HiGNN: Hierarchical Left Ventricle Landmark Detection with Graph Neural Networks

Masoud Mokhtari
masoud@ece.ubc.ca
14186167

Mobina Mahdavi
mobina@ece.ubc.ca
91592170

Hooman Vaseli
hoomanv@ece.ubc.ca
54157152

Abstract

The function of the Left Ventricle (LV) chamber of the heart is often examined by measuring certain dimensions in echocardiography (echo) cine series including the LV internal dimension and the thickness of the LV walls. The manual clinical procedure to quantify these measurements involves pinpointing critical pixel locations through visual inspection, which is extremely noisy due to the inherent inter-observer variability. The need to reduce this variability and the emergence of point-of-care ultrasound (POCUS) imaging devices, which are often used by non-expert users, has sparked interest in automatic landmark detection methods. Prior works in this domain are convolution-based where models are trained in a supervised manner with pixel-level labels. We argue that since only a small number of pixels are to be detected in a high-dimensional image, this task can benefit from a hierarchical approach. More specifically, dividing the image into patches of different granularity and performing patch-level landmark detection as an auxiliary task can aid the main pixel-level task. Therefore, in this work, we introduce a hierarchical landmark detection network that relies on the representative power of graph neural networks (GNNs) to enable information propagation between these auxiliary tasks and the main task. For each ultrasound image, we create a single pixel-level grid graph and multiple graphs containing nodes corresponding to patches of different sizes. GNNs are then used to perform node prediction to indicate which patches and pixels include landmarks while enabling communication between the tasks. Across the three different landmark measurements, our model achieves an average mean absolute error of 1.5 mm and a mean percent error rate of 9.9% on a private LV Landmark dataset containing ultrasound images from 23755 patients, which performs on par with state of the art without reliance on sample rejections and any pretraining tasks. This shows that LV landmark detection benefits from this hierarchical approach. Our code is publicly available at: `https://github.com/RCL-LVLD/gnn_lvl`.

1 Introduction

Ultrasound imaging of the heart, also known as echocardiography (echo), is one of the fastest growing imaging modalities used for diagnosis and intervention in clinical and point of care settings due to its inherent safety, affordability, and real-time nature (1). In echo imaging, Parasternal Long Axis (PLAX) view, one of the several standard echo views, is the optimal view to assess the function of Left Ventricle (LV) through measuring certain dimensions such as: Left Ventricle Internal Diameter (LVID), Interventricular Septal (IVS), and Left Ventricular Posterior Wall (LVPW). LVID and wall thicknesses (IVS and LVPW) are then used to calculate LV mass and Relative Wall Thickness (RWT), which help diagnose LV hypertrophy and risk of a stroke (2).

Preprint. Under review.

The aforementioned measurements are done via placing clinical landmarks defining the two ends of a linear distance. This procedure is subject to significant inter-observer variability due to differences in operator experience. Moreover, variations in image quality across patients can majorly affect the placement of clinical landmarks. Therefore, there is great need in echo imaging for the automation of LV measurements from PLAX view, specifically in the context of point-of-care ultrasound (POCUS), where non-expert users are usually involved. Automation of cardiac landmark detection, however, proves to be challenging because the annotations are noisy and are sparsely recorded throughout the echo. For instance, LVID is conventionally labeled only in two time steps.

The automatic approaches proposed to-date mainly use U-Nets(3) and Convolutional Neural Networks (CNNs) as the backbone for their network architecture for either a direct regression of the landmark coordinates (4; 5; 6; 7; 8; 9) or generation of heatmaps localizing the landmarks (10; 11; 12; 13; 14). In this setting, there are no direct means of information propagation among the different landmark locations. Additionally, the model has to detect only a few positive pixels among all the pixels in a high-dimensional image.

In this work, to address these problems, we propose a hierarchical framework based on Graph Neural Networks (GNNs) to detect LV landmarks in ultrasound images. This framework includes a main pixel-level task and multiple patch-level auxiliary tasks with each task containing patches of different granularity. For the main task, the ultrasound image is represented as a grid-graph where each pixel is a node and has connections to its vertical and horizontal neighbour pixels/nodes. For each auxiliary task, a similar grid-graph is constructed with the difference that the nodes correspond to patches in the image rather than single pixels. The auxiliary graphs communicate with each other and with the nodes in the main graph through the use of virtual nodes (15). This framework allows the model to build inductive bias by learning simpler auxiliary tasks in conjunction with the difficult pixel-level task. On a private LV landmark dataset containing ultrasound images for 23755 patients, our model achieves mean absolute errors of 2.3 mm, 1.1 mm, 1.2 mm, and mean percent error rates of 5.1%, 11.8%, 12.9% for the landmarks of LVID, IVS, and LVPW respectively.

Our contributions are twofold:

- We introduce a novel, GNN-based, hierarchical framework for LV landmark detection and pixel segmentation.
- We explore different approaches to building a hierarchical framework through the use of CNNs, U-Nets and Average Pooling.

2 Related work

The first work to tackle the task of LVID detection was proposed by Sofka et al. (16), where they performed regression of the corresponding clinical landmark coordinates using a CNN, with the addition of Long Short-Term Memory (LSTM) units as a temporal regularizer. Gilbert et al. (17) proposed a modified U-Net to effectively estimate LVID, IVS and LVPW measurements. Lin et al. (18) proposed a landmark detection and tracking system with a cycle consistency loss to track the landmarks through unlabeled echo frames. Finally, Jafari et al. (19) proposed an uncertainty-driven video landmark and key frame detection framework.

These works are all convolution-based in a supervised learning setting for direct pixel-level predictions, which require the model to predict only a few positive pixels in a high-dimensional image. To address this, prior work smoothes the pixel labels by adding a Gaussian distribution around landmarks, which introduces bias in the learning process. Additionally, during the training process, the location of each landmark location is independent of other landmarks, which is not a correct assumption as landmarks identifying LV walls are highly correlated. Our model aims to remedy the former problem by the introduction of more tractable auxiliary tasks to guide the main pixel-level task so that the model learns the location of the landmarks without reliance on Gaussian label smoothing and addresses the latter problem by allowing communication among pixels through the message passing operations inherent in GNNs (20).

Lastly, to the best of our knowledge, while GNNs have never been applied to the task of LV landmark detection, they have been used for landmark detection in other domains. Li et al. (21) proposed to perform landmark detection via modeling the landmarks with a graph, and performing a cascaded

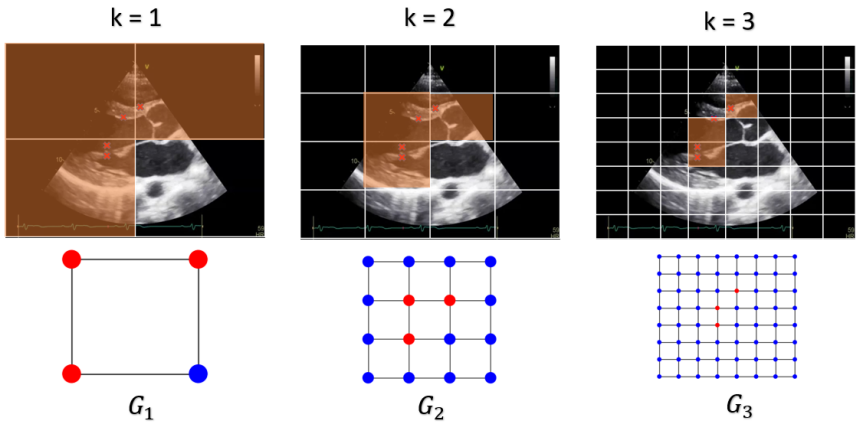

Figure 1: Hierarchical graph generation for K=3. Blue nodes are the non-landmark nodes, and red nodes are nodes where the corresponding patch contains a landmark pixel.

regression of the locations. These models, however, do not have the same hierarchical approach as our framework where auxiliary (and easier) tasks are learned in conjunction with the main task.

3 Method

3.1 Problem Setup

We consider the following supervised setting for LV wall landmark detection: We have a dataset $D = \{X, Y\}$ where $|D| = n$ is the number of $\{x^i, y^i\}$ pairs such that $x^i \in X$ and $y^i \in Y$ where $i \in [1..n]$. Each $x^i \in \mathbb{R}^{H \times W}$ is an ultrasound image of the heart where H and W are height and width of the image respectively, and each y^i is the set of 4 point coordinates $[(h_1^i, w_1^i), (h_2^i, w_2^i), (h_3^i, w_3^i), (h_4^i, w_4^i)]$ indicating the landmark locations in x^i.

3.2 Hierarchical Graph Creation

Let us denote a graph with $G(V, E)$ where V is the set of nodes, and E is the set of edges in the graph such that if $v_i, v_j \in V$ and there is an edge from v_i to v_j then $e_{i,j} \in E$. During the training phase, for each image $x^i \in X_{train}$, K different graphs $G_k(V_k, E_k)$ are constructed using the following steps where for each $k \in [1..K]$:

1. $2^k \times 2^k = 4^k$ nodes are added to V_k to represent each patch in the image. Note that larger k corresponds to graphs of finer resolution, while smaller k corresponds to coarser graphs.
2. Undirected edges are added in a grid like manner such that $e_{l-1,s}, e_{l+1,s}, e_{l,s-1}, e_{l,s+1} \in E_k$ for each $l, s \in [1..2^k]$.
3. A patch feature embedding h_j^k where $j \in [1..4^k]$ is associated with that patch's node $v_j \in V_k$. Different patch feature generation techniques are described in Section 3.3.
4. Binary node labels $\hat{y}_k^i \in \{0, 1\}^{4^k}$ are generated such that $\hat{y}_{kj}^i = 1$ if at least one of point coordinates in y^i falls within the patch associated with node $v_j \in V_k$.

An example with $K = 3$ is shown in Figure 1.

3.3 Node Feature Creation

As described in Section 3.2, each node in an auxiliary graph corresponds to a patch in the input image. Different methods of generating features for these nodes are explored in the following subsections. Additionally, an ablation study is provided in Section 4.5 to show the effects of each one of these approaches to generating node features.

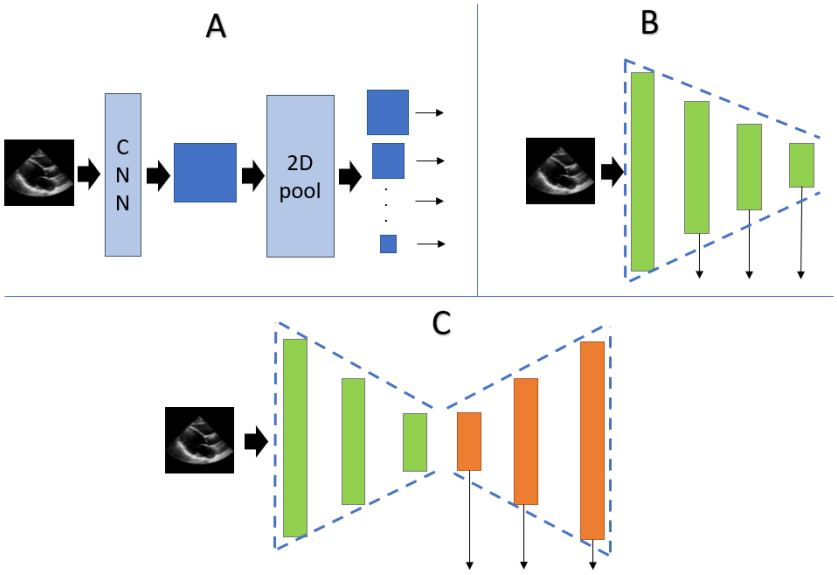

Figure 2: Different approaches to node feature generation: (A) 2D average pooling layers with different kernel sizes are used to generate features for nodes of auxiliary graphs with different coarseness levels. (B) Multiple CNN layers are used to transform the image, and the intermediate features are used for node features such that deeper layers contain the features for coarser graphs. (C) The intermediate features of the decoder part of a U-Net are used as node features such that deeper representations correspond to node features of finer graphs.

2D Average Pooling: Inspired by (22), we can directly generate embeddings for each node by extracting patches of different sizes from the image. To achieve this, depending on the coarseness of the corresponding graph, we use 2D average pooling layers with different kernel sizes to summarize patch-level information. That is, as shown in Figure 2A, for each $k \in K$, we use a 2D average pooling layer with a kernel size of $(\lfloor H/2^k \rfloor, \lfloor W/2^k \rfloor)$ and a step size of $(\lfloor H/2^k \rfloor, \lfloor W/2^k \rfloor)$.

Intermediate CNN Features: As shown in Figure 2B, a multi-layer CNN is used to generate the features for the auxiliary graphs such that deeper layers contain the features for coarser graphs. The kernel size for each CNN layer is determined so that the resulting intermediate feature map's dimension matches the number of nodes in the corresponding graph. The intuition behind this approach is that deeper features have larger receptive fields corresponding to large patches of the original image.

Intermediate U-Net Features: In the CNN-approach to generating node features, the initial feature embeddings are used for the main pixel-level graph, while deeper embeddings provide the features for the auxiliary tasks. This may be a sub-optimal approach as the main graph lacks the more abstract features obtained from the deeper CNN layers. To remedy this, as shown in Figure 2C, the decoder part of a U-Net can be used to obtain node features such that deeper layer embeddings correspond to the node features for the finer graphs. This means that the main pixel-level graph would have the features from the last layer of the network.

3.4 Virtual Nodes and Inter-Task Communication

to summarize the information in each of the hierarchical graphs G_k, we add a virtual node V_{Gk} and assume virtual edges between V_{Gk} and all nodes $v_j \in V_k$ exist. To allow inter-task communication, all possible edges between these virtual nodes are also added as shown in Figure 3. The combination of auxiliary graphs, the main pixel-level graph and the virtual nodes create a larger graph representing each sample denoted by G_i where $i \in [1..n]$.

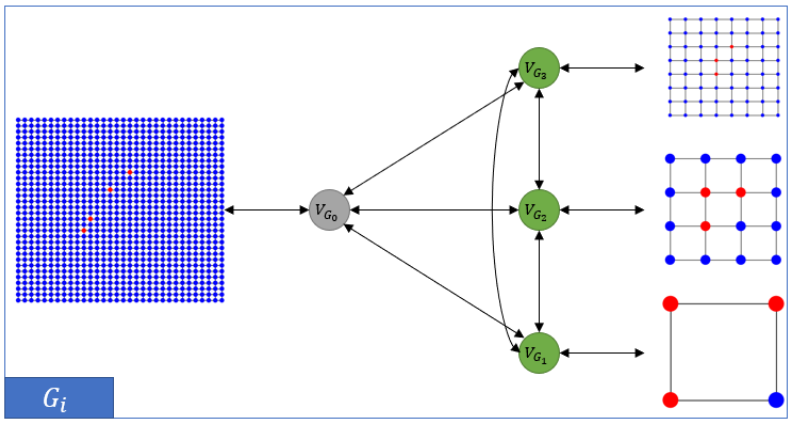

Figure 3: Overall graph for each sample: It consists of the main pixel-level graph and all auxiliary graphs. Inter-task communication is enabled through the use of virtual nodes and connections between them.

3.5 Training and Objective Functions

The graphs created for each sample are fed into GNN layers followed by an MLP layer such that:

$$h_{\text{nodes}} = \text{ReLU}(\text{GNN}_l(G_i)), \text{for l in [1..L]} \tag{1}$$

$$h_{\text{out}} = \sigma(\text{MLP}(h_{\text{nodes}})) \tag{2}$$

Where σ is the Sigmoid function, $h_{\text{nodes}} \in \mathbb{R}^{|VG_i| \times d}$ is the set of d-dimensional embeddings for all nodes in the graph, and $h_{\text{out}} \in [0,1]^{|VG_i| \times 4}$ is the 4-channel prediction for each node with each channel corresponding to one of the pixel landmarks. In the inference time, the expected value of each of the heatmaps are taken in x and y directions to extract the landmark coordinates.

To train the network the following objective function are considered: **Binary Cross Entropy (BCE):** We use a BCE loss since each node should either be classified as a landmark or a non-landmark. **Weighted BCE:** Since the number of landmark locations is much smaller than non-landmark locations, we use a higher weight for landmark nodes. **BCE + L2 regression of landmark coordinates:** To give the model better training signals, we add a regression loss term on top of a node classification objective (such as BCE). The regression objective is the L2 loss on the predicted coorindantes vs. the ground truth labels. The impact of these different training objectives is studied in Section 4.5.

4 Experiments

4.1 Dataset

For this study, a private dataset of 28,577 echo cine series of the PLAX from 23,755 patients is used. The data is collected using cart-based ultrasound machines from various manufacturers and annotated by experts for LVID, IVS, and LVPW, each marked by two landmarks on the frame. As shown in Figure 4, landmarks of the lower IVS and the upper LVID, as well as the lower LVID and upper LVPW are shared. Therefore, we summarize the problem to finding the location of four landmarks. We split the dataset with the ratio of 80-10-10 percents for training, validation, and test sets respectively while ensuring that the splits are patient-exclusive.

4.2 Quantitative Results

The four landmark coordinate predictions of the model and the expert annotations are used to create the predicted and ground truth LVID, IVS and LVPW measurements. The unit for these measurements is pixels and must be converted to millimeters (mm) using pixel to mm ratios, which are available and specific to every frame.

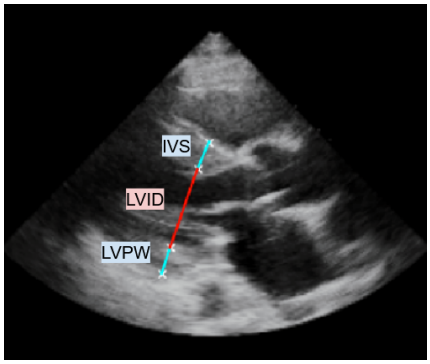

Figure 4: An example echo frame showing a line containing IVS, LVID and LVPW measurments from top to bottom. These measurements require 6 pixel locations to be identified. However, as seen in the figure, two pixel locations are shared between these lines, and therefore, 4 landmarks are enough to characterize these measurements.

The model is evaluated in terms of how close predicted and ground truth errors are. More specifically, the error is calculated using Mean Absolute Error (MAE) in mm, and Mean Percent Error (MPE) in percents as follows:

$$\text{MPE} = 100 \times \frac{|L_{\text{pred}} - L_{\text{true}}|}{L_{\text{true}}} \tag{3}$$

$$\text{MAE} = |L_{\text{pred}} - L_{\text{true}}| \tag{4}$$

where $L_{\text{pred}}, L_{\text{true}}$ are the prediction and ground truth values for every measurements. Our results and comparisons to prior work are shown in Table 1. Additionally, scatter plots showing true and predicted landmark locations are provided in Figure 6. A discussion of these results is provided in Section 4.4.

Table 1: Quantitative results in terms of MAE and MPE on the test set. Lower values are better for MAE and MPE. The code for some models is not available; therefore, results outside the ones reported in the original papers cannot be obtained as indicated by "-" in the table. Modified U-Net uses a U-Net with elongated Gaussian labels, while RDT attempts to track the landmarks through consecutive frames using a cyclic consistency loss. Multi-Task U-Net, an orthogonal work of ours, uses a multi-headed U-Net architecture and is semi-supervised on the annotated frames of the video. Even though Multi-Task U-Net uses sample rejection and video data, we see that our model beats their performance in some measurements and performs closely for others without any sample rejections and only using individual frames.

Model	MAE [mm]			MPE [%]		
	LVID	IVS	LVPW	LVID	IVS	LVPW
Modified U-Net (17)	34	-	-	6.0	13.4	**10.8**
RDT (18)	8.1	-	-	-	-	-
Multi-Task U-Net	2.4	**1.1**	**1.1**	5.3	11.9	12.4
HiGNN (Ours)	**2.3**	**1.1**	1.2	**5.1**	**11.8**	12.9

4.3 Qualatative Results

As seen in Figure 5(LEFT), for an input echo frame with higher quality and clearer boundaries, the model generates focused heatmaps at the locations of the landmarks, and the predictions are very close to the ground truth. In Figure 5(RIGHT), a low quality case is shown where the model has higher uncertainty. In this case, the prediction is diffused along the direction of the walls of the LV, showing that the model has learned the approximate area of the landmarks, while not being confident due to the noisy nature of the frame.

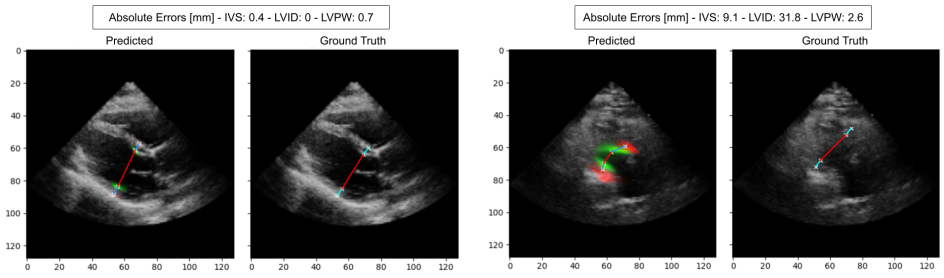

Figure 5: Visualisation of model prediction heatmaps and ground truth labels. (LEFT) We see an example where the image has high quality and the LV wall boundaries are clear. In this case, the model's heatmaps are more focused, and predicted landmarks have low error. (RIGHT) We see an example where the image is noisy and of low quality, which results in the model being less confident and its heatmaps being more diffused.

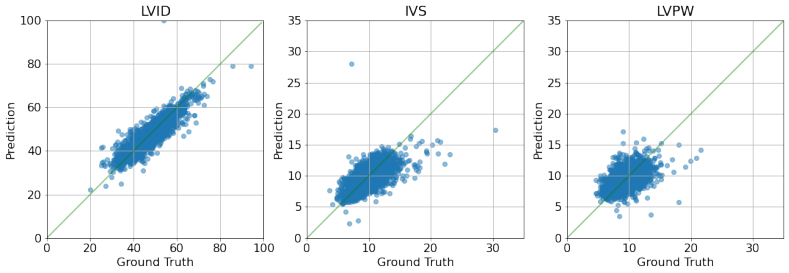

Figure 6: Scatter plots of the predicted measurements [mm] vs. ground truth [mm]. The ideal line is shown in green.

4.4 Discussion

As shown in Table 1, our model outperforms the state of art in most measurements without any reliance on sample rejections and only using individual frames rather than videos. Most prior work put a Gaussian distribution around ground truth landmarks in order to ease the training procedure and avoid having to detect a single positive pixel among many negative ones, which can introduce bias in the learned coordinates. Additionally, sample rejection is usually used to ignore out of distribution samples and increase test time performance. Furthermore, most models rely on pretraining to increase convergence speed. Our model, however, is not using pretraining or any sample rejection tricks and is using a Weighted BCE objective that does not rely on smoothed labels. Therefore, we argue that the performance benefits of our model come from the following components:

- GNNs: Representing each frame and its coarser versions as grid graphs that have connections between them, as shown in Figure 3, allows widespread message passing and communication between pixels and tasks. Such communication is not possible in prior works that solely rely on CNNs and U-Nets. This inter and intra-task communication allows the model to properly capture dependencies among pixels.
- Hierarchical Framework: Allowing the model to attend to simpler tasks that correspond to coarser versions of the main frame makes the learning process more tractable. This is because the model applies the inductive bias learned from simpler tasks to the main difficult task, allowing it to capture nuances and important information in the data.

4.5 Ablation Studies

As shown in Table 2, the impact of different node feature extraction methods on model performance is studied. We see that the U-Net based model significantly outperforms others. We postulate that the CNN and Average Pooling-based methods perform poorly because the shallow representations

Table 2: Quantitative evaluation results for different node feature extraction methods on the validation set. We see that the U-Net-based method outperforms others due to extracting features from deeper layers of the model. Please note that all these models use the Weighted BCE variant of the training objective

Model	MAE [mm]			MPE [%]		
	LVID	IVS	LVPW	LVID	IVS	LVPW
Average Pooling	26.6	9.4	11.6	56.7	100	129.9
CNN	24.7	10.4	10.9	52.6	111.1	121.5
U-Net	2.4	1.3	1.2	5.2	13.9	13.4

for the main-level graph are insufficient in providing enough information for the task. On the other hand, the U-Net-based method provides more abstract features for the graphs since the decoder is positioned in deeper layers of the model.

As seen in Table 3, we also provide ablation results on different objective functions introduced in Section 3.5. We see that the Weighted BCE + Regression variant outperforms all other objective functions. As expected, the BCE objective performs the weakest due to the number of negative pixels being larger than the positive ones, while the Weighted BCE accounts for such imbalance.

Table 3: Quantitative evaluation results for different training loss functions on the validation set. Please note that all these models use the Average Pooling variant of the feature extractor. We see that Weighted BCE + Regression outperforms all other model variants. Both Weighted BCE and Weighted BCE + Regression variants outperform BCE because they account for the number of positive pixels compared to negative ones.

Model	MAE [mm]			MPE [%]		
	LVID	IVS	LVPW	LVID	IVS	LVPW
Weighted BCE + Regression	11.5	5.8	6.5	24.4	61.5	71.8
Weighted BCE	26.6	9.4	11.6	56.7	99.9	129.9
BCE	44	6.4	5.8	94.4	65.1	59.8

5 Conclusion and Future Work

In this work, we introduce a novel medical landmark detection model that incorporates GNNs into a hierarchical framework. The model performs better than the state of the art on most measurements without reliance on network pretraining, sample rejection or label smoothing and by only using a single frame rather than an entire video. We postulate that the performance benefits of our model arise from two architectural choices. Firstly, the use of GNNs allows communication between tasks and pixels, enabling information to be propagated in ways that prior work lack. Secondly, the hierarchical approach allows the model to build better inductive bias by solving simpler tasks in conjunction with the main task and learning nuances in the data.

While the model shows promising performance, we believe that there are certain shortcomings that must be addressed in future work. As an instance, for each task, the model builds a grid graph where each node is connected to its horizontal and vertical neighbours. This assumption may not be suitable since the pixel dependencies go beyond the 1-hop vertical and horizontal relationships. Moreover, the content of echo frames have a conical shape, and reflecting this shape in the way the graphs are built can help with the performance of the model. Another shortcoming of the model is that it must learn proper initial node features along with the learning of landmark locations, which makes convergence more difficult. We argue that using pretraining for the frame extraction network (e.g. the initial U-Net model), significantly aids the model in finding an optimal solution by providing more meaningful initial features for the nodes in the graph. Lastly, the model must be expanded to train and infer on echo videos rather than individual frames, which is a challenging task in terms of how the hierarchical graphs are created for video data.

References

[1] L. Brattain, B. Telfer, M. Dhyani *et al.*, "Machine learning for medical ultrasound: status, methods, and future opportunities. abdom radiol 43: 786–799," 2018.

[2] T. M. McFarland, M. Alam, S. Goldstein, S. D. Pickard, and P. D. Stein, "Echocardiographic diagnosis of left ventricular hypertrophy," *Circulation*, vol. 57, no. 6, pp. 1140–1144, 1978. [Online]. Available: http://ahajournals.org

[3] 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, 2015, pp. 234–241.

[4] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng, "Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 8236–8246.

[5] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3691–3700.

[6] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3476–3483.

[7] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1653–1660.

[8] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4177–4187.

[9] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4998–5006.

[10] R. Chen, Y. Ma, N. Chen, D. Lee, and W. Wang, "Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham: Springer International Publishing, 2019, pp. 873–881.

[11] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 483–499.

[12] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using cnns," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, pp. 230–238.

[13] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5686–5696.

[14] Z. Zhong, J. Li, Z. Zhang, Z. Jiao, and X. Gao, "An attention-guided deep regression model for landmark detection in cephalograms," *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, p. 540–548, 2019. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-32226-7_60

[15] E. Hwang, V. Thost, S. S. Dasgupta, and T. Ma, "Revisiting virtual nodes in graph neural networks for link prediction," 2022. [Online]. Available: https://openreview.net/forum?id=ETiaOyNwJW

[16] M. Sofka, F. Milletari, J. Jia, and A. Rothberg, "Fully convolutional regression network for accurate detection of measurement points," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 258–266.

[17] A. Gilbert, M. Holden, L. Eikvil, S. A. Aase, E. Samset, and K. McLeod, "Automated left ventricle dimension measurement in 2d cardiac ultrasound via an anatomically meaningful cnn approach," in *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis*. Springer, 2019, pp. 29–37.

[18] J. Lin, G. Sahebzamani, C. Luong, F. T. Dezaki, M. Jafari, P. Abolmaesumi, and T. Tsang, "Reciprocal landmark detection and tracking with extremely few annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 170–15 179.

[19] M. H. Jafari, C. Luong, M. Tsang, A. N. Gu, N. Van Woudenberg, R. Rohling, T. Tsang, and P. Abolmaesumi, "U-land: Uncertainty-driven video landmark detection," *IEEE Transactions on Medical Imaging*, 2021.

[20] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," *CoRR*, vol. abs/1704.01212, 2017.

[21] W. Li, Y. Lu, K. Zheng, H. Liao, C. Lin, J. Luo, C.-T. Cheng, J. Xiao, L. Lu, C.-F. Kuo *et al.*, "Structured landmark detection via topology-adapting deep graph learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 266–283.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929