Learning Stethoscope Placement for Heartbeat Detection Using Convolutional Neural Networks

Bryan Tiang
Stanford University
tiang@stanford.edu

Cristian Galdeano
Stanford University
galdeano@stanford.edu

Abstract

We propose a vision-based deep learning framework to automate stethoscope placement for cardiac auscultation. Accurate positioning is essential for capturing clean heart sounds, yet traditionally requires clinical expertise. Our approach leverages convolutional neural networks to localize the optimal placement point directly from a 2D RGB image of a person in a frontal pose.

We construct a custom dataset of annotated torso images and formulate the task as a dense prediction problem using a Gaussian affordance map centered on the ground truth location. As a baseline, we implement a standard U-Net architecture and compare it against several variants, including MultiResUnet and our own modified U-Net with residual connections, group normalization, and reduced depth.

Our model achieves the lowest localization error of 9.22 pixels on a fixed test set, outperforming both the baseline U-Net and MultiResUnet, while requiring fewer parameters. We present both quantitative and qualitative results, as well as training loss curves to support the model's convergence and generalization.

1. Introduction

In this project, we aim to use a convolutional neural network (CNN) to determine the correct placement of a stethoscope on the human body for measuring heartbeats. Accurate stethoscope positioning is critical for obtaining clean heart sound signals (auscultation), yet remains a manual and expertise-driven process. While prior work has explored stethoscope placement using rule-based models, 3D sensing, or even acoustic feedback, our project focuses on learning the optimal location directly from a 2D image using a CNN. This allows us to simplify the setup while aiming for accurate, fully automated placement. Some systems have used tele-operated robots to perform auscultation tasks [1, 9], but these still require a human operator to guide the process remotely. In contrast, our fully automated approach aims to reduce interaction time and eliminate the need for clinician supervision. We demonstrate that this lightweight, vision-only setup can achieve accurate localization while remaining efficient and easy to deploy in robotic systems.

As the use of robots in clinical settings continues to grow, automating routine tasks, such as measuring vital signs, becomes increasingly valuable. This project explores a vision-based deep learning approach to bridge the gap between robotic perception and physical medical examination. To achieve this, we use our own modified version of the widely adopted U-Net architecture [6], which consists of a contracting path that captures contextual features and a symmetric expanding path that enables precise localization. For training, we constructed our own dataset by collecting images of a diverse set of people from online sources and manually annotating the correct stethoscope placement, which serves as ground truth during training.

The input to our model is a single RGB image of a person facing the camera. The output is an affordance map, where each pixel is assigned a probability indicating how appropriate that location is for stethoscope placement. The final predicted location is chosen as the pixel with the highest probability value. We use a CNN based on the U-Net architecture to learn this mapping. As a baseline, we implement the original U-Net model described in [6] without modifications. Our proposed model includes several architectural changes that considerably improves the performance for our specific application.

In this work we show that the models we train on the dataset we collected are able to obtain a reasonable predicted location for placing a stethoscope regardless of the orientation of the image. Our architecture obtained the best performance out of the five models we experimented with.

2. Related work

In addition to the U-Net paper, a relevant work in this space is ARSteth [3], which assists home users in placing a stethoscope using augmented reality and acoustic feedback. Their system uses an analytical method to estimate

auscultation points by detecting shoulder landmarks and computing placement based on anatomical ratios. While their approach involves user interaction and multimodal guidance, our method focuses on fully automated, vision-based stethoscope placement, making it more suitable for autonomous robotic applications in clinical settings.

Similarly, [10] propose a system for autonomous robotic auscultation of heart and lung sounds. Their method involves capturing a 3D point cloud of the patient, registering a human body model, and estimating anatomical landmarks to guide stethoscope placement. While their system also uses audio feedback and Bayesian optimization to refine placement, our approach focuses on a simpler and more lightweight setup by using a single 2D RGB image as input. This makes our method more accessible and easier to integrate into vision-based robotic systems without requiring 3D scanning infrastructure.

Another relevant extension of U-Net is MultiResUNet [4], which introduces multi-resolution convolutional blocks and residual connections to improve feature extraction at different scales. Their approach enhances segmentation performance in complex biomedical images where fine localization is critical. Inspired by this work, we experimented with architectural modifications to better adapt U-Net for predicting precise stethoscope placement in RGB images of the human body. Together, these works highlight the growing use of deep learning for medical tasks and help motivate our lightweight, image-only approach to automated stethoscope placement.

3. Dataset

Our dataset consists of cropped full-body and torso-only images of individuals standing in an upright, frontal pose. Most images were sourced from online clothing stores, where models typically appear in consistent front-facing positions, which is ideal for our application. All images were resized to 128×128 pixels for consistency and to reduce computational cost. We manually annotated each image by marking the correct stethoscope placement location using OpenCV, producing pixel-level ground truth labels for training. Currently, the size of our dataset is 2450 images. We also included some negative samples where the correct stethoscope placement location is not present within the image. Figure 1 shows some examples of images in the dataset.

(a) (b) (c)

Figure 1. Examples of images from the collected dataset. (a) Closeup image of a torso. (b) Image of entire person. (c) Image of a negative example.

We split the dataset into 80% for training and 20% for validation. To increase data diversity and quantity, each training image is augmented five times using random rotations, and each validation image is augmented twice. This results in a total of 11760 images for training and 1470 images for validation after augmentation.

The test set that we use consists of 58 images. Some were intentionally chosen to differ significantly from the training data, with cluttered background, more complex poses, or cropped and rotated views. A few test samples are randomly rotated variants of images from the training set. Figure 2 shows some examples of these images.

(a) (b) (c)

Figure 2. Examples from the test dataset. (a) Original image. (b) Rotated variant of image (a). (c) Another example.

4. Methods

Directly training a model to predict the correct pixel location can be challenging due to the sparsity of the reward signal. To address this, we formulated the task as a pseudo segmentation problem by generating an affordance map. Specifically, we overlay a 2D Gaussian centered on the labeled ground-truth location, where each pixel's value represents the probability of it being the correct stethoscope placement. This transforms the task into a dense prediction problem, enabling smoother learning through pixel-wise loss.

We have currently trained 5 different models on the dataset that we have accumulated. The details for the models will be mentioned in the following subsections.

4.1. Baseline U-net model

As a baseline, we implemented the original U-Net architecture [6], which is widely used in biomedical image

segmentation due to its strong performance on localization tasks. Our implementation keeps the core components of the original design, including the symmetric encoder-decoder structure with skip connections, 3x3 convolutions, and ReLU activations. The only architectural change we made was to add padding to the convolutions in both the encoder and decoder paths to preserve spatial dimensions. This change was necessary otherwise the dimensions of the image would not fit the convolutions for the final encoder.

We did not include any normalization layers, residual connections or other enhancements in this baseline model. The output of the network is a 2D affordance map, and we used a binary cross-entropy (BCE) loss to compare the prediction with the ground truth. For training, we used the SGD optimizer with a momentum of 0.99, a learning rate of 0.1 and a batch size of 32. The model was trained over 101 epochs. This baseline provided a clear point of comparison for evaluating the benefits of our architectural modifications.

4.2. MultiResUnet

One type of model that we took ideas for this task was the MultiResUnet [4]. While the architecture has the same overall structure as the U-net architecture, its main difference lies in how it replaces the convolutions in the decoder and encoder paths of the U-net with MultiRes blocks, and uses residual paths for the shortcut paths between the encoder and the decoder. The MultiRes block takes inspiration from Inception blocks that used convolutions of different sizes in parallel to emphasize different scales of detail in the image [8]. The MultiResUnet model that we tested had 5 encoder and decoder paths. One modification that we made was to use Leaky Rectified Linear Unit (LeakyReLU) activation layers instead of ReLU activation layers between the convolution and upconvolution layers since that seemed to give us slightly better validation losses. We found that setting the gradient of the negative portion of the LeakyReLU to 0.1 yielded the best results. We used a composite loss function for this model which is as detailed:

1. **Binary Cross-Entropy Loss:** Applied to all samples for pixel wise affordance prediction.

2. **Conditional Coordinate Regression Loss:** This loss term is only activated for positive samples (samples with a ground truth pixel target). The predicted coordinates are computed via soft-argmax on the output of the model as shown in the equation below:

$$(x, y) = \left(\sum_{i,j} \sigma(\beta \cdot \mathbf{S})_{ij} \cdot x_{ij}, \sum_{i,j} \sigma(\beta \cdot \mathbf{S})_{ij} \cdot y_{ij} \right) \tag{1}$$

   Where $\mathbf{S}$ is the output of the model, β is a temperature term used to control the peak sharpness and σ denoting the softmax function. Both the predicted and ground truth coordinates are normalized by the image height and width. The regression loss is then computed using mean square error (MSE) loss and scaled by a factor of 0.2. A low weighting factor was chosen since higher values were observed to slow down training or cause it to be unstable.

The purpose of the additional term in this loss function is to further refine the positional accuracy of the predicted locations from the model. Note that the subsequent versions of the MultiResUnet models were trained using the same composite loss function.

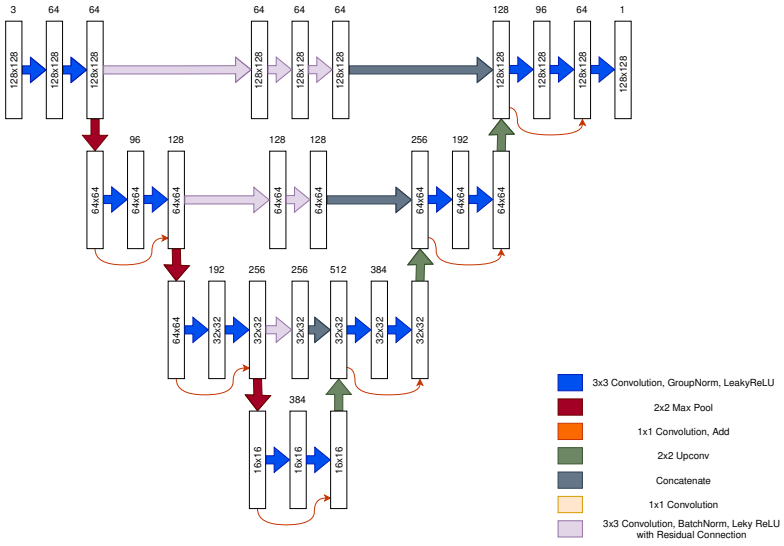

Figure 3. Architecture diagram for our modified U-net

To train this model we used a one cycle learning rate [7] with cosine annealing with a warm up of 30 percent and a max learning rate of 0.001. The model was trained over 101 epochs with a batch size of 32 with the AdamW optimizer [5] set with a weight decay of 0.01.

4.3. MultiResUnet - no ResPaths

This version of the model is very similar to the MultiResUnet model in the previous section. The main difference with this model is that we removed the Res paths in the shortcut. We wanted to test if removing that component of the model would impact its performance. To keep the comparison fair we also used the same learning rate and training parameters as those in section 4.2.

4.4. Small MultiResUnet

This version of the MultiResUnet model is a smaller with 4 encoder and decoder paths instead of 5 and without the Res path shortcuts. We decided to omit the Res path since this results in a smaller model and our initial tests only showed a minimal drop in performance. We wanted to test out a smaller model in order to have a better idea as to how complex the model would have to be for our specific task. Using a smaller model would also mean that the inference time of our model would be much faster, and that would be useful for real time applications especially like those in robotics. As usual, the learning rate and training parameters used were the same as those detailed in section 4.2.

4.5. Our model

Our modified model is a smaller version of the original U-Net model. The overall architecture for this model is shown in Figure 3. We reduced the depth of the U-Net by using three contracting and expanding stages instead of four, as we found that removing the fourth contracting path had no effect on the model's performance. We further modified this model by adding residual paths within the encoder and decoder layers, since they have been shown to help with gradient flow for neural networks [2]. From our initial tests, we observed minor improvements in having the Res paths when testing different versions of the MultiResUnet models, therefore we decided to include those Res paths into the shortcut connections between the encoder and decoder layers for this model as well. LeakyReLU activation layers with a negative gradient of 0.1 were also used for this model. In terms of normalization we found that using group norm gave us the best results. We increased the number of groups in GroupNorm layers in deeper decoder blocks (8, 16, and 32) to normalize finer-grained features more effectively.

For training the model, we used the same learning rate scheduler, optimizer, and parameters mentioned in section 4.2. The composite loss function that was detailed in Section 4.2 was also utilized for this model.

5. Results

We evaluate the models using the mean Euclidean distance between the predicted pixel and the ground truth annotation. This distance is computed in image pixel space and averaged across the test set. A lower mean pixel distance indicates more accurate localization of the stethoscope point. That is, given ground truth pixel coordinates (x_i, y_i) and predicted coordinates $(\hat{x}_i, \hat{y}_i)$ for image i, we define the evaluation metric as:

$$\text{Mean pixel distance} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2} \tag{2}$$

where N is the number of test images.

5.1. Loss Curves

Figures 4 and 5 show the training and validation loss for the baseline U-net and our modified model, respectively. Both models exhibit stable convergence, with validation loss closely following the training loss, which suggests that neither model is overfitting. However, it is difficult to directly compare the performance of the two models by only looking at their validation losses, considering that they are trained using different loss functions as mentioned in section 4.1 and 4.5. Even if both models used the same loss function, we would need a significant gap in validation losses to draw meaningful conclusions. In this case, the loss curves primarily serve to verify that training is proceeding smoothly and that the models are not diverging. See appendix Figure 8, 9, and 10 for the remaining loss curves.

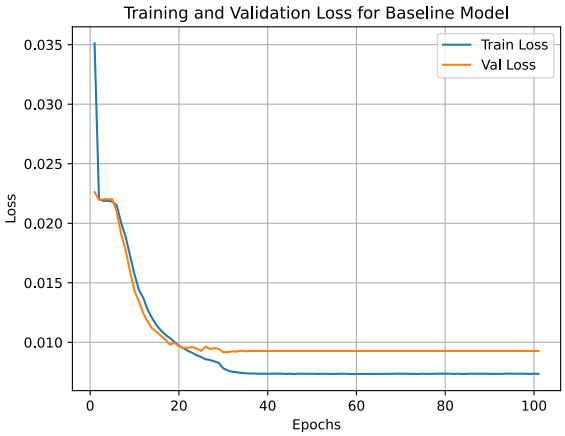

Figure 4. Training and Validation Loss Curves for Baseline model

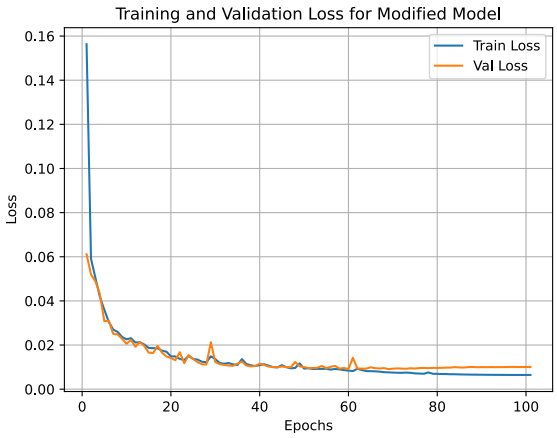

Figure 5. Training and Validation Loss Curves for our modified U-net model

5.2. Qualitative Results

Figures 6 and 7 show qualitative predictions on sample test images. The modified model's predictions are visibly closer to the ground truth location (red dot) compared to those of the baseline model. Some prediction errors occur in cases where the background color or texture is visually similar to the subject's torso, introducing ambiguity. Errors also appear in a few images with partial occlusion of the torso, such as an arm crossing in front, which are not as common in the training data.

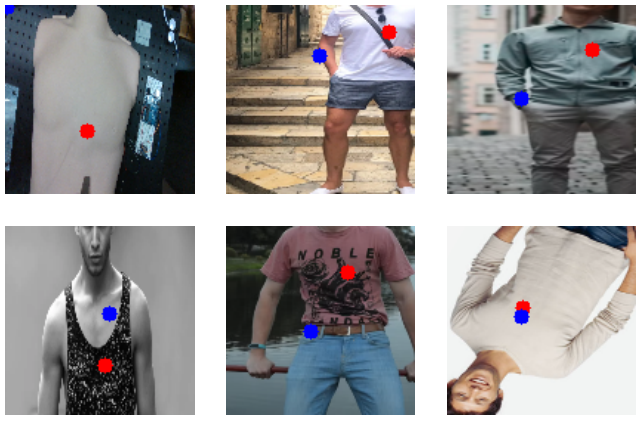

Figure 6. Sample baseline model predictions on test data. Ground truth location is labeled as a red dot and predicted location as a blue dot.

5.3. Quantitative Comparison

We evaluated all models on a fixed test set of 58 images. Table 1 reports the mean pixel distance for each model. We can see that our modified model achieves the lowest error of 9.22 pixels, outperforming both the baseline U-net and the original MultiResUnet models. Interestingly, the smaller version of the MultiResUnet also performs well despite having fewer parameters. This may suggest that architectural efficiency and properly tuned normalization layers could outweigh the importance of model size in this particular task.

Figure 7. Sample predictions from our model on the test dataset. Ground truth location is labeled as a red dot and predicted location as a blue dot

Model	Mean Pixel Distance
Baseline U-Net	15.50
Our model	**9.22**
MultiResUnet	9.73
MultiResUnet - no Res paths	10.80
MultiResUnet smaller	10.03

Table 1. Comparison of model performance based on mean pixel distance between predicted and ground truth stethoscope placement. Lower values indicate better localization accuracy.

5.4. Parameter Efficiency

Table 2 shows the number of parameters in each model. Our model balances performance and parameter count, obtaining superior accuracy with only 9.0 M parameters, which is less than a third of the baseline U-net. It is worth mentioning that the smaller MultiResNet also achieves comparable performance, despite having far fewer parameters than the other models.

Model	Param Count
Baseline U-Net	31.0 M
Our model	9.0 M
MultiResUnet	13.9 M
MultiResUnet - no Res paths	9.8 M
MultiResUnet smaller	2.4 M

Table 2. Model parameter count

6. Conclusion

In this project, we propose a fully automated approach for predicting optimal stethoscope placement using a vision-based deep learning model. By training on RGB images of human torso, we demonstrated that our version of the U-net architecture can outperform both a baseline U-net and a MultiResUnet model in terms of localization accuracy, while using fewer parameters. Our results show that targeted architectural modifications and proper normalization choices can significantly improve performance in this task.

There are several directions for future work. First, expanding the training dataset to include more diverse body types, age groups, and images sources could improve model generalization. Incorporating segmentation of the torso region might also help focus the model's attention on relevant features and reduce potential distraction from background clutter. Additionally, given the promising results from the smaller network, further exploration of lightweight architecture could help optimize the model for real time deployment. Finally, the framework could be extended to predict multiple standard heart auscultation points rather than a single location, aligning more closely with real clinical practice. This could be achieved by modifying the output layer to predict multiple heatmaps, one for each anatomical point, with each point treated as a separate class. Additionally, incorporating lung auscultation points would broaden the system's applicability to respiratory assessments and further increase its clinical utility.

7. Appendix

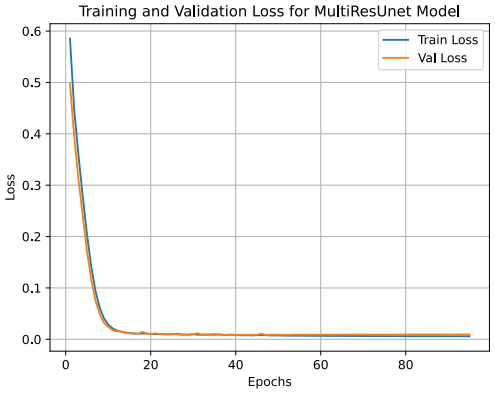

Figure 8. Loss curve for MultiResUnet

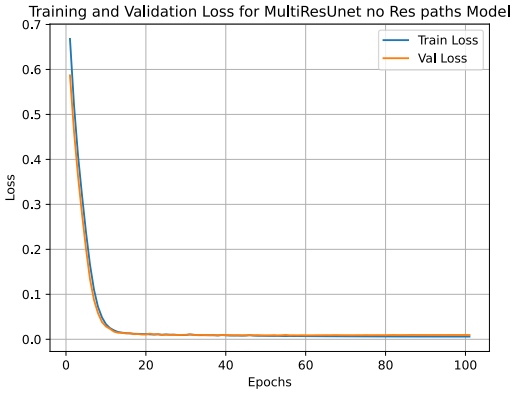

Figure 9. Loss curve for MultiResUnet with no Res paths

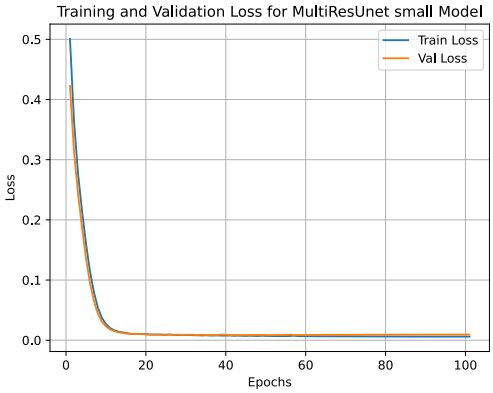

Figure 10. Loss curve for small MultiResUnet

8. Contributions

Bryan Tiang and Cristian Galdeano contributed equally to the project and report. Bryan did data collection, model training and architecture testing. Cristian did model architecture development and analysis, literature review, report writing and editing.

References

[1] M. Giuliani, D. Szczęśniak-Stańczyk, N. Mirnig, G. Stollnberger, M. Szyszko, B. Stańczyk, and M. Tscheligi. User-centred design and evaluation of a tele-operated echocardiography robot. *Health and Technology*, 10:649–665, 2020.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[3] K. Hou, S. Xia, E. Bejerano, J. Wu, and X. Jiang. Arsteth: Enabling home self-screening with ar-assisted intelligent stethoscopes. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*, IPSN '23, page 205–218, New York, NY, USA, 2023. Association for Computing Machinery.

[4] N. Ibtehaz and M. S. Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87, Jan. 2020.

[5] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019.

[6] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[7] L. N. Smith and N. Topin. Super-convergence: Very fast training of neural networks using large learning rates, 2018.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, 2014.

[9] G. Yang, H. Lv, Z. Zhang, L. Yang, J. Deng, S. You, J. Du, and H. Yang. Keep healthcare workers safe: application of teleoperated robot in isolation ward for covid-19 prevention and control. *Chinese Journal of Mechanical Engineering*, 33:1–4, 2020.

[10] Y. Zhu, A. Smith, and K. Hauser. Automated heart and lung auscultation in robotic physical examinations. *CoRR*, abs/2201.09511, 2022.