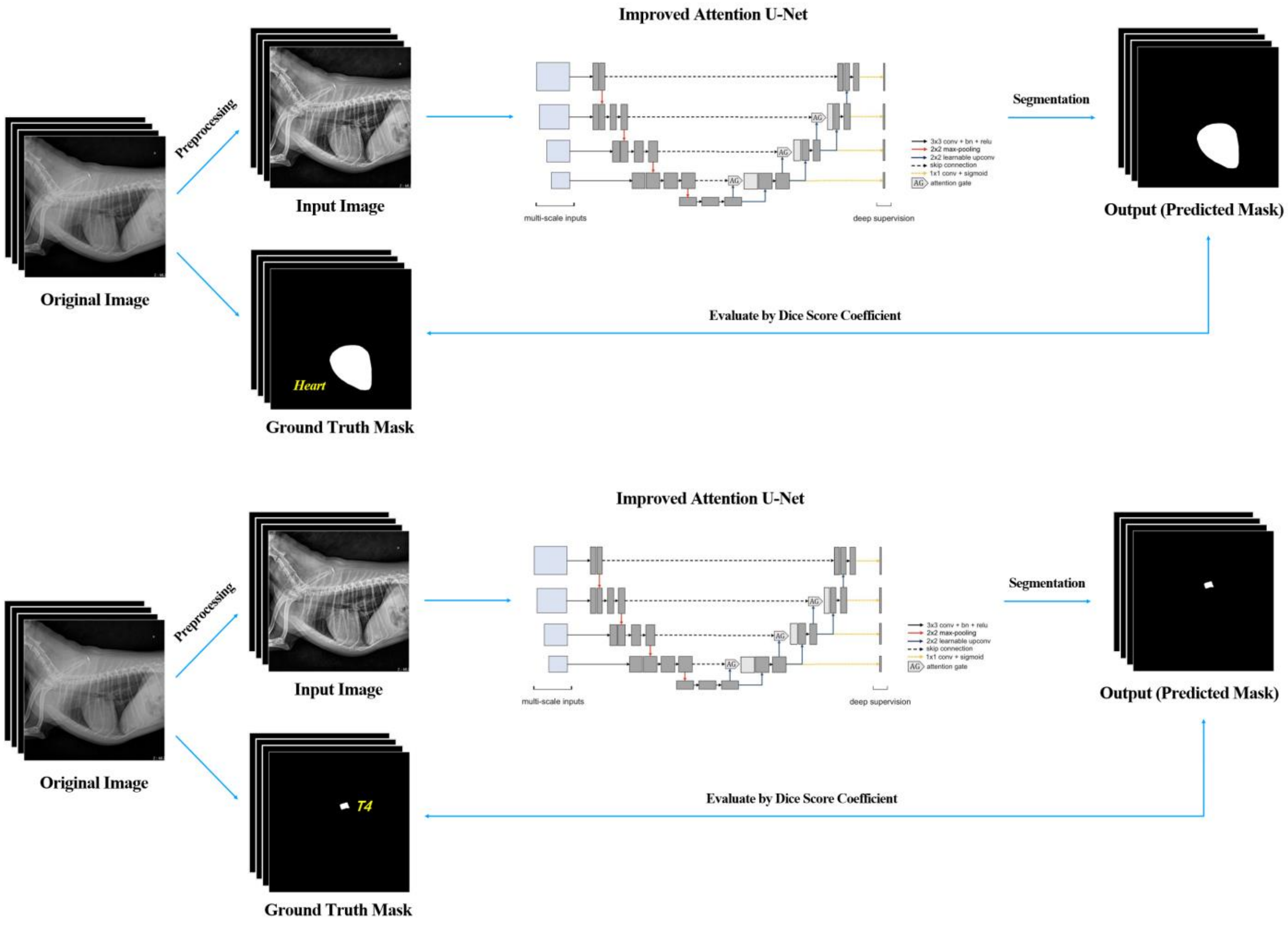


Figure 6. Flow of segmentation DL network training and evaluation. Improved Attention U-Net [18] receives both preprocessed input radiographic images and ground truth mask for heart (top) and T4 body (bottom). Network is trained by tversky loss and focal tversky loss. Ouput of the network is predicted mask of heart and T4 body. Performance of network is evaluated by dice score coefficient between ground truth mask and predicted mask.

2.2.5.1 DL Model for Heart Segmentation

DL model was implemented with Keras [24]. 1,000 images were randomly splitted into 800 training, 100 validation, 100 test images. The training set was used to update parameters of network, validation set was tune hyperparameters and callbacks (described below) to prevent overfitting of DL model, and test set was used to evaluate performance of trained network. The training and validation sets were randomly shuffled after every epoch. Stochastic gradient (SGD) optimizer with initial learning rate 0.01 and momentum value of 0.9 was used as an optimization algorithm. We set batch size as 16 and epoch size as 100. For callbacks, we used learning rate scheduler which automatically decrease learning rate by factor 0.1 when performance on validation set does not improve over 3 epochs, and early stopping to stop at optimal epoch based on performance on validation set. Tversky loss and Focal tversky loss were used as loss function. Intermediate layers used focal tversky loss, and final layer used tversky loss. For loss function of heart segmentation network, we set hyperparameter α for tversky loss as 0.7 to strongly penalize FNs. Hyperparameter γ for focal tversky loss was set to 4/3, which was chosen from paper [18]. Finally, for last layer we used a combined loss function which was weighted sum of tversky loss and boundary loss :

$$\text{Final Layer Loss} = (1-\delta) * \text{Tversky Loss} + \delta * \text{Boundary Loss}$$

We used δ value of 0.05. Plain dice score coefficient (DSC) was used as evaluation metrics.

2.2.5.2 DL Model for T4 Body Segmentation

Same training, validation, test images were used for T4 body segmentation. The training and validation sets were randomly shuffled after every epoch. Since T4 body segmentation task needed much more precise training than heart segmentation due to smaller ROI, the Adaptive Moment Estimation (Adam) optimizer with initial learning rate 0.0001 was used as an optimization algorithm. Batch size, callbacks, evaluation metrics were same as heart segmentation network. For loss function of T4 body segmentation network, we set α as 0.2 to strongly penalize FPs. γ was set to 4/3, which was chosen from paper [18]. In T4 body segmentation task, we did not used boundary loss since it decreased segmentation accuracy. Comparison between heart segmentation and T4 body segmentation network is described in Table 1.

Type	Heart Segmentation Network	T4 Body Segmentation Network
Network Architecture	Improved Attention U-Net [18]	
Data Split	Training 800, Validation 100, Test 100	
Input Dimension	(256, 256, 1)	
Output Dimension	(256, 256, 1)	
Hyperparameter α for Tversky Loss	0.7 (Penalize False Negative)	0.2 (Penalize False Positive)
Hyperparameter γ for Focal Tversky Loss	4/3 (Focus on data with higher loss)	
Intermediate Layer Loss Function	Focal Tversky Loss	
Final Layer Loss Function	0.95 * Tversky Loss + 0.05 * Boundary Los	Tversky Loss
Optimizer	SGD (Learning rate 0.1, Momentum 0.9)	Adam (Learning rate 0.0001)
Batch Size	16	
Callbacks	Reduce Learning Rate on Plateau (factor = 0.1, Monitor validation loss) Early Stopping (Stop at epoch with minimal validation loss)	
Evaluation Metrics	Dice score coefficient	

Table 1. Description of segmentation network for heart and T4.

2.3 Volumetric Measurement of Heart

2.3.1 Analysis of Binary Mask

For analysis of heart volume adjusted by T4 body length, we measured area of heart, vertical length of heart and T4 body length of binary mask (ground truth mask and predicted mask from DL model). Using opencv library [22] in Python, we measured 1) area of heart, 2) vertical length of heart and 3) drawed minimal bounding box of T4 body and measured width of bounding box (Figure 7).

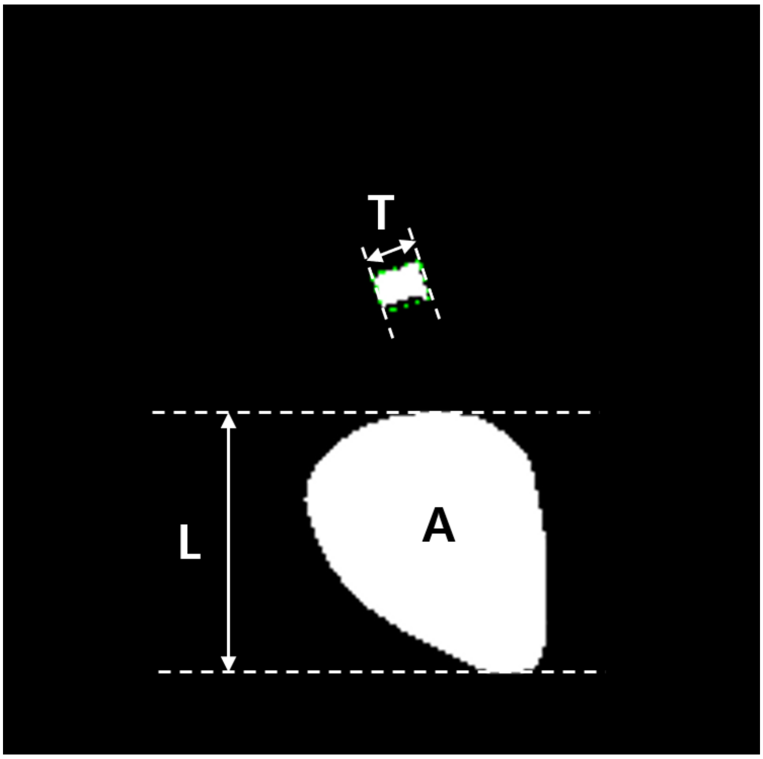


Figure 7. Analysis of binary mask. Area (A) and vertical length (L) of heart is measured, and width of bounding box (dashed) is measured as length of T4 body (T). Vertical length of heart (L) was calculated as difference between maximum y point and minimum y point of heart region. T4 body length (T) was measured as width of minimum area bounding box of T4 body.

2.3.2 Vertebra-adjusted Heart Volume (VaHV)

To calculate volume of heart using area of heart, we applied single plane volume calculation method of modified Simpson's rule [25]. The equation used to calculated heart volume is as follows :

$$\text{Volume} = (\text{Area of Heart})^2/(\text{Vertical Length of Heart})$$

Since scale of radiographs is not fixed, this volumetric measurement needs to be adjusted. Earlier study have shown that T4 body length has good linear correlation with short and long axis of heart [4]. To estimate relatioship between T4 body length and calculated volume, we used allometric equation which was proposed in study to scale Left Ventricular End Diastolic Diameter (LVID_d) with body weight [26]. Allometric equaion $Y = aX^b$ was used to fit a straight line to the relationship and identify the correct exponent for T4 body length. Y represents calculated volume of heart, and X represents T4 body length. For correct estimation, area of heart, vertical length of heart and T4 body lengths from ground truth mask (from 1,000 images) were used. With logarithmic form of the allometric equation ($\log Y = \log a + b\ \log X$), b value of 0.75 showed highest statistical significance (P-val < 0.001). So we defined new method to calculate vertebra-adjusted heart volume (VaHV):

$$\text{VaHV} = [(\text{Area of Heart})^2/(\text{Vertical Length of Heart})]/(\text{T4 Body Length})^{0.75}$$

Description of VaHV calculation is shown in Figure 8.

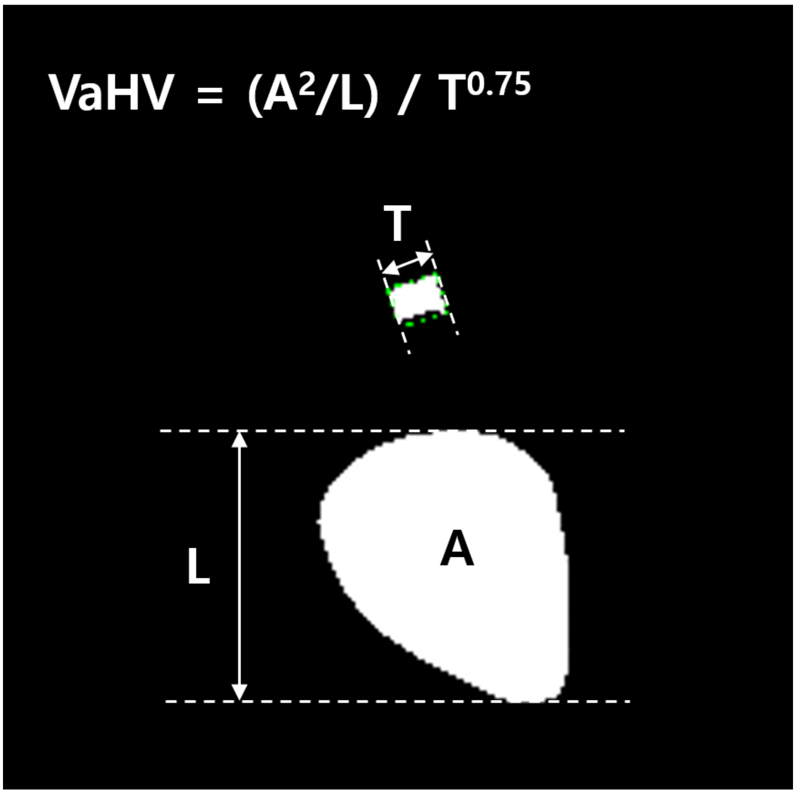


Figure 8. VaHV calculation method. A = area of heart, L = vertical length of heart, and T = length of T4 body.

2.3.3 Calculation of VaHV from DL Model Prediction

After training, the output of segmentation DL model was a score map between 0 and 1 which represent the probability that a given pixel belongs to heart or T4. We segmented heart and T4 body applying threshold at 0.5. Using method described in 2.3.1, we measured area of heart, vertical length of heart and T4 body length from predicted segmentation mask made by trained DL model. Then we calculated VaHV of 100 test images.

2.4 Statistical Analysis

2.4.1 Segmentation DL Model Performance

Statistical analysis was performed using R software (Vienna, Austria) and Python in test set of 100 radiographic images. Dice score coefficient, precision, recall between predicted mask (segmented by applying threshold of 0.5) and ground truth mask were obtained. After calculating VaHV on predicted mask, summary descriptive indexes (mean, standard deviation) and distribution of VaHV were obtained.

2.4.2 Correlation between VaHV and VHS

To compare VaHV and VHS, we investigated correlation between VaHV and VHS of same image. Pearson's correlation (r), Scatter plot were used to assess correlation. α level for determination of statistical significance was 0.05.

2.4.3 Evaluation of Cardiomegaly using Echocardiographic Measurement

To evaluate clinical utility of VaHV in detecting cardiomegaly, we collected 188 studies that had concurrent (within 1 month) radiographic examination and echocardiographic examination. VaHV of 188 images were calculated using trained DL model. LA/Ao ratio and LVIDDN were used as indicator of left atrial enlargement and left ventricular enlargement, as guided in ACVIM consensus guideline [8]. Using 1.6 and 1.7 for cutoff of LA/Ao and LVIDDN, student's t test result was obtained. To measure accuracy of VaHV and VHS in determining left atrial and ventricular enlargement ($LA/Ao \geq 1.6$ and $LVIDDN \geq 1.7$), Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) was obtained.

3. Results

3.1 DL Model

3.1.1 Heart Segmentation

With early stopping callback, training of heart segmentation model stopped at 20 epochs. Loss (combination of tversky loss, focal tversky loss and boundary loss) curve and dice score coefficient curve is shown in Figure 9. By applying segmentation threshold of 0.5, dice coefficient score between predicted mask and ground truth mask was 0.956, recall was 0.978 and precision was 0.936. Results of predicted mask from traind DL model is shown in Figure 10.

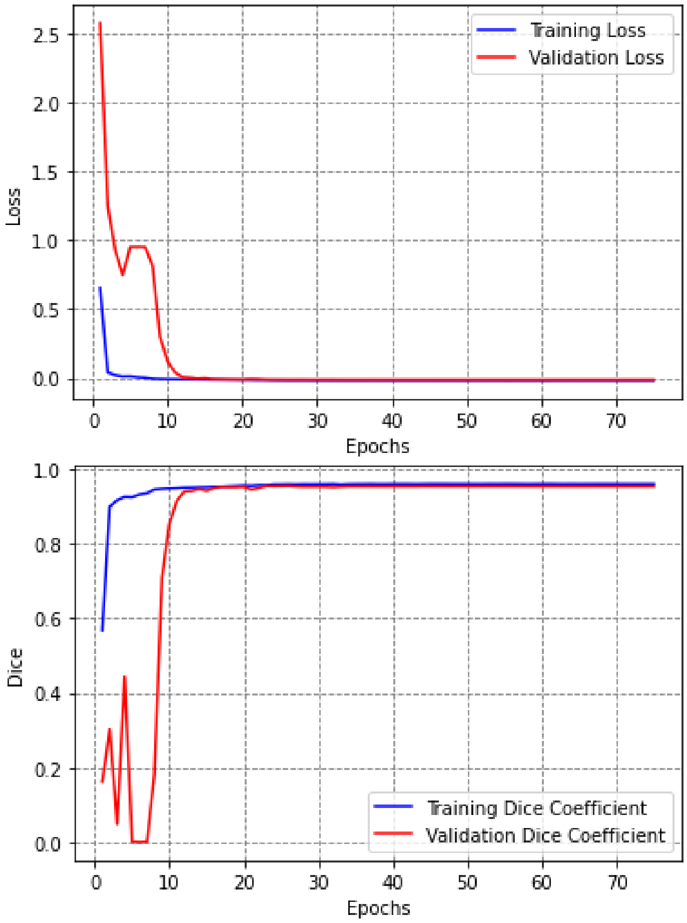
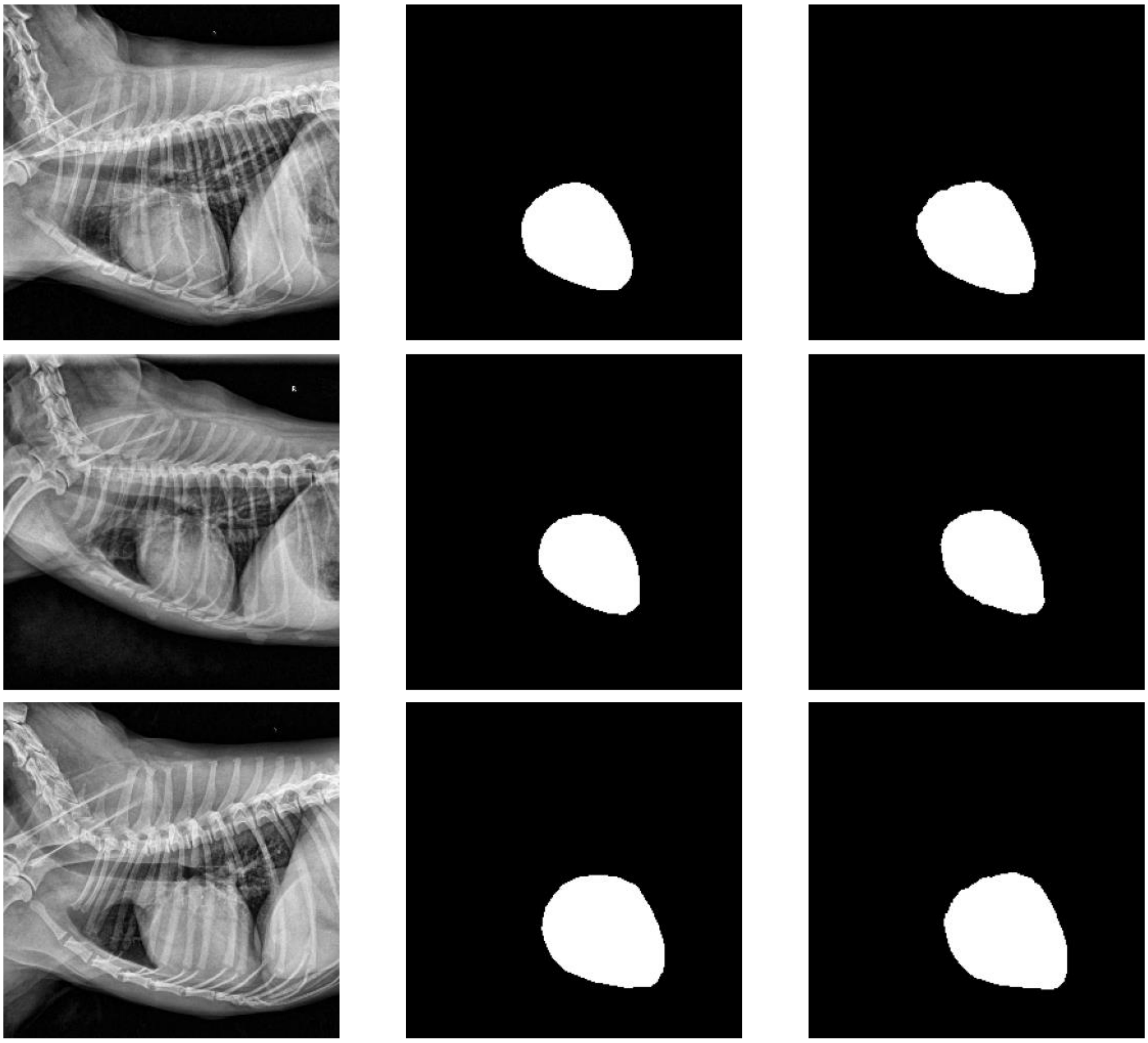


Figure 9. Loss (top) and Dice score coefficient (bottom) curve in training history of heart segmentation DL model.



(a) Input Image (b) Ground Truth of heart (c) Predicted mask of heart

Figure 10. Examples of input image (left column), ground truth mask (middle column), predicted mask of heart (right column).

3.1.2 T4 Body Segmentation

With early stopping callback, training of T4 segmentation model was stopped at 60 epochs. Loss (combination of tversky loss and focal tversky loss) curve and dice score coefficient curve is shown in Figure 11. By applying segmentation threshold of 0.5, dice coefficient score between predicted mask and ground truth mask was 0.844, recall was 0.812 and precision was 0.905. Results of predicted mask from traind DL model is shown in Figure 12.

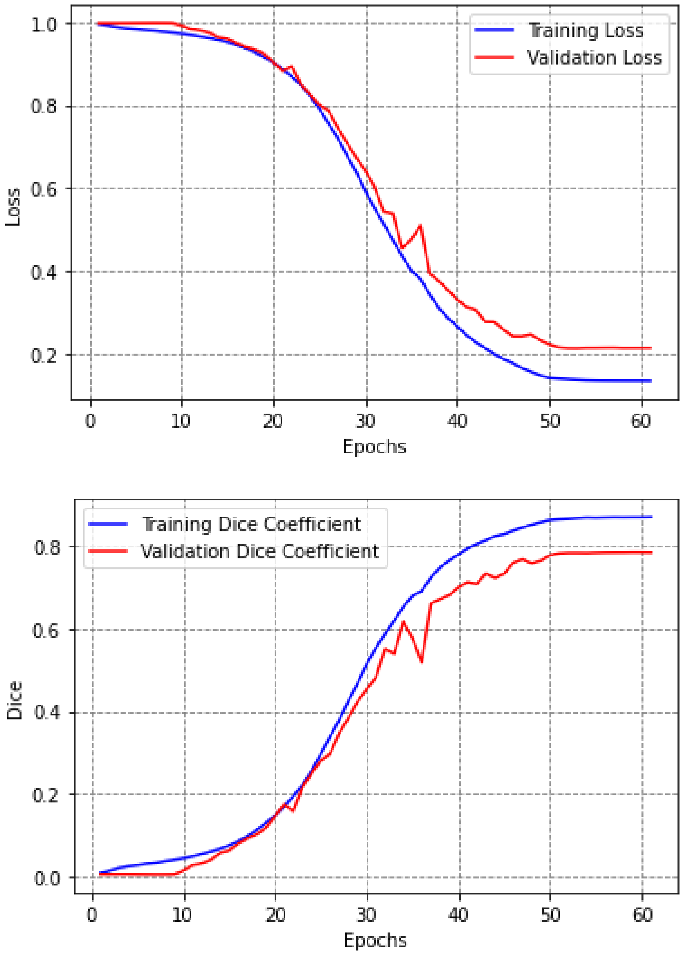
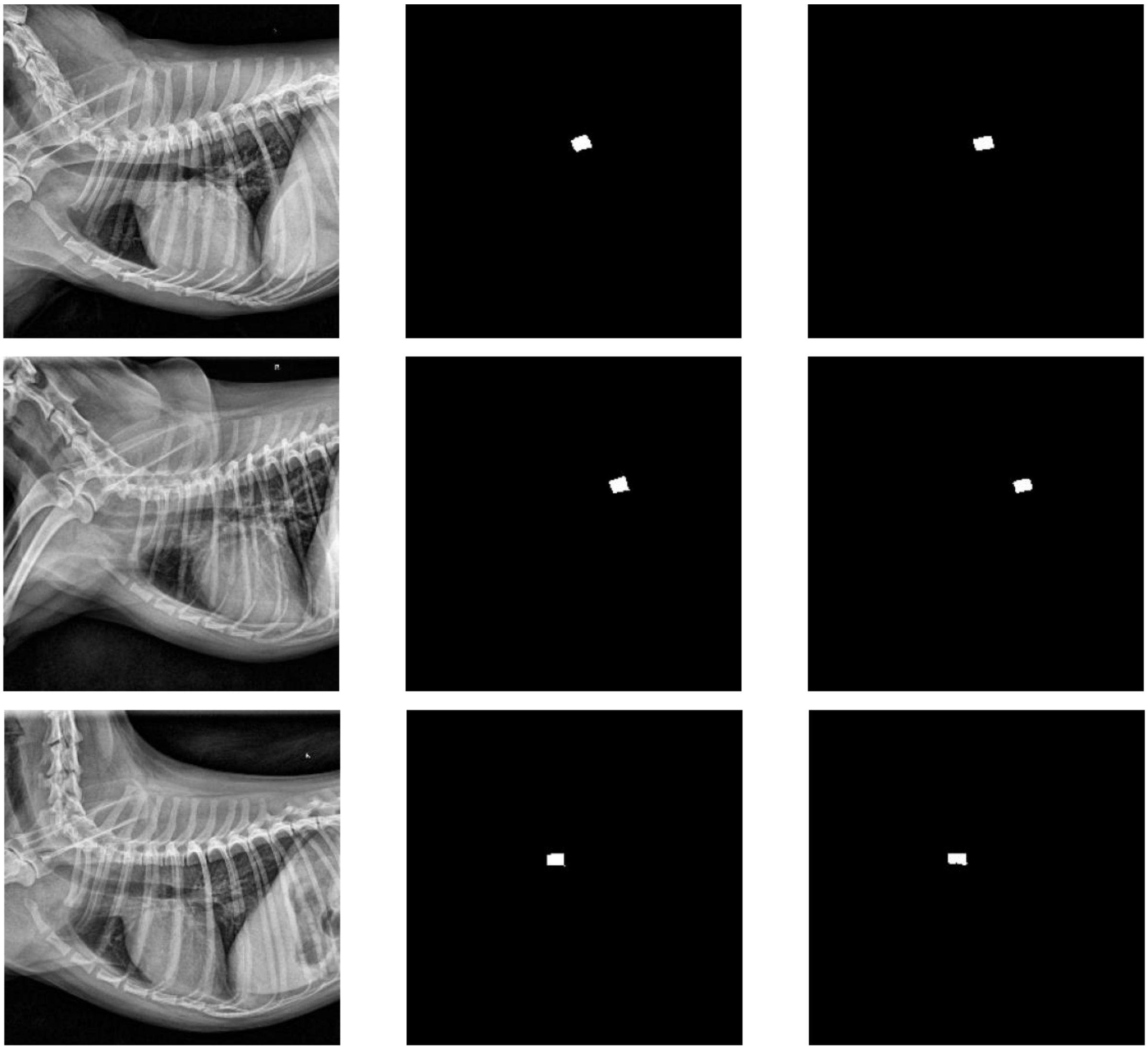


Figure 11. Loss (top) and Dice score coefficient (bottom) curve in training history of T4 segmentation DL model.



(a) Input Image (b) Ground Truth of T4 body (c) Predicted mask of T4 body

Figure 12. Examples of input image (left column), ground truth mask (middle column), predicted mask for T4 body (right column).

3.2 Descriptive Statistics of VaHV

In test set of 100 radiographic images, we evaluated VaHV using predicted mask from trained DL model. Histogram of test set VaHV is shown in Figure 13. Summary statistics of 100 VaHV are described in Table 2.

Mean	Standard Deviation	Max Value	Min Value
50741.8	16931.43	127972.1	23174.72

Table 2. Descriptive statistics of 100 VaHV obtained from DL Model prediction in test set

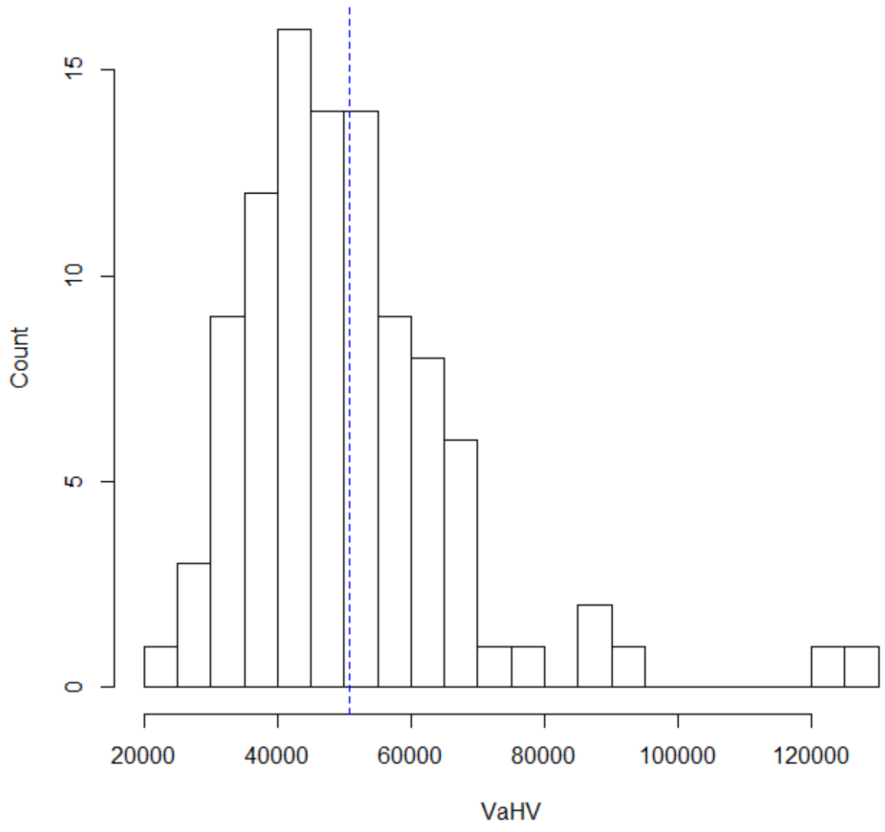


Figure 13. Histogram of VaHV evaluated in test set of 100 radiographic images. Blue dashed line indicates mean value.

3.3 Correlation between VaHV and VHS

To investigate correlation between VaHV and VHS, we performed spearman's correlation test in 100 test set images. Scatter plot of VaHV and VHS, and spearman's correlation coefficient is shown in Figure 14. A positive correlation (r = 0.69, *P* < 0.001) was observed between VaHV and VHS.

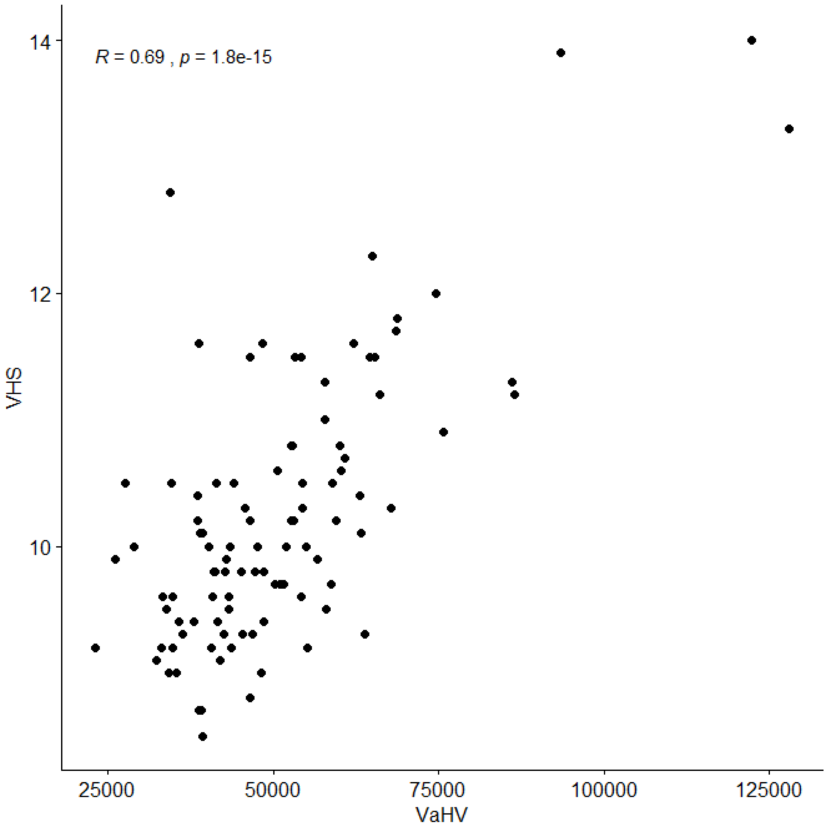


Figure 14. Scatter plot between VaHV and VHS values from 100 test set. Spearman's correlation coefficient and corresponding *P*-value is shown (top left).

3.4 Diagnostic Performance of VaHV for Detecting Cardiomegaly

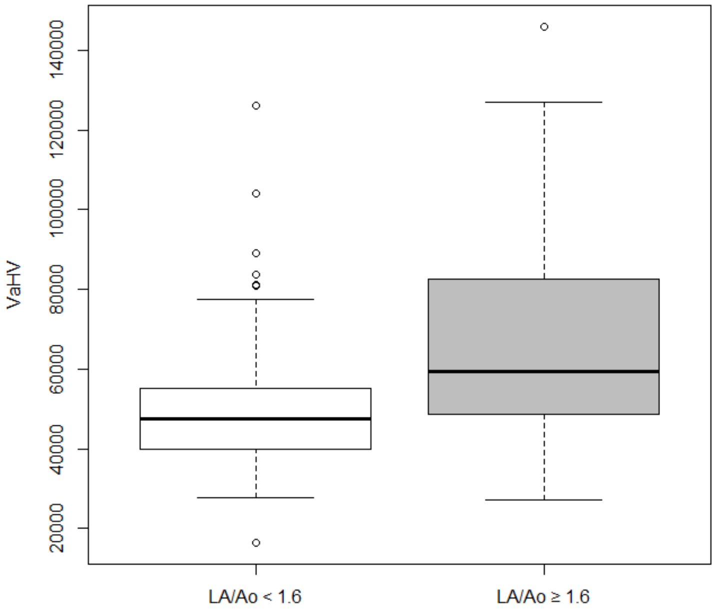
Using 1.6 and 1.7 as cutoff for LA/Ao and LVIDDN, we performed student's t-test for VaHV of 188 patients with echocardiographic record. I compared VaHV values by 3 criteria :

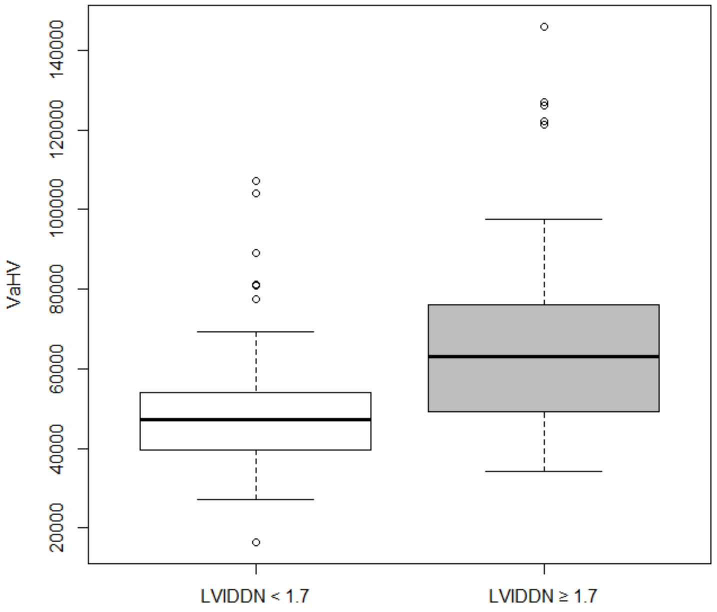
Criterion 1) LA/Ao < 1.6 group vs. LA/Ao ≥ 1.6 group,

Criterion 2) LVIDDN < 1.7 group vs. LVIDDN ≥ 1.7 group,

Criterion 3) LA/Ao < 1.6 or LVIDDN < 1.7 group vs. LA/Ao ≥ 1.6 and LVIDDN ≥ 1.7 group.

VaHV showed greater values for all 3 criteria, with *P*-value of 2.838 x 10^{-5}, 2.91 x 10^{-7} and 7.039 x 10^{-6}. Box plots for VaHV in all 3 criteria are shown in Figure 15.





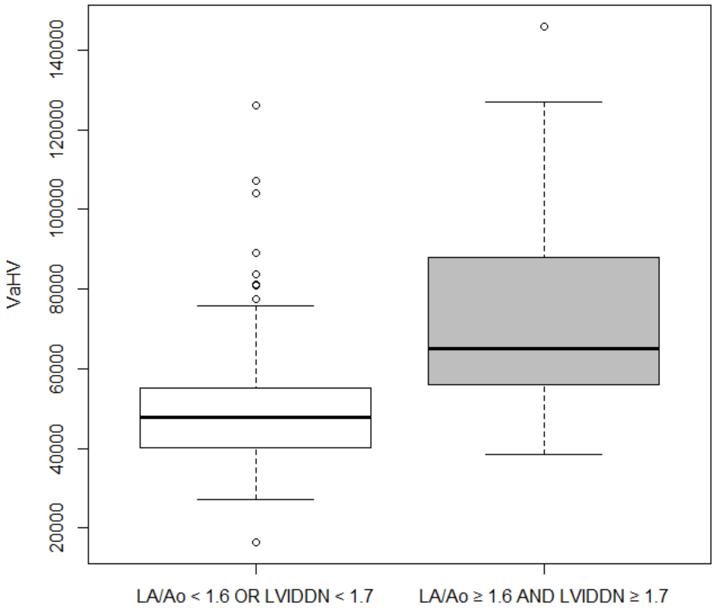
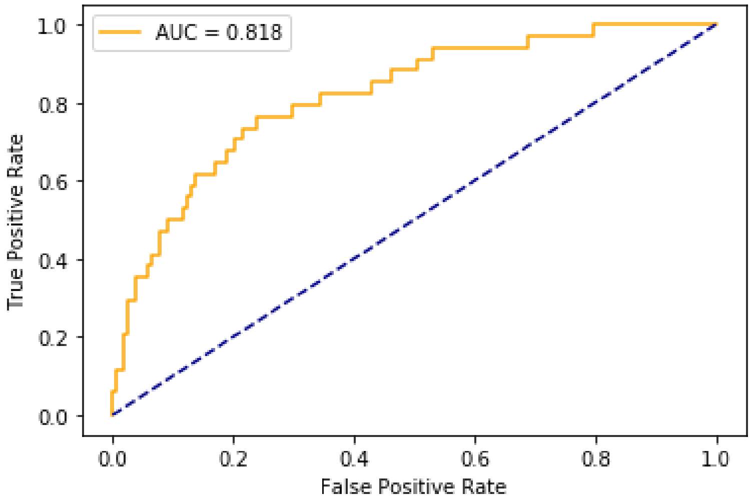


Figure 15. Boxplots comparing VaHV values for groups divided by absence or presence of left atrial enlargement (defined as LA/Ao < 1.6 and LA/Ao ≥ 1.6, top), left ventricular enlargement (defined as LVIDDN < 1.7 and LVIDDN ≥ 1.7, middle) and both left atrial and ventricular enlargement (defined as LA/Ao < 1.6 or LVIDDN < 1.7 and LA/Ao ≥ 1.6 and LVIDDN ≥ 1.7, bottom).

Receiver Operating Curve (ROC) and area under the curve (AUC) of VaHV and VHS in classifying left atrial and ventricular enlargement (LA/Ao ≥ 1.6 and LVIDDN ≥ 1.7) is shown in figure 16. The AUC and confidence limits for VaHV and VHS were respectively 0.818 and 0.805

a. VaHV



b. VHS

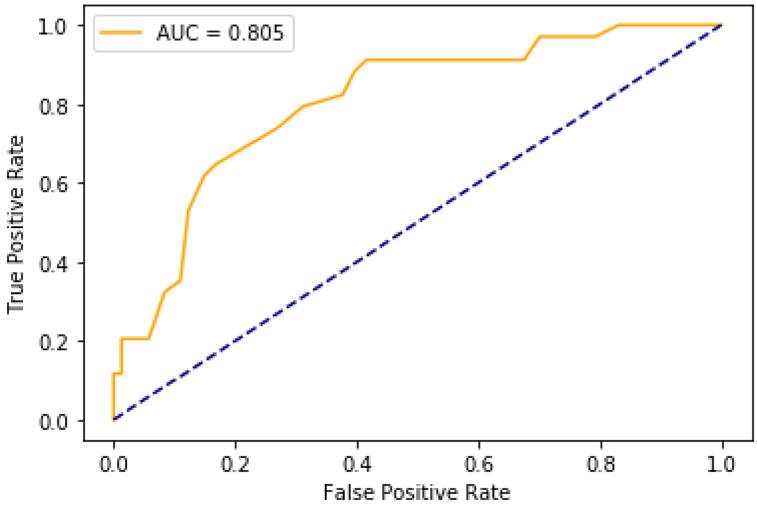


Figure 16. ROC curves and AUC of VaHV (a), VHS (b) in detecting Left atrial and ventricular enlargement, defined as LA/Ao over 1.6 and LVIDDN over 1.7.

4. Dicsussion

In this study we presented 1) deep learning for segmentation of heart and 4th thoracic vertebrae (T4) body, and 2) a novel objective measurement (VaHV) to estimate heart volume in digital radiograph of dog.

Including VHS, there are several more studies (e.g., Vertebral Left Atrial Size (VLAS) [27], RLAD [28]) that utilize 1-dimensional length in radiographic image to detect cardiomegaly. But to date there is no study that used 2-dimensional heart area measurement to assess volume of heart. To measure volumetrics measurement of heart, we used semantic segmentation by applying fully convolutional network to thoracic radiographs. Since heart and T4 body have highly unbalanced ROI (regions of interest) size, we trained two improved Attention U-Net [18] with different hyperparameters to segment heart and T4 body separately. Chosen parameters effectively improved dice score coefficient, precision and recall of DL networks. The deep learning model output showed very good agreement with ground truth mask (Dice score coefficient : heart = 0.956, T4 body = 0.844).

With measured heart area and T4 body length, we developed new method to calculate volume of heart and adjust it by exponentiated T4 body length, and named this novel volumetric measurement vertebra-adjusted heart volume (VaHV). Calculated VaHV showed statistically significant correlation with VHS ($r = 0.69$, $P < 0.001$), indicating that it realistically reflects size of heart. VaHV also showed great performance in detecting left atrial enlargement (LA/Ao over 1.6) and left ventricular enlargement (LVIDDN over 1.7), which are official indicators of left heart enlargement in staging myxomatous mitral valve disease (MMVD) in dogs. The usefulness of VaHV was shown by high AUC value (0.818) in classification of left heart enlargement (LA/Ao over 1.6 and LVIDDN over 1.7). AUC of VaHV was higher than that of VHS. VaHV showed potential usefulness to assess and monitor cardiomegaly. Combined with gold standard methods (VHS and echocardiography), VaHV may help veterinary clinicians diagnose cardiac enlargement.

This study showed the potential of deep learning as a tool for cardiomegaly assessment in clinical practice. This study was the first attempt to apply semantic segmentation to radiographic images in veterinary medicine. With more advanced deep learning architectures applied to veterinary images in further studies, field of veterinary artificial intelligence will make continuous development. Furthermore, methodologies used in this study can be expanded to assess diverse cardiac diseases in chest radiographs of human.

This study has some limitations. We only used 1,000 cases to train, validate and test deep learning model. Although U-Net reach great performance with less than 500 datasets [13], larger number of data will improve deep learning model and make more robust model by preventing overfitting. In addition, since VHS, LA/Ao and LVIDDN values which were used to evaluate accuracy and diagnostic performance of VaHV were reported by many radiologists in SNU-VMTH, there may be biases per radiologist.

5. Conclusion

To asssess volumtric measurement of heart in lateral thoracic radiographic images of dogs, improved attention U-Net with different hyperparameters were trained to segment heart and T4 body. The new radiographic measurement named VaHV was calculated by using area of heart, vertical length of heart and T4 body length. VaHV showed great performance in detecting left atrial and ventricular enlargement. VaHV, volumetric measurement of heart obtained from deep learning would provide clinicians a effective tool for detection and monitoring of cardiomegaly in dogs.

6. References

[1] McComb, B. L., Chung, J. H., Crabtree, T. D., Heitkamp, D. E., Iannettoni, M. D., Jokerst, C., ... & Ravenel, J. G. (2016). ACR appropriateness criteria® routine chest radiography. Journal of Thoracic Imaging, 31(2), W13-W15.

[2] Speets, A. M., van der Graaf, Y., Hoes, A. W., Kalmijn, S., Sachs, A. P., Rutten, M. J., ... & Mali, W. P. (2006). Chest radiography in general practice: indications, diagnostic yield and consequences for patient management. British Journal of General Practice, 56(529), 574-578.

[3] Li, Z., Hou, Z., Chen, C., Hao, Z., An, Y., Liang, S., & Lu, B. (2019). Automatic cardiothoracic ratio calculation with deep learning. IEEE Access, 7, 37749-37756.

[4] Buchanan, J. W., & Bücheler, J. (1995). Vertebral scale system to measure canine heart size in radiographs. JOURNAL-AMERICAN VETERINARY MEDICAL ASSOCIATION, 206, 194-194.

[5] Nakayama, H., Nakayama, T., & Hamlinxya, R. L. (2001). Correlation of cardiac enlargement as assessed by vertebral heart size and echocardiographic and electrocardiographic findings in dogs with evolving cardiomegaly due to rapid ventricular pacing. Journal of veterinary internal medicine, 15(3), 217-221.

[6] Detweiler, D. K., & Patterson, D. F. (1965). The prevalence and types of cardiovascular disease in dogs. Annals of the New York Academy of Sciences, 127(1), 481-516.

[7] Borgarelli, M., Savarino, P., Crosara, S., Santilli, R. A., Chiavegato, D., Poggi, M., ... & Tarducci, A. (2008). Survival characteristics and prognostic variables of dogs with mitral regurgitation attributable to myxomatous valve disease. Journal of veterinary internal medicine, 22(1), 120-128.

[8] Keene, B. W., Atkins, C. E., Bonagura, J. D., Fox, P. R., Häggström, J., Fuentes, V. L., ... & Uechi, M. (2019). ACVIM consensus guidelines for the diagnosis and treatment of myxomatous mitral valve disease in dogs. Journal of veterinary internal medicine, 33(3), 1227-1240.

[9] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.

[10] Li, Z., Hou, Z., Chen, C., Hao, Z., An, Y., Liang, S., & Lu, B. (2019). Automatic cardiothoracic ratio calculation with deep learning. IEEE Access, 7, 37749-37756.

[11] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[12] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[13] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[14] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[15] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Glocker, B. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

[16] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).

[17] https://towardsdatascience.com/a-detailed-explanation-of-the-attention-u-net-b371a5590831

[18] Abraham, N., & Khan, N. M. (2019, April). A novel focal tversky loss function with improved attention u-net for lesion segmentation. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (pp. 683-687). IEEE.

[19] Ma, J. (2020). Segmentation Loss Odyssey. arXiv preprint arXiv:2005.13449.

[20] Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., & Ayed, I. B. (2019, May). Boundary loss for highly unbalanced segmentation. In International conference on medical imaging with deep learning (pp. 285-296). PMLR.

[21] Pizer, S. M. (1986). Psychovisual issues in the display of medical images. In Pictorial information systems in medicine (pp. 211-233). Springer, Berlin, Heidelberg.

[22] OpenCV. (2015). Open Source Computer Vision Library.

[23] Kentaro Wada. 2016. labelme: Image Polygonal Annotation with Python.

[24] Chollet, F. & others, 2015. Keras. Available at: https://github.com/fchollet/keras

[25] Schiller, N. B., Shah, P. M., Crawford, M., DeMaria, A., Devereux, R., Feigenbaum, H., ... & Silverman, N. H. (1989). Recommendations for quantitation of the left ventricle by two-dimensional echocardiography. Journal of the American Society of Echocardiography, 2(5), 358-367.

[26] Cornell, C. C., Kittleson, M. D., Torre, P. D., Häggström, J., Lombard, C. W., Pedersen, H. D., ... & Wey, A. (2004). Allometric scaling of M-mode cardiac measurements in normal adult dogs. Journal of veterinary internal medicine, 18(3), 311-321.

[27] Malcolm, E. L., Visser, L. C., Phillips, K. L., & Johnson, L. R. (2018). Diagnostic value of vertebral left atrial size as determined from thoracic radiographs for assessment of left atrial size in dogs with myxomatous mitral valve disease. Journal of the American Veterinary Medical Association, 253(8), 1038-1045.

[28] Salguero, X. S., Prandi, D., Llabrés-Díaz, F., Manzanilla, E. G., & Bussadori, C. (2018). A radiographic measurement of left atrial size in dogs. Irish veterinary journal, 71(1), 25.

초록

개의 심장질환 중 가장 높은 유병률을 나타내는 이첨판 폐쇄부전증을 포함하여 다양한 심장질환이 점진적인 심비대를 특징으로 하기에, 개의 흉부 방사선 영상에서 심장 크기를 측정하여 심비대를 진단하는 것은 심장질환을 조기에 발견하고 적절한 치료시기를 계획하는 데 있어 매우 중요한 부분을 차지한다. 현장에서 바로 잴 수 있는 지표로서 기존에는 vertebral heart score (VHS)가 널리 사용되고 있으나, 이는 1 차원 길이의 합으로 이루어진 지표이기에 심비대를 진단하는 데 한계가 있을 수 있다. 본 연구의 목적은 개의 흉부 방사선 영상에서 심장 면적과 척추체 길이를 자동으로 산출하는 딥러닝 모델을 구축하고, 이를 이용하여 심장 용적을 추정할 수 있는 지표를 개발하는 것이었다.

본 연구는 서울대학교 수의과대학 동물병원 검진자료로부터 수집된 총 1,188 건의 자료를 바탕으로 수행되었다. 1,000 건의 영상은 심장과 척추체의 면적을 자동으로 분할 (semantic segmentation) 해주는 딥러닝 모델을 훈련시키고 평가하기 위해 사용되었으며, 이를 이용하여 새로운 심장 용적 지표인 vertebra-adjusted heart volume (VaHV) 를 산출했다.

추가로 1 달 미만 간격의 방사선 촬영 기록과 심장초음파 검진 기록을 가진 188 건의 영상을 수집하여 훈련된 딥러닝 모델을 이용해 계산한 VaHV 와 심장초음파 기록 (LA/Ao, LVIDDN) 을 비교하여 VaHV 의 심비대 진단능을 평가하였다.

심장과 척추체의 면적 불균형을 보완하기 위해 서로 다른 hyperparameter 를 가진 Improved Attention U-Net 이 사용되었으며, 두 개의 신경망 모두 시험용 데이터셋에서 정답 면적과 높은 일치율 (dice score coefficient 0.956, 0.844) 를 보였으며, 신경망의 예측결과에서 계산된 VaHV 는 기존에 기록된 VHS 와 통계적으로 유의한 상관계수를 (r = 0.69, $P < 0.001$) 보여 VaHV 가 실제로 심장 면적을 잘

대변함을 확인하였다. 또한 188 건의 자료에서 VaHV 가 좌심비대 (LA/Ao > 1.6, LVIDDN > 1.7) 에 대해 높은 예측력을 가짐을 확인하였으며 (AUC 0.818), 기존에 사용되던 VHS 의 예측력 (AUC 0.805) 보다 우수한 성능을 보임을 확인하였다.

본 연구는 수의방사선에서 최초로 딥러닝을 이용한 의미론적 면적 분할 (semantic segmentation) 을 적용하여 수의 영상에서 기존보다 더 다양한 신경망 알고리즘이 활용될 수 있는 가능성을 보여주었다. 또한 심장의 2 차원 면적이 심비대를 진단함에 있어 기존의 길이 기반 심장 크기 측정 지표를 보완할 수 있다는 가능성을 보여주었다.

주요어 : 딥러닝, 의미론적 분할, 수의 방사선 영상, 심장 면적 분석

학번 : 2019-24084