2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340526

Segmenting Cardiac Ultrasound Videos Using Self-Supervised Learning

Erik Lamoureux[1*], Sana Ayromlou[2*], Seyedeh Neda Ahmadi Amiri[3*], and Helge Rhodin[4]

***Abstract*— Deep learning models trained with an insufficient volume of data can often fail to generalize between different equipment, clinics, and clinicians or fail to achieve acceptable performance. We improve cardiac ultrasound segmentation models using unlabeled data to learn recurrent anatomical representations via self-supervision. In addition, we leverage supervised local contrastive learning on sparse labels to improve the segmentation and reduce the need for large amounts of dense pixel-level supervisory annotations. Then, we implement supervised fine-tuning to segment key temporal anatomical features to estimate the cardiac Ejection Fraction (EF). We show that pretraining the network weights using self-supervised learning for subsequent supervised contrastive learning outperforms learning from scratch, validated using two state-of-the-art segmentation models, the DeepLabv3+ and Attention U-Net.**

***Clinical relevance*—This work has clinical relevance for assisting physicians when conducting cardiac function evaluations. We improve cardiac ejection fraction evaluation compared to previous methods, helping to alleviate the burden associated with acquiring labeled images.**

I. INTRODUCTION

Cardiovascular disease is the leading cause of death globally and is commonly assessed using ultrasound imaging. By measuring cardiac health in a simple and accessible manner, more people with cardiac conditions could be identified to receive appropriate treatment, preventing premature deaths. Cardiac ejection fraction (EF) is the ratio of the blood pumped out of the ventricle (stroke volume) to the maximum amount of blood in the ventricle (end-diastolic volume). It is a commonly used metric for determining functional cardiac health and is used for various clinical evaluations, and diagnoses [1]–[3]. By accurately measuring EF in an accessible manner, clinicians have easy access to critical information that can help diagnose and treat cardiac patients.

Automated methods to assess cardiac health include using deep convolutional neural networks (CNNs) for segmenting cardiac anatomical features and assessing the ejection fraction [4]. Specifically, U-Net, a well-known medical image segmentation network, has been used to assess the end-systolic (ES) and end-diastolic (ED) frames from ultrasound (US) videos to segment the left ventricle [5]. Other previous work includes calculating left ventricle EF without segmentation with Conv3D [6], ResNet(2+1D), and CNN followed by LSTM [7] using regression loss. Other segmentation techniques include using co-learning from appearance and shape to increase both temporal and spatial accuracy [8]. Recently, transformers have been used to estimate ejection fraction [9]. In [9], the authors first applied a Residual AutoEncoder Network on an input video to reduce its dimensionality. Then, a BERT model [10] was adapted for token classification, providing reasoning ability in the spatiotemporal domain. However, the heart's sophisticated anatomical structure and low resolution in most medical imaging modalities makes training and tuning these type of network architectures difficult. Further, BERT and related transformer networks require a large training set and a long time to train. This is disadvantageous for medical imaging as these requirements are unlikely to be satisfied in routine clinical settings.

Although it is relatively easy for the average observer to label natural data, annotating medical data requires input from domain experts, which is costly, time-consuming, and laborious. Further, due to the domain shift among data gathered by different medical devices with differing experimental apparatuses and due to the need to respect patient privacy, it is difficult to acquire large, uniform, and labeled data from several institutions. However, unlike the rarity of labeled medical data, unlabeled data is inherently informative and easier to acquire. One way to overcome the barriers associated with a lack of labeled data is to pretrain the network using transfer learning. Transfer learning provides the model with proper weight initialization [11], providing an improved starting point for gradient descent optimization compared to the use of standard random values. This helps the network converge faster with fewer data. Transfer learning can be implemented using self-supervised learning, where the model is primed with unlabeled data to extract intrinsic information. After this pretraining, the model is subsequently trained with labeled data to achieve supervised feature learning [12], [13]. Advances in self-supervised learning show promise for improving segmentation and classification outcomes by reducing the need for large amounts of annotated images [14].

Recently, self- and semi-supervised learning has been coupled with contrastive learning for medical image segmentation [14]–[22]. Mechanistically, contrastive learning forces the embedding features of similar images to be near to each other in the latent space, whereas those that are dissimilar are more apart. Introducing contrastive learning loss improves performance, demonstrated by achieving state-of-the-art in

This work was not supported by any organization.

∗ Equal contribution.

[1]Mechanical Engineering, University of British Columbia Vancouver, BC V6T 1Z4, Canada erik.lamoureux@ubc.ca

[2]Electrical and Computer Engineering, University of British Columbia Vancouver, BC V6T 1Z4, Canada s.ayromlou@ece.ubc.ca

[3]Electrical and Computer Engineering, University of British Columbia Vancouver, BC V6T 1Z4, Canada nedaahmadi@ece.ubc.ca

[4]Computer Science, University of British Columbia Vancouver, BC V6T 1Z4, Canada rhodin@cs.ubc.ca

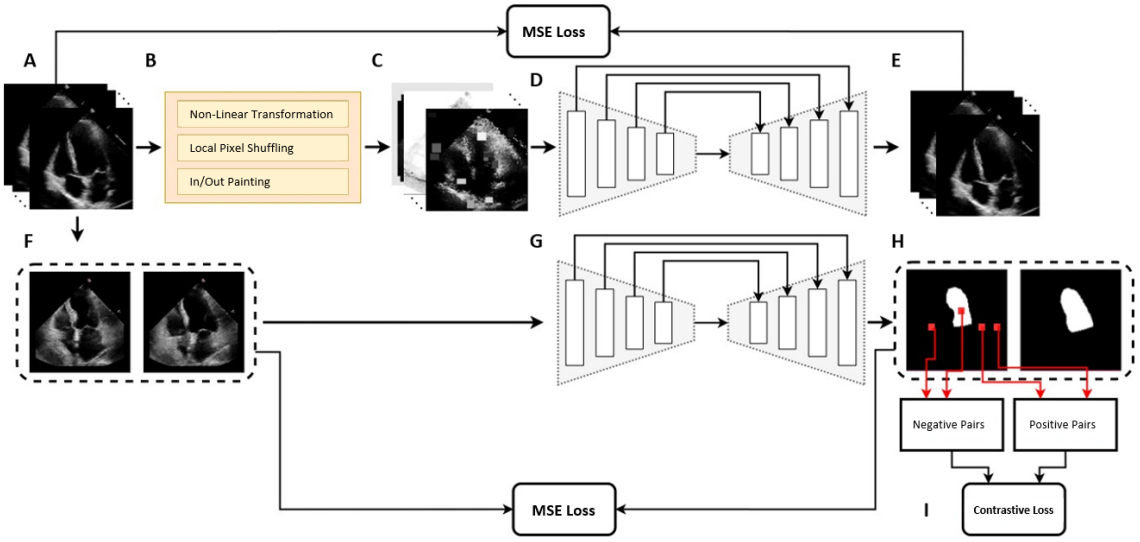

Fig. 1. Overview. (A-E) Self-supervised learning. (A) Input images. (B) Self-supervision techniques to produce deformed images. (C) Self-supervision outputs are input to the encoder-decoder network (D). (E) Decoder outputs are used to calculate mean-squared error (MSE) loss. (F-H) Segmentation. (F) Labeled input images for segmentation are run through the pretrained encoder-decoder network (G). Note that the pretrained weights determined in encoder-decoder D are used in G. The outputs are used to calculate MSE loss and contrastive loss based on negative and positive pairs (H). (I) Contrastive loss for contrastive learning.

self- and semi-supervised classification problems [23]. There are many applications of self-supervised and contrastive learning including using global and local contrastive loss for medical image segmentation [16], image reconstruction [24] and remote sensing scene representation [25]. Its powerful ability to extract essential features for downstream tasks from unlabeled data makes it a great candidate for label-efficient learning.

Here, we use self-supervised learning to leverage model pretraining to learn the anatomical structure of the heart by reconstructing deformed versions of unlabeled video frames. In this way, the encoder-decoder blocks are trained to obtain intermediate feature maps. Next, the self-supervised pretrained model is finetuned by applying supervised learning using the labeled End Diastolic (ED) and End Systolic (ES) frames of videos from the EchoNet-Dynamic dataset [26]. This supervised approach uses pixel-level contrastive learning, where pixel-level embeddings are used to calculate the local contrastive loss. By including this information for supervision, contrastive learning can better extract important image features. Overall, this work demonstrates self-supervised pretraining and supervised local contrastive learning to assess the ejection fraction of cardiac ultrasound images.

II. METHODS

In this paper, we apply three cardiac ultrasound segmentation evaluation types—baseline segmentation, self-supervised learning, and contrastive learning—using two different models: the DeepLabv3+ [27] and Attention U-Net [28]. An overview of our approach is displayed in Fig. 1.

A. Self-Supervised Learning

We used the EchoNet-Dynamic dataset [26], consisting of 10,036 echocardiogram videos and human expert annotations.

Within each video, only two frames, the End Diastolic (ED) and End Systolic (ES) have segmentation labels. These frames capture the end of diastole (relaxation) and the end of systole (contraction) of the heart, respectively. We used four self-supervised learning training schemes to utilize the available unlabeled frames. These are used to train source segmentation models for yielding high-performance supervision signals *via* transfer learning.

Influenced by Model Genesis's deformation techniques applied to MRI and CT images [29], we developed a self-supervised training scheme using transformed versions of randomly selected frames within each ultrasound video. Then, we train the source model to reconstruct the original clean frame given the transformed version. This approach causes the model to learn common image representations, creating a robust, adaptive, transferable, and generalizable model that is well able to handle variations common in ultrasound images.

1) Learning Appearance via Image Transformation: Since semantic segmentation methods require pixel-wise annotations, learning absolute or relative intensity values can be informative. Using pixel intensity values for supervision, a non-linear, monotonous, one-to-one mapping was applied to pixel values using a Bézier Curve (Fig. 2A). By trying to restore the original intensities from the transformed images, the model learns the structure and appearance of anatomical cardiac features.

2) Learning Texture via Local Pixel Shuffling: The second deformation method involves sampling a random window

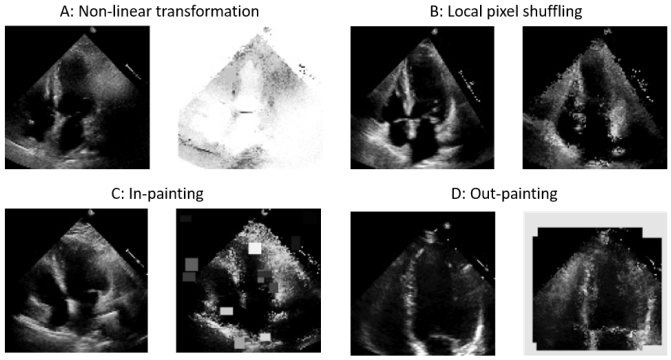

Fig. 2. Example image transformation and deformation techniques. (A) Non-linear transformation. (B) Local pixel shuffling. (C) In-painting. (D) Out-painting.

and shuffling the order of contained pixels (Fig. 2B). As the size of chosen window increases, it is more difficult for the model to recover the original image. Further, if the window is too large, it can change the global content of the image. Therefore, the window size was chosen to be smaller than the model's receptive field. This deformation approach teaches the model about segmentation boundaries and texture.

3) Learning Context via In-Painting: In deformation by learning context *via* in-painting, the model learns to fill in missing data due to the introduction of noise (Fig. 2C). The noise consists of random small windows within the image that are replaced with constant random values. This forces the model to learn the local continuities of anatomical heart structures.

4) Learning Context via Out-Painting: For out-painting deformation, multiple windows of various sizes and aspect ratios were generated and placed on top of each other, resulting in a single window of a complex shape. Then, pixels outside the window get a random value (Fig. 2D). This helps the model learn global geometry and spatial layout.

5) Self-Supervision Summary: By priming the model with images based on non-linear transformations, local pixel shuffling, and in/out painting, the model learns the sophisticated and recurrent anatomy based on these free supervision signals *via* self-supervision. With a larger training dataset, we reduce the chance of over-fitting and improve overall performance. Overall, training the encoder-decoder for reconstruction (Fig. 1D) provides the initial weights for the subsequent encoder-decoder framework for target segmentation (Fig. 1G).

B. Supervised Local Contrastive Learning

To bolster the supervised segmentation training, we use local contrastive learning to enable the model to extract distinctive local representations. The l-th upper-most layer of the decoder has the same dimensions as the input image and is input to a two-layer pixel-wise convolution, h. This produces a 128-dimensional embedding for each pixel position, feature map, f, and contrastive loss, L,

$$f(\tilde{x}_i) = h(D_l(E(x_i))) \tag{1}$$

$$L(x_i) = -\frac{1}{|\Omega|}\sum_{(u,v)\in\Omega}\frac{1}{|P(u,v)|} \log\frac{\sum_{(u_p,v_p)\in P(u,v)}\exp(f_{u,v}\cdot f_{u_p,v_p}/\tau)}{\sum_{(u',v')\in N(u,v)}\exp(f_{u,v}\cdot f_{u',p'}/\tau)}, \tag{2}$$

where, $P(u,v)$ denotes the positive set and $N(u,v)$ denotes the negative set of $f_{u,v}(x_i)$. For a given point (u,v) in the feature map, the positive set $P(u,v)$ describes all features in the feature maps of a given batch that share the same annotation as (u,v) [18]. The negative set is all the features in the feature maps that have different labels. Further, as background pixels make up a large number of positive pairs for comparison to left ventricle segmentation pixels, Ω only contains points with non-background annotation. These background pixels are not lost, rather they are included in the negative set $N(u,v)$. Using labels, supervised contrastive loss provides additional information on the similarity of features derived from the same class and the dissimilarity of features from different classes. Then, the total local contrastive loss is calculated by averaging all losses

$$L_c = \frac{1}{|A|}\sum_{x_i\in A} L(x_i). \tag{3}$$

C. Finetuning the Target Model for Segmentation

Finally, by using the ES and ED frames of each video, their annotations, and initialized weights, the target model is finetuned for segmentation. Finetuning is a type of transfer learning that starts with a pretrained model and updates all of the model's parameters when applied to a new task, essentially retraining the entire model. Here, we pretrain the model using the self-supervised-derived deformed images, then apply the supervised local contrastive learning for left ventricle segmentation.

III. RESULTS

In order to follow the typical trend in machine learning research, we conducted multiple experiments on a convincingly large real-world dataset to compare our approach to prior work. Moreover, we conducted an ablation study by adding other experiments to analyze the importance of specific design aspects.

A. Data & Technical Requirements

EchoNet-Dynamic. The EchoNet-Dynamic dataset [26] is a large echocardiography dataset of the left ventricle's motion during cardiac cycles. It consists of apical four-chamber (A4C) view videos (112x112 pixels) from 10,036 patients acquired between 2016 and 2018 at Stanford Health Care. Each video is labeled with the corresponding left ventricle border tracing, EF, ED, and ES frame indexes, and volume of the left ventricle at the end-systole (ES) and end-diastole (ED) frames, determined by expert sonographers. Each video consists of 24-1,002 frames, providing large unlabeled datasets for self-supervision.

Computational Resources. We used two NVIDIA Tesla V100 GPUs with 16GB of memory for training and evaluation.

Experimental Details. We evaluated the proposed methods using a collection of experiments. Specifically, we assessed baseline segmentation methods (DeepLabv3+ and Attention U-Net), applied self-supervised learning to compare the performance of the models with and without initially pretrained weights, and conducted supervised local contrastive learning.

Here, the EchoNet-Dynamic dataset is applied to both the DeepLabv3+ and Attention U-Net model. We use the two labeled frames in each video, the End Diastolic (ED) and End Systolic (ES) frames, to conduct supervised learning for segmentation. This collection of images is split 70/10/20 for training, validation, and testing, respectively. Optimal hyperparameters for the model are determined automatically using "Sweep" in Weights and Biases (https://wandb.ai/site). The performance of the baseline segmentation approaches is compared with each other, with self-supervised learning using pretrained network weights, and with contrastive learning.

B. Baseline Models

Here, we use semantic segmentation methods (Fig. 3) for pixel-specific labeling using the DeepLabv3+ and Attention U-Net models.

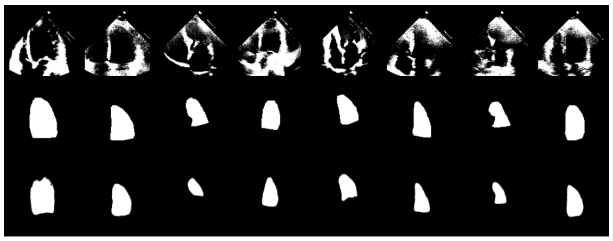

Fig. 3. Example segmentation results using DeepLabv3+. The top row is the input image, the middle row is the output predictions for the left ventricle's segmentation, and the bottom row is the batch end diastolic frames.

1) DeepLabv3+ Segmentation: We find the DeepLabv3+ model behaves well, with training converging in 10 epochs over a training runtime of 3 hours. The trained model produces a testing dice coefficient of 90.64% (Fig. 4A). Our implementation of DeepLabv3+ slightly under-performs the EchoNet-Dynamic paper's use of DeepLabv3 (dice coefficient 92% [30]). However, these performance differences are minor and could be due to differences in the models, hyperparameter tuning (e.g. our 10 epochs compared to their 50), computational resources, or training time.

2) Attention U-Net Segmentation: We find that the Attention U-Net model achieves better outcomes compared to the DeepLabv3+ model as the segmentation is more accurate (Fig. 4B, Table II). We suspect this is because the attention gate helps the model segment the border pixels better by assigning it more weight, leading to more robust results than DeepLabv3+.

C. Self-Supervised Learning

The self-supervised learning experiments seek to improve upon the baseline segmentation methods with the ultimate

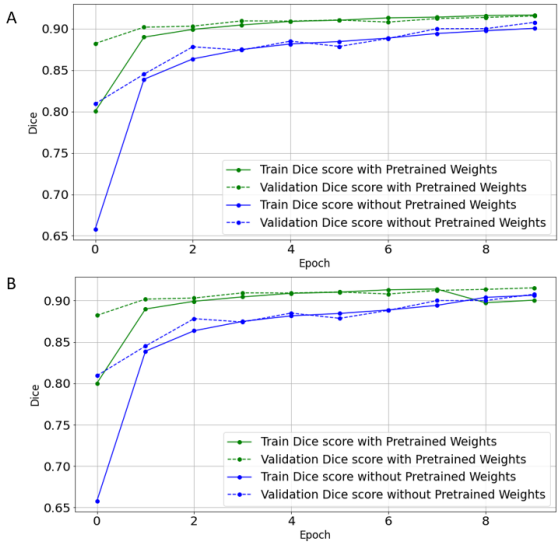

Fig. 4. DeepLabv3+ (A) and Attention U-Net (B) segmentation with and without pretrained weights.

aim of surpassing the performance of the EchoNet-Dynamic paper [30] and Saeed et al. (2022) [20]. Self-supervised learning utilizes a variety of methods to produce deformed images from the unlabeled dataset (see II): non-linear transformation, local pixel shuffling, in-painting, and out-painting (Fig. 5).

By expanding the dataset presented to the model, the model learns from a larger batch of images with increased variation. Additionally, the deformation procedures force the network to learn how to fill in gaps in the images as vital information is withheld. This provides the trained model with additional adaptability to images that deviate from the training set. Here, the network is pretrained using deformed images from the 70% training set and the 10% validation and 20% testing sets are held-out from training. MSE loss metrics for the self-supervision deformation procedures are found in Table I. For DeepLabv3+ (Attention U-Net), we see local pixel shifting results in the lowest MSE loss at 0.243 (0.226), while non-linear transform and all techniques combined are the highest at 0.402 (0.497) and 0.404 (0.428), respectively. It makes sense that all techniques combined result in comparable performance to the worst deformation technique because more information is lost, beyond the point where it benefits the model.

1) Pretrained vs. Non-Pretrained Weights: We find that for both DeepLabv3+ and Attention U-Net, the pretrained self-supervised networks outperform the baseline segmentation approaches (from III-B).

Finding optimal hyperparameters for both models was done using the Weights & Biases sweep function for both models. We find that self-supervised DeepLabv3+ improves on the baseline method by ~0.5% for the Dice coefficient (Fig. 4A, Table II). Also, we compare the impact of each augmentation on final accuracy and found all four augmentations made the model more robust (Table I, Table II). In fact,

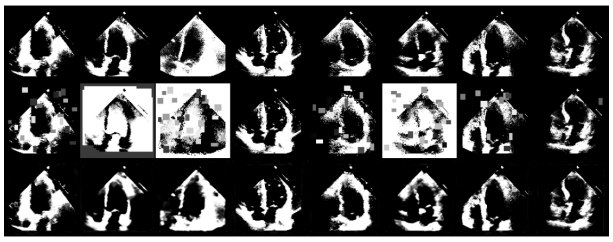

Fig. 5. Top row: a batch of random frames from various videos. Second row: deformed version of selected frames. Final row: reconstructed images after the model has been trained for multiple epochs.

TABLE I
MEAN SQUARED ERROR LOSS (MSE) FOR SELF-SUPERVISION

Model	Self-Supervised Techniques	MSE Loss
DeepLabv3+	Non-Linear Transform	0.402
DeepLabv3+	Local pixel shifting	0.243
DeepLabv3+	In/Out painting	0.315
DeepLabv3+	All-combined	0.404
Attention U-Net	Non-Linear Transform	0.497
Attention U-Net	Local pixel shifting	0.226
Attention U-Net	In/Out painting	0.372
Attention U-Net	All-combined	0.428

local pixel shuffling and in/out painting had the most overall impact due to these techniques more directly augmenting the relevant left ventricle area. The self-supervised approach using Attention U-Net has an advantage of ∼1% over the non-self-supervised method (Fig. 4B, Table II). Further, similar to the baseline segmentation methods, we find that self-supervised Attention U-Net outperforms the self-supervised DeepLabv3+ in terms of training time, training, validation, and testing loss, and Dice coefficient for ejection fraction segmentation.

Overall, we find that the Attention U-Net self-supervised model outperforms the DeepLabv3+ model and performs similarly to the DeepLabv3 model used in the EchoNet-Dynamic paper [30]. This provides us with an indication of the validity of our approach. The improved performance of the self-supervised approach demonstrates its value for medical image segmentation. In a few key areas, we outperformed the EchoNet-Dynamic model, namely: time and memory cost per prediction. Overall, our method displays improved performance compared to previously published work on this dataset [20], [30], with lower computational cost.

D. Visualization and Validation

To further assess our model, we implemented Gradient-weighted Class Activation Mapping (Grad-CAM) [31]. Grad-CAM produces visual explanations from deep networks using gradient-based localization. It uses the gradients of the left ventricle segmentation flowing into the final convolutions layer, resulting in a coarse localization map of the important image regions for determining segmentation [31]. We implemented Grad-CAM using Captum (https://captum.ai/api/layer.html#gradcam). This allows us to plot both positive attributions (green pixels in the third row of Fig. 6) and negative attributions (red pixels in the final row of Fig. 6). Since we aim to segment the left ventricle, boundary pixels should have a higher gradient, as seen in Fig. 6; the green pixels are clustered around anatomical border structures, especially those around the left ventricle. Conversely, red pixels are more randomly distributed. This visualization provides support for the validity of our model, indicating that the model is learning what we expect.

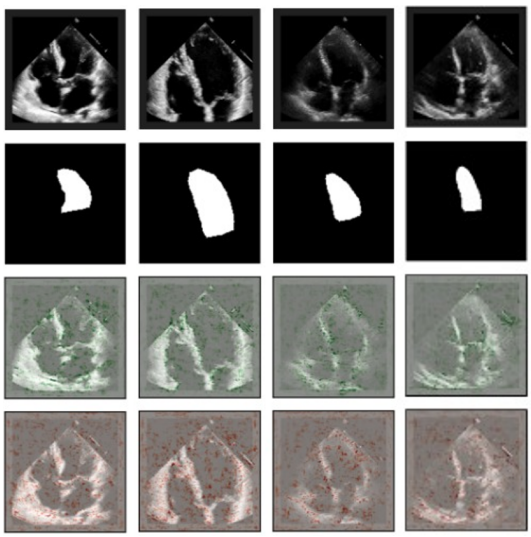

Fig. 6. GradCAM visualization producing localization map expressing important regions of the image for segmentation. First row: original echocardiography images at end-diastolic phase. Second row: predicted left ventricle segmentation. Third row: green pixels indicate areas that contributed more to segmentation learning. Final row: red pixels indicate areas that contributed less to learning.

E. Ablation Study

An ablation study seeks to remove a component of a deep learning network to study its influence on the model's performance. We perform three ablation studies: (1) Determine the impact of each augmentation method (Table I), (2) assess the model with and without pretrained weights (see III-C.1, Table II, and Fig. 4), and (3) compare models using different Dice losses (MSE or contrastive, Table II), and different contrastive learning approaches (Table III). Briefly, we find that pretraining with self-supervision better initialize the models' weights prior to learning on the training set. Additionally, introducing contrastive learning improves the time to convergence and the performance metrics.

For (3), we compare the self-supervision performance using Dice coefficient MSE loss compared to contrastive learning loss. Dice coefficient loss evaluates the performance of a segmentation task by comparing the predicted mask to the ground truth mask. Alternatively, contrastive learning loss provides a measure of the similarity and differences of elements in an image. As expected, we find that the introduction of contrastive learning loss improves the models' performance by ∼1-2% (see Table II). The introduction of contrastive learning loss outperforms the DeepLabv3 model

used in comparable work, EchoNet-Dynamic [30] and Saeed et al. (2022) [20] (Table III).

TABLE II
COMPARISON OF PERFORMANCE METRICS, INCLUDING DICE COEFFICIENT AND LOSS (MSE OR CONTRASTIVE)

Evaluation Type	Model	Performance Metrics	
		Dice	Loss
Baseline Segmentation	DeepLabv3+	90.64%	0.037
	Attention U-Net	90.99%	0.0034
Self-Supervised Learning	DeepLabv3+	91.08%	0.035
	Attention U-Net	91.71%	0.0033
Contrastive Learning	DeepLabv3+	92.60%	0.050
	Attention U-Net	92.93%	0.034

TABLE III
COMPARISON TO STATE-OF-THE-ART METHODS

Method	Model	Pretraining	Dice
EchoNet [30]	DeepLabv3	-	92%
Saeed et al. [20]	DeepLabv3	-	92.04%
Saeed et al. [20]	DeepLabv3	SimCLR [32]	92.52%
Saeed et al. [20]	DeepLabv3	BYOL [33]	92.09%
Saeed et al. [20]	U-Net	-	91.51%
Saeed et al. [20]	U-Net	SimCLR [32]	91.85%
Saeed et al. [20]	U-Net	BYOL [33]	90.70%
Our Contrastive	DeepLabv3+	Our Self-Supervision	**92.60**%
Our Contrastive	Att. U-Net	Our Self-Supervision	**92.93**%

IV. DISCUSSION

We implemented self-supervised and contrastive learning methods for the segmentation of echocardiography videos to assess cardiac ejection fraction. Compared to the baseline supervised segmentation methods, self-supervised and contrastive learning approaches provide significantly improved performance (Table II). Our self-supervised approach meets the previous state-of-the-art and combined with our contrastive learning method provides superior performance (Table III).

In addition to assessing ejection fraction demonstrated here, self-supervised learning has been applied to echocardiography images to conduct view synchronization [34], diagnose mitral regurgitation [35], register echocardiogram dense tracking [36], and to detect atrial fibrillation [37]. Similar to our work, these authors find that pretraining using self-supervision improves echocardiogram segmentation compared to baseline segmentation methods. To further improve self-supervised learning methods, future research should use generative adversarial networks (GANs) for data augmentation. The use of GANs for data augmentation has demonstrated substantially improved performance compared to conventional augmentation techniques when used to segment MRI images [14].

In further pursuit of alleviating the medical annotation bottleneck, contrastive learning approaches have been used to assess cardiac ultrasound images [20]. Contrastive learning applied to echocardiography has been used to assess view classifications [38], to segment the left ventricle [20], for self-supervised transfer learning [39], and for image regression for disease assessment [40]. To better understand the best practices for contrastive learning, it would be worthwhile to further compare different techniques: a simple framework for contrastive learning (SimCLR) [32] (used in [20]), Bootstrap Your Own Latent (BYOL) [33] (used in [20]), supervised local contrastive loss [18] (our contrastive method), learning global and local features [16], momentum contrastive voxel-wise representation learning [21], and local contrastive loss with pseudo-label [17].

Although self-supervised and contrastive learning methods provide high performance especially when annotated data is limited, they require massive amounts of unlabeled data for pretraining to achieve good results [20]. Therefore, semi-supervised contrastive learning can help reduce the medical image labeling burden [14], [17], [18], [21] and should be compared to the supervised local contrastive learning approach used here.

Overall, self-supervised and contrastive learning approaches provide performance benefits for cardiac ultrasound segmentation. These approaches provide additional value for segmenting medical imaging data by reducing the need for expensive high-quality labeled medical images.

DATASET USE AGREEMENT

For use of the EchoNet Dynamic dataset, we adhered to the Stanford University School of Medicine EchoNet-Dynamic Dataset Research Use Agreement (https://echonet.github.io/dynamic/index.html#access). The development and release of this dataset was approved by the Stanford University Institutional Review Board (IRB) [30].

ACKNOWLEDGMENT

We would like to thank Dr. Purang Abolmaesumi of the Robotics and Control Laboratory at the University of British Columbia for providing the computational resources used.

REFERENCES

[1] P. F. Cohn, R. Gorlin, L. H. Cohn, and J. J. Collins Jr, "Left ventricular ejection fraction as a prognostic guide in surgical treatment of coronary and valvular heart disease," The American Journal of Cardiology, vol. 34, no. 2, pp. 136–141, 1974.

[2] S. P. Murphy, N. E. Ibrahim, and J. L. Januzzi, "Heart failure with reduced ejection fraction: a review," JAMA, vol. 324, no. 5, pp. 488–504, 2020.

[3] O. Vedin, C. S. Lam, A. S. Koh, L. Benson, T. H. K. Teng, W. T. Tay, O. O. Braun, G. Savarese, U. Dahlstrom, and L. H. Lund, "Significance of ischemic heart disease in patients with heart failure and preserved, midrange, and reduced ejection fraction: a nationwide cohort study," Circulation: Heart Failure, vol. 10, no. 6, pp. e003875, 2017.

[4] J. H. Park, S. K. Zhou, C. Simopoulos, J. Otsuki and D. Comaniciu, "Automatic Cardiac View Classification of Echocardiogram," IEEE 11th International Conference on Computer Vision, pp. 1-8, 2007.

[5] E. Smistad, A. Østvik, I. M. Salte, S. Leclerc, O. Bernard, and L. Lovstakken, "Fully automatic real-time ejection fraction and mapse measurements in 2d echocardiography using deep neural networks," IEEE International Ultrasonics Symposium (IUS), 2018.

[6] A. Shalbaf, H. Behnam, P. Gifani, and Z. Alizadeh-Sani, "Automatic detection of end systole and end diastole within a sequence of 2-d echocardiographic images using modified isomap algorithm," 1st Middle East Conference on Biomedical Engineering, pp. 217-220, 2011.

[7] A. M. Fiorito, A. Østvik, E. Smistad, S. Leclerc, O. Bernard, and L. Lovstakken, "Detection of cardiac events in echocardiography using 3d convolutional recurrent neural networks," IEEE International Ultrasonics Symposium (IUS), 2018.

[8] H. Wei, H. Cao, Y. Cao, Y. Zhou, W. Xue, D. Ni, and S. Li, "Temporal consistent segmentation of echocardiography with co-learning from appearance and shape," Medical Image Computing and Computer Assisted Intervention, pp. 623–632, 2020.

[9] H. Reynaud, A. Vlontzos, B. Hou, A. Beqiri, P. Leeson, and B. Kainz, "Ultrasound video transformers for cardiac ejection fraction estimation," Medical Image Computing and Computer Assisted Intervention (MICCAI), pp 495-505, 2021.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[11] M. R. Hosseinzadeh Taher, F. Haghighi, R. Feng, M. B. Gotway, and J. Liang, "A systematic benchmarking analysis of transfer learning for medical image analysis," MICCAI Workshop on Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health. Lecture Notes in Computer Science, vol 12968, Springer, pp. 3–13, 2021.

[12] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," IEEE International Conference on Computer Vision, pp. 2051–2060, 2017.

[13] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 11, pp. 4037-4058, 2020.

[14] K. Chaitanya, N. Karani, C. F. Baumgartner, A. Becker, O. Donati, and E. Konukoglu, "Semi-supervised and task-driven data augmentation," 26th International Conference on Information Processing in Medical Imaging, pp. 29–41, 2019.

[15] J. Z. Bengar, J. van de Weijer, B. Twardowski, and B. Raducanu, "Reducing label effort: Self-supervised meets active learning," IEEE/CVF International Conference on Computer Vision, pp. 1631–1639, 2021.

[16] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," Advances in Neural Information Processing Systems, vol. 33, pp. 12546–12558, 2020.

[17] K. Chaitanya and E. Erdil and N. Karani and E. Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," arXiv preprint arXiv:2112.09645, 2021.

[18] X. Hu, D. Zeng, X. Xu, and Y. Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 481–490, 2021.

[19] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 857-876, 2021.

[20] M. Saeed, R. Muhtaseb, and M. Yaqub, "Contrastive pretraining for echocardiography segmentation with limited data," Medical Image Understanding and Analysis, pp. 680–691, 2022.

[21] C. You, R. Zhao, L. Staib, and J. S. Duncan, "Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation," arXiv preprint arXiv:2105.07059, 2021.

[22] X. Zhao, R. Vemulapalli, P. A. Mansfield, B. Gong, B. Green, L. Shapira, and Y. Wu, "Contrastive learning for label efficient semantic segmentation," IEEE/CVF International Conference on Computer Vision, pp. 10623–10633, 2021.

[23] B. Kim, J. Choo, Y.-D. Kwon, S. Joe, S. Min, and Y. Gwon, "Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning," arXiv preprint arXiv:2101.06480, 2021.

[24] F. Paredes-Vall es and G. C. de Croon, "Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3446–3455, 2021.

[25] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1182–1191, 2021.

[26] D. Ouyang, B. He, A. Ghorbani, M. P. Lungren, E. A. Ashley, D. H. Liang, and J. Y. Zou, "Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning," NeurIPS ML4H Workshop: Vancouver, BC, Canada, 2019.

[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European conference on computer vision, pp. 801–818, 2018.

[28] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al., "Attention u-net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.

[29] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis," in Medical Image Computing and Computer Assisted Intervention, pp. 384–393, 2019.

[30] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley et al., "Video-based ai for beat-to-beat assessment of cardiac function," Nature, vol. 580, no. 7802, pp. 252–256, 2020.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," IEEE International Conference on Computer Vision, pp. 618–626, 2017.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," International Conference on Machine Learning, pp. 1597–1607, 2020.

[33] J.-B. Grill, F. Strub, F. Altche, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar et al., "Bootstrap your own latent-a new approach to self-supervised learning," Advances in Neural Information Processing Systems, vol. 33, pp. 21 271–21 284, 2020.

[34] F. T. Dezaki, C. Luong, T. Ginsberg, R. Rohling, K. Gin, P. Abolmaesumi, and T. Tsang, "Echo-syncnet: self-supervised cardiac view synchronization in echocardiography," IEEE Transactions on Medical Imaging, vol. 40, no. 8, pp. 2092–2104, 2021.

[35] F. Yang, J. Zhu, J. Wang, L. Zhang, W. Wang, X. Chen, X. Lin, Q. Wang, D. Burkhoff, S. K. Zhou et al., "Self-supervised learning assisted diagnosis for mitral regurgitation severity classification based on color doppler echocardiography," Annals of Translational Medicine, vol. 10, no. 1, 2022.

[36] W. Zhu, Y. Huang, M. A. Vannan, S. Liu, D. Xu, W. Fan, Z. Qian, and X. Xie, "Neural multi-scale self-supervised registration for echocardiogram dense tracking," arXiv preprint arXiv:1906.07357, 2019.

[37] F. T. Dezaki, T. Ginsberg, C. Luong, H. Vaseli, R. Rohling, K. Gin, P. Abolmaesumi, and T. Tsang, "Echo-rhythm net: Semi-supervised learning for automatic detection of atrial fibrillation in echocardiography," IEEE International Symposium on Biomedical Imaging, pp. 110–113, 2021.

[38] A. Chartsias, S. Gao, A. Mumith, J. Oliveira, K. Bhatia, B. Kainz, and A. Beqiri, "Contrastive learning for view classification of echocardiograms," in International Workshop on Advances in Simplifying Medical Ultrasound. Springer, pp. 149–158, 2021.

[39] K. Wickstrøm, M. Kampffmeyer, K. Ø. Mikalsen, and R. Jenssen, "Mixing up contrastive learning: Self-supervised representation learning for time series," Pattern Recognition Letters, vol. 155, pp. 54–61, 2022.

[40] W. Dai, X. Li, W. H. K. Chiu, M. D. Kuo, and K.-T. Cheng, "Adaptive contrast for image regression in computer-aided disease assessment," IEEE Transactions on Medical Imaging, vol. 41, no. 5, pp. 1255–1268, 2021.