Is Contrastive Learning Suitable for Left Ventricular Segmentation in Echocardiographic Images?

Mohamed Saeed[*] MOHAMED.SAEED@MBZUAI.AC.AE
Rand Muhtaseb[*] RAND.MUHTASEB@MBZUAI.AC.AE
Mohammad Yaqub MOHAMMAD.YAQUB@MBZUAI.AC.AE
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Editors: Under Review for MIDL 2022

Abstract

Contrastive learning has proven useful in many applications where access to labelled data is limited. The lack of annotated data is particularly problematic in medical image segmentation as it is difficult to have clinical experts manually annotate large volumes of data. One such task is the segmentation of cardiac structures in ultrasound images of the heart. In this paper, we argue whether or not contrastive pretraining is helpful for the segmentation of the left ventricle in echocardiography images. Furthermore, we study the effect of this on two segmentation networks, DeepLabV3, as well as the commonly used segmentation network, UNet. Our results show that contrastive pretraining helps improve the performance on left ventricle segmentation, particularly when annotated data is scarce.We show how to achieve comparable results to state-of-the-art fully supervised algorithms when we train our models in a self-supervised fashion followed by fine-tuning on just 5% of the data.We also show that our solution achieves better results than what is currently published on a large public dataset (EchoNet-Dynamic) and we compare the performance of our solution on another smaller dataset (CAMUS) as well.

Keywords: Contrastive learning, segmentation, echocardiography, ultrasound, SimCLR, BYOL, self-supervised

1. Introduction

Echocardiography is a valuable diagnostic tool in cardiovascular disease as it can rapidly locate the presence of any abnormalities within the heart. This involves the quantification of heart structures such as the left ventricle. However, there is a lot of room for error in this process due to factors such as human variability or low image quality as ultrasound images are often very noisy. (Alsharqi et al., 2018)

Deep learning solutions can help automate the annotation process, but they are limited by the quantity and quality of labelled training data which can be difficult to obtain. For the problem of left ventricle segmentation in particular, previous works have had some success but there is room for improvement, potentially with the acquisition of more annotated data (Kusunose et al., 2019). However, self-supervision helps in bridging this gap by making use of unlabelled data that does not require input from clinical experts. In similar tasks such as view classification of echocardiography images, contrastive pretraining on unlabelled data showed impressive improvements in results (Chartsias et al., 2021). This indicates potential utility for segmentation problems given that the features learned for classification should not be too dissimilar.

[*] Contributed equally

2. Related Work

In this section, we aim to give a brief revisit to important concepts which our paper investigates. We believe this is important to make our work clearer to a wide audience.

2.1. Segmentation Networks

We investigate two well-known segmentation networks, UNet (Ronneberger et al., 2015) and DeepLabV3 (Chen et al., 2017), which have demonstrated huge success in many segmentation problems and this is why we have chosen them. UNet is a fully convolutional network that consists of a contracting path (encoder) and an expanding path (decoder) in a U-shaped architecture. Features are extracted by the contracting path and then upsampled gradually by the expanding path, with skip connections between corresponding layers in the contracting and expanding paths. The second network is DeepLabV3 which had initially shown great performance on semantic segmentation of natural images. It introduces an atrous spatial pyramid pooling (ASPP) module that utilizes atrous (dilated) convolutions at different rates to solve the problem of object scale variations in addition to expanding the receptive field while keeping the feature maps' spatial dimensions. ASPP consists of multiple dilated convolutions at different rates stacked in parallel followed by a concatenation of the outputs of said convolutions. Features from the encoder are passed through the ASPP module before upsampling back to the original resolution. In the following subsection, we review the use of these two networks in echocardiographic left ventricle segmentation.

2.2. Ventricular Segmentation

One example pertaining to the use of deep learning in ventricular segmentation employed a UNet to segment the left ventricle in more than 1500 images from ultrasound videos of 100 patients. The network was trained on the output of another segmentation algorithm that used a Kalman filter. Expert annotation was only available for 52 of the images, so the dataset was expanded by automatically annotating more examples using the Kalman filter based algorithm. Consequently, the UNet trained on this data was able to achieve a Dice score of 0.87, outperforming the previous algorithm. (Smistad et al., 2017)

Later work by (Moradi et al., 2019) proposed a modification to the UNet architecture by combining it with a feature pyramid network. This was trained for left ventricle segmentation on the publicly available CAMUS dataset (Leclerc et al., 2019) which consists of two- and four-chamber ultrasound images from 500 patients. Testing was then done on an external dataset of 137 four-chamber view images. Results showed that this architecture outperformed other state-of-the-art methods, achieving a Dice score of 0.953 on the test set.

Furthermore, (Ouyang et al., 2020) attempted the same task, training on their large publicly available EchoNet-Dynamic dataset (Ouyang et al., 2019a), containing 20,060 annotated images from 10,030 patients. A DeepLabV3 (Chen et al., 2017) network was chosen for this task and obtained a Dice score of 0.9211.

2.3. Contrastive Learning

Whilst there are multiple published contrastive learning algorithms in the literature, we have chosen to investigate two commonly used ones, namely SimCLR and BYOL.

2.3.1. SimCLR

SimCLR (Chen et al., 2020) is a popular framework for self-supervised contrastive learning, used to learn representations from unlabelled data. In essence, SimCLR creates two augmented versions of every input image. For each minibatch, one pair of augmented images coming from the same original image is chosen as the positive pair. All other pairs coming from different input images are considered negative pairs. The aim then becomes to maximize the agreement within the positive pair while simultaneously maximizing the disagreement between the positive pair and all the negative pairs. The framework begins with a base encoder which is a typical feature extractor such as a ResNet-50 (He et al., 2016). A projection head is added on top of this to map the encoded representation to a new space in which a contrastive loss based on cosine similarity is applied.

2.3.2. BYOL

Meanwhile, Bootstrap Your Own Latent (BYOL) (Grill et al., 2020) uses a similar contrastive approach to SimCLR but without negative pairs and this is why we chose it to compare the effect of this difference on the contrastive pretraining. It always uses a single pair of images which are transformed versions of the same input. The framework allows representation learning by making use of two networks (called the *online* and *target* network). The online network is trained to predict the output of the target network. Meanwhile, the target network's weights are just an exponential moving average of the online network. The two networks are mostly identical, having an encoder (usually a ResNet-50), followed by a projection head which linearly projects the encoder's features onto a different space. The only difference is that the online network has an added predictor head, which is simply another linear projection. During training, the online network learns by attempting to maximize the agreement between the outputs from the two networks by minimizing a contrastive loss which simplifies to twice the negative of the cosine similarity between the two networks' outputs.

3. Methods

In this paper, we developed a solution to segment the left ventricle in echocardiography images that is based on self-supervised contrastive learning. We argue why this could be a better approach than full supervision. This section describes the used data, the setup and the conducted experiments.

3.1. Datasets

3.1.1. EchoNet-Dynamic

The EchoNet-Dynamic dataset (Ouyang et al., 2019b) consists of 10,036 videos of apical four-chamber (A4C) view for patients who had echocardiography between 2016 and 2018 at Stanford Health Care. Each video consists of a sequence of 112 x 112 2D grayscale images extracted from the Digital Imaging and Communications In Medicine (DICOM) file and labeled with the corresponding left ventricle tracing, ejection fraction (EF), volume at end-systole (ES) and volume at end-diastole (ED) by expert sonographers. For each video, two frames (ES and ED) are annotated with manual segmentation. To the best of our knowledge this is currently the largest publicly available dataset for left ventricle segmentation, making it ideal for our contrastive task, given that it has a large amount of both labelled and unlabelled data.

3.1.2. CAMUS

The Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) dataset (Leclerc et al., 2019) contains scans of 500 patients who underwent echocardiography at the University Hospital of St Etienne in France. Each patient's data is labelled with the corresponding left ventricle ejection fraction (EF) and volumes at end-systole (ES) and end-diastole (ED). Annotations include tracings of the left ventricle endocardium, myocardium and left atrium (LA) for both apical two-chamber (A2C) and apical four-chamber (A4C) views of the heart. Training and testing sets consist of 450 annotated and 50 unannotated videos, respectively. We found that 50 patients are missing from the training set, resulting in data of only 400 patients for the training set. We have chosen this small dataset to investigate the importance of contrastive learning when having limited data.

3.2. Experimental setup

We experiment with SimCLR and BYOL pretraining *(pretext task)* for left ventricle segmentation on the EchoNet-Dynamic and CAMUS datasets. First, we pretrained a DeepLabV3 backbone (ResNet-50 with atrous convolutions (Chen et al., 2017)) and a UNet backbone (original UNet encoder) with both SimCLR and BYOL. For the pretraining, unlabelled frames from the datasets are used. Thereafter, the pretrained backbones were used to train the segmentation networks, DeepLabV3 and UNet *(downstream task)*. The downstream segmentation experiments were done with 100%, 50% 25% and 5% of the available labelled data. In addition, we compare the SimCLR and BYOL pretrained backbones to randomly initialized and ImageNet pretrained (fully supervised) ones to see if self-supervision is beneficial. For evaluation, the Dice similarity coefficient (DSC) is used as a metric.

$$\mathrm{DSC} = 2 * \frac{\text{intersection}}{\text{intersection} + \text{union}} \tag{1}$$

All images were resized to 224x224 pixels for both the pretext and downstream tasks. Bilinear interpolation was used for the input images and nearest neighbour interpolation was used for the masks.

Pretext task: All backbones were pretrained for 300 epochs on two NVIDIA A6000 GPUs. DeepLabV3 backbones were trained with a batch size of 128 (64 per device) and UNet backbones were trained with a batch size of 256 (128 per device). An Adam optimizer was used for the pretraining with a learning rate of 1e-3 for SimCLR and 0.2 for BYOL. These were chosen experimentally. For both SimCLR and BYOL, we use the augmentation strategy proposed in the SimCLR paper to see if these contrastive learning algorithms work out of the box for ventricular segmentation. The augmentations consist of random resized cropping, color distortions and Gaussian blurring.

Downstream task: The segmentation tasks were performed on a single A6000 GPU with a batch size of 128. A Madgrad (Defazio and Jelassi, 2021) optimizer was used because it was found to converge better and faster than other optimizers and hyperparameters were selected experimentally. The base learning rate was 1e-4 for DeepLabV3 experiments and 1e-5 for UNet experiments.

3.3. EchoNet-Dynamic Experiments

The two annotated ES and ED frames from every video were used for the downstream task, resulting in 14,920 images for training, 2,576 for validation and 2,552 for testing. This is the same setup as the original EchoNet-Dynamic paper to allow a fair comparison. Meanwhile, for pretraining, the unlabelled frames in between ES and ED were used. One random frame between ES and ED was used for each patient. This was done to avoid having frames that are too similar to each other. As a result, the pretraining training set consisted of 7460 images, and the validation set contained 1288 images.

3.4. CAMUS Experiments

For the downstream task, the 400 available *annotated* videos were split into 300 for training, 50 for validation and 50 for testing. Two frames (ES and ED) were taken from each video, leading to a training set of 600 images, a validation set of 100 images and a testing set of 100 images. For pretraining, a random frame (not including ES or ED frame) was taken from each of the 300 training videos. In addition, to create a validation set, a random frame from each of the videos in the *unannotated* CAMUS test set was used. These are samples from the held out CAMUS test set that were not used anywhere else in our experiments. Overall, the pretraining task used 300 training images and 50 validation images.

Table 1: Summary of experiments conducted on the EchoNet-Dynamic Dataset with different fractions of data for the downstream task

No.	**Pretraining**	**Network**	**Dice (100%)**	**Dice (50%)**	**Dice (25%)**	**Dice (5%)**
1	-	DeepLabV3	0.9204	0.9164	0.9090	0.8920
2	ImageNet	DeepLabV3	0.9229	0.9175	0.9142	0.8968
3	SimCLR	DeepLabV3	**0.9252**	**0.9242**	**0.9190**	**0.9125**
4	BYOL	DeepLabV3	0.9209	0.9042	0.8938	0.8816
5	-	UNet	0.9151	0.9100	0.9046	0.8915
6	SimCLR	UNet	0.9185	0.9157	0.9078	0.9048
7	BYOL	UNet	0.9070	0.8959	0.8768	0.8318

Table 2: Summary of experiments conducted on the CAMUS Dataset with different fractions of data for the downstream task

No.	Pretraining	Network	Dice (100%)	Dice (50%)	Dice (25%)	Dice (5%)
1	-	DeepLabV3	0.9095	0.8941	0.8731	0.7803
2	ImageNet	DeepLabV3	0.9286	0.9217	0.9120	0.8539
3	SimCLR (C)	DeepLabV3	0.9105	0.8862	0.8851	0.8450
4	SimCLR (E)	DeepLabV3	**0.9311**	0.9219	0.9234	**0.9123**
5	BYOL (C)	DeepLabV3	0.8189	0.6202	0.5727	0.0084
6	BYOL (E)	DeepLabV3	0.8347	0.7552	0.6321	0.5729
7	-	UNet	0.9125	0.8921	0.8883	0.8006
8	SimCLR (C)	UNet	0.9102	0.8965	0.8597	0.8013
9	SimCLR (E)	UNet	0.9296	**0.9224**	**0.9248**	0.9077
10	BYOL (C)	UNet	0.8162	0.7810	0.7063	0.0520
11	BYOL (E)	UNet	0.8824	0.8366	0.7984	0.7256

(E): Pretrained on EchoNet data,* *(C):*** *Pretrained on CAMUS data*

4. Results

Tables 1 and 2 show the quantitative results of the experiments that were conducted, for the EchoNet and CAMUS datasets respectively. Qualitative results from selected models are also shown in Figure 1.

4.1. EchoNet-Dynamic

As Table 1 shows, DeepLabV3 with a SimCLR pretrained backbone outperformed all other methods (including the EchoNet-Dynamic (Ouyang et al., 2019b) baseline - 0.9211 dice), regardless of the amount of data. In fact, with only 5% of the data, SimCLR produces results (0.9125 dice) that are close to fully supervised training with all of the available data (0.9229 dice). Futhermore, with the UNet architecture, SimCLR was found to be beneficial although the improvement was minor. We also found ImageNet pretraining to perform better than random initialization. SimCLR aside, BYOL did not have any significant benefit over ImageNet pretraining or even random initialization. Furthermore, DeepLabV3 performed better than UNet in the segmentation task.

4.2. CAMUS

Results on the CAMUS dataset are shown in Table 2. When pretrained on CAMUS, SimCLR backbones (0.9105 dice) were found to perform worse than ImageNet pretrained backbones (0.9286 dice). However, SimCLR backbones pretrained on the EchoNet dataset showed better performance (0.9311 dice), exceeding both random initialization and ImageNet pretrained backbones. This was the case for both DeepLabV3 and UNet. Meanwhile, BYOL backbones continued to show worse performance on the CAMUS dataset as well, especially when pretrained on the CAMUS dataset itself. When finetuned on only 5% of the

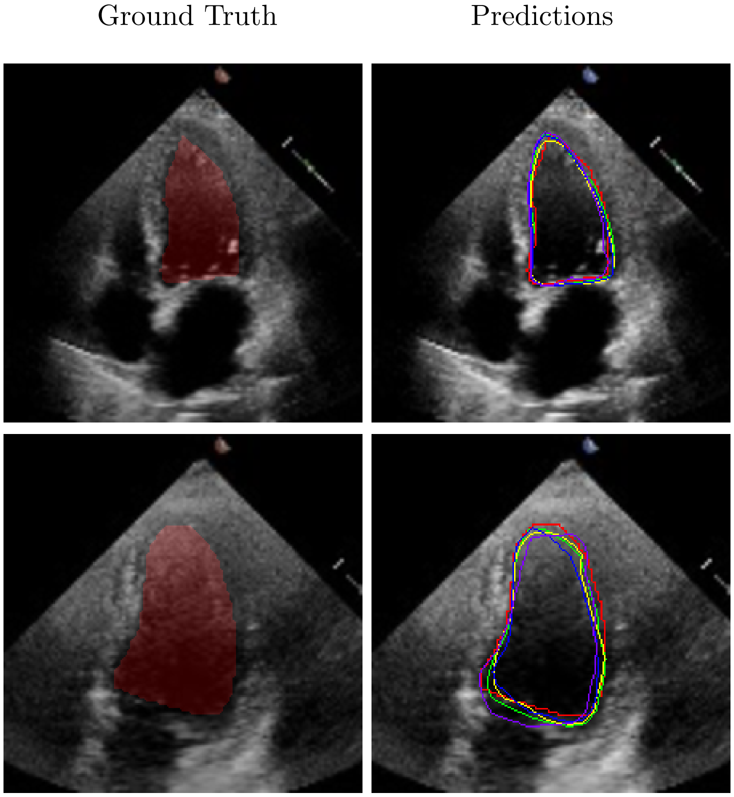

Figure 1: Qualitative results using an ImageNet backbone with 100% data (blue), a SimCLR backbone with 100% data (green), a SimCLR backbone with 5% data (yellow) with DeepLabV3 and a BYOL backbone with 100% data (purple) with DeepLabV3 on the EchoNet-Dynamic dataset for two cases. **Top:** Case where predictions are close to the ground truth (red). **Bottom:** More difficult case where predictions have more discrepancy.

data, these backbones showed extremely poor performance, failing the downstream segmentation task. Pretraining with EchoNet improved the BYOL backbones, which achieved a dice score of up to 0.7256 when finetuned on 5% of the data and up to 0.8824 when finetuned on 100% of the data.

5. Discussion

The experiments have shown that SimCLR outperforms BYOL when it comes to pretraining backbones for left ventricle echocardiography segmentation. We also noticed that BYOL is less stable than SimCLR. However, the purpose of the experiments was to study the use of these models with minimal changes and see how they perform out-of-the-box, without extensive tuning. This may be part of the reason why BYOL has shown suboptimal performance.

The main difference between the two frameworks is the fact that BYOL only uses positive pairs, trying to maximize agreement between two augmented versions of a single image (positive pair) and hence find a common representation for them. Conversely, SimCLR tries to maximize agreement between the differently augmented versions of an image (positive pair) while also maximizing disagreement between that image and augmented versions of other images (negative pairs). Meanwhile, BYOL's contrastive learning is implicit and indirectly dependent on the differences between the original images. In our experiments, we use a random frame from each video to introduce some dissimilarity between the original images but it seems like this is not as effective as the transformations that SimCLR uses.

Furthermore, constrastive learning requires large amounts of data to produce good results, which is why pretraining on the CAMUS dataset with only 400 samples was not beneficial (rows 3 & 6 in Table 2). Hence, it makes sense for EchoNet-Dynamic pretraining to be more beneficial and its capability to work with a different dataset shows that generalizable features were learned from the pretraining (rows 4 & 9 in Table 2).

Apart from contrastive learning, the experiments also suggest that DeepLabV3 is more effective than UNet for echocardiography segmentation (compare rows 3 & 6 in Table 1). In general, what makes DeepLabV3 perform well is its atrous spatial pyramid pooling module that captures multi-scale representations of high level features extracted by the encoder, making it more resistant to changes in object scales - in this case the size of heart structures -, which do vary depending on the heart cycle and the anatomy of the patient's heart.

6. Conclusion

While contrastive learning is an open research problem, we conclude from our experiments that vanilla SimCLR pretraining could lead to an improvement in cardiac ultrasound segmentation, especially when annotated data for the downstream task is limited. However, it is crucial to pretrain on a large enough dataset to provide good results. Further experimentation could lead to a better understanding of contrastive learning frameworks in the context of cardiac ultrasound imaging. For example, there is room for improvement on the augmentation strategy used in the SimCLR paper because it was targeted at natural images not medical images. Optimizing this choice might lead to more significant improvements and is a good direction for future work in this area. Additionally, both SimCLR and BYOL are sensitive to batch sizes and require very large batch sizes for optimal performance. Regardless, our work has shown that SimCLR does work with minimal changes and moderate resources.

References

Maryam Alsharqi, WJ Woodward, JA Mumith, DC Markham, Ross Upton, and Paul Leeson. Artificial intelligence and echocardiography. *Echo research and practice*, 5(4):R115–R125, 2018.

Agisilaos Chartsias, Shan Gao, Angela Mumith, Jorge Oliveira, Kanwal Bhatia, Bernhard Kainz, and Arian Beqiri. Contrastive learning for view classification of echocardiograms.

In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 149–158. Springer, 2021.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Aaron Defazio and Samy Jelassi. Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization. *arXiv preprint arXiv:2101.11075*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kenya Kusunose, Akihiro Haga, Takashi Abe, and Masataka Sata. Utilization of artificial intelligence in echocardiography. *Circulation Journal*, pages CJ–19, 2019.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

Shakiba Moradi, Mostafa Ghelich Oghli, Azin Alizadehasl, Isaac Shiri, Niki Oveisi, Mehrdad Oveisi, Majid Maleki, and Jan Dhooge. Mfp-unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. *Physica Medica*, 67:58–69, 2019.

David Ouyang, Bryan He, Amirata Ghorbani, Matt P Lungren, Euan A Ashley, David H Liang, and James Y Zou. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, 2019a.

David Ouyang, Bryan He, Amirata Ghorbani, Matthew P. Lungren, Euan A. Ashley, David H. Liang, and James Y. Zou. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. 2019b.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Erik Smistad, Andreas Østvik, et al. 2d left ventricle segmentation using deep learning. In *2017 IEEE international ultrasonics symposium (IUS)*, pages 1–4. IEEE, 2017.