# IDEAL: Improved DEnse LocAL Contrastive Learning for Semi-Supervised Medical Image Segmentation

**Hritam Basak[1], Soumitri Chattopadhyay[*2], Rohit Kundu[*3], Sayan Nag[*4], Rammohan Mallipeddi[5]**

[1] Department of Computer Science, Stony Brook University
[2] Department of Information Technology, Jadavpur University
[3] Department of Electrical and Computer Engineering, UC Riverside
[4] Department of Medical Biophysics, University of Toronto
[5] Department of Artificial Intelligence, School of Electronics, Kyungpook National University

**PAPER ID : 538**

Project Page: https://rohit-kundu.github.io/IDEAL-ICASSP23/

# Background


In the biomedical domain, it is difficult to acquire large quantities of annotated data as the annotations are to be done by trained medical professionals.

Self-supervised pre-training alleviates the annotation problem by utilizing unlabeled data to learn distinctive information which can further be utilized in downstream applications.

In medical image segmentation, contrastive self-supervised learning has been leveraged naively in a few prior works, which being a global function fails to capture spatially local features.

# Motivation

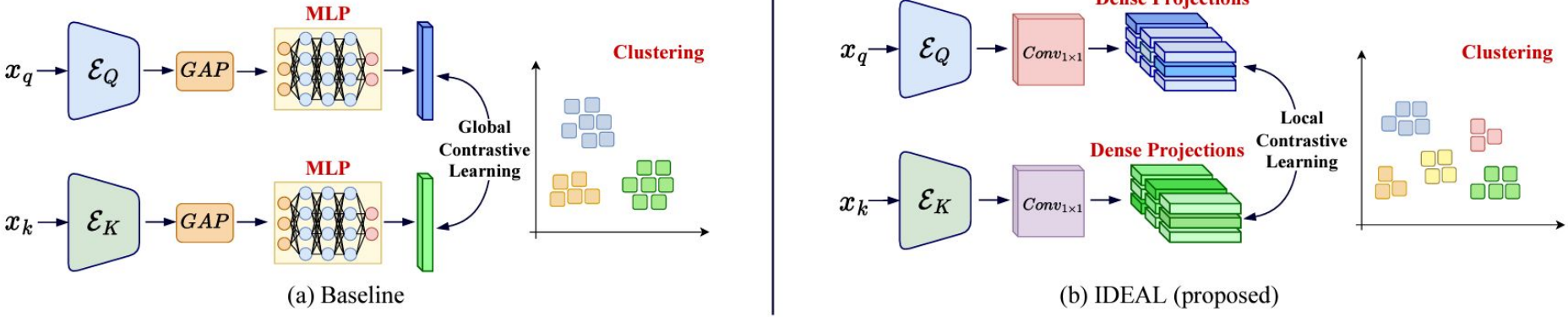



Fig.(a) shows Chaitanya *et al.*'s baseline work employing global CL that learns global representations. Contrary to this, our proposed model utilises dense projections to learn local spatial semantics, as shown in Fig.(b).

We hypothesize that learning local features aid pixel-level tasks such as segmentation, whereas global features are suitable for classification tasks.

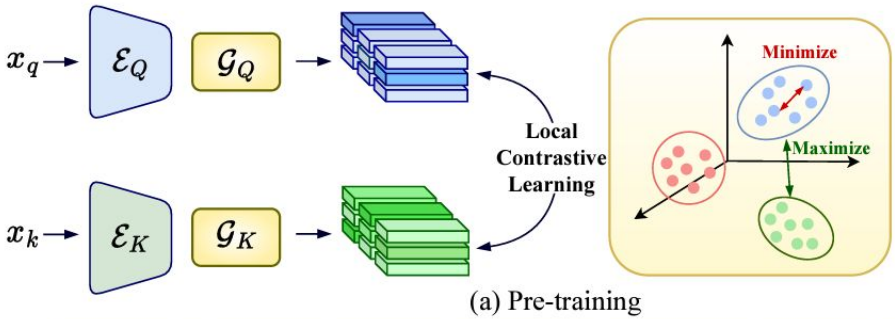
# Contributions

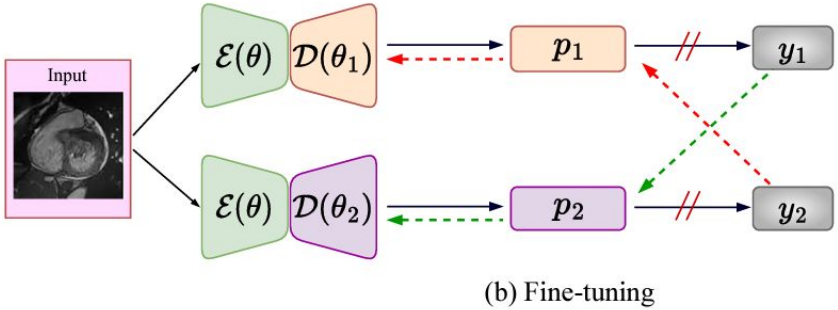

1. We propose a self-supervised strategy that leverages dense projection head representations for contrastive learning for learning robust local features.

2. We re-define ’positive’ and ’negative’ samples in contrastive learning and extend the InfoNCE loss to adapt to dense representations during the pre-training phase.

3. Our work also employs a uniquely devised cross-consistency regularization for fine-tuning the network to the downstream segmentation task.

# Proposed Method



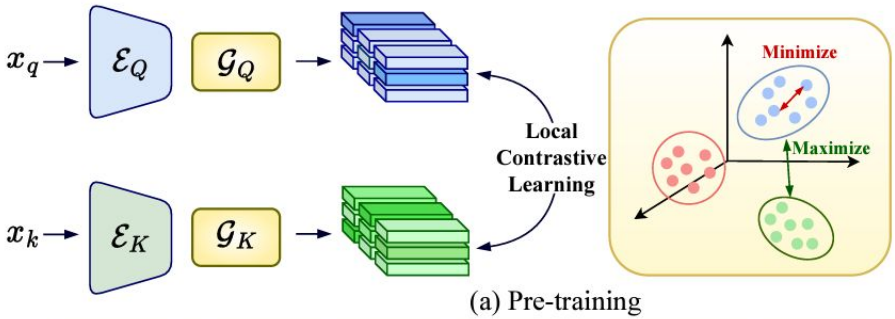

// ⟹ Stop gradient flow through this path  ⟹ Backpropagation path through first model  ⟹ Backpropagation path through second model

$$\mathcal{L}_{loc} = \frac{-1}{D_h D_w} \sum_{i=1}^{D_h D_w} \log \frac{\exp(q_i \cdot k_i^+ / \tau)}{\exp(q_i \cdot k_i^+ / \tau) + \sum_{k_i^-} \exp(q_i \cdot k_i^- / \tau)}$$

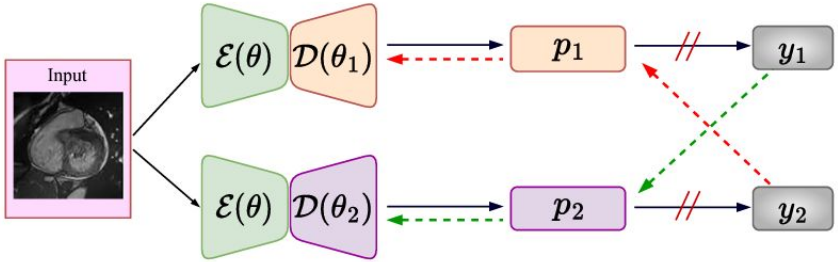
$$\mathcal{L}_{SSF} = \begin{cases} CE(p_1, m) + CE(p_2, m), & \text{for labeled set} \\ CE(p_1, y_2) + CE(p_2, y_1), & \text{for unlabeled set} \end{cases}$$

In the pre-training phase, the projection head shown employs a 1x1 convolution layer instead of a traditional MLP for dense feature extraction, which are used for computing the local contrastive loss, a modified version of the InfoNCE objective.

# Proposed Method




(a) Pre-training




(b) Fine-tuning

// ⟹ Stop gradient flow through this path  ⟹ Backpropagation path through first model  ⟹ Backpropagation path through second model

$$\mathcal{L}_{loc} = \frac{-1}{D_h D_w} \sum_{i=1}^{D_h D_w} \log \frac{\exp(q_i \cdot k_i^+ / \tau)}{\exp(q_i \cdot k_i^+ / \tau) + \sum_{k_i^-} \exp(q_i \cdot k_i^- / \tau)}$$

$$\mathcal{L}_{SSF} = \begin{cases} CE(p_1, m) + CE(p_2, m), & \text{for labeled set} \\ CE(p_1, y_2) + CE(p_2, y_1), & \text{for unlabeled set} \end{cases}$$

In fine-tuning phase, the pre-trained encoder is shared with two differently perturbed decoder branches. The predicted output segmentation maps are thresholded to obtain the respective outputs for which gradients flow in the alternate branch, thus enforcing cross-consistency in segmentation.

# Experimental Setup


**Datasets:** (1) ACDC dataset, comprising 100 cardiac 3D shot-axis MRI volumes with annotations for three structures; (2) MMWHS dataset, consisting of 20 cardiac 3D MRI volumes with annotations for seven structures.

**Pre-training:** 78(train)+2(valid) for ACDC; 8(train)+2(valid) for MMWHS
**Evaluation:** 20(test) for ACDC; 10(test) for MMWHS

**Implementation:** ResNet50 encoder is used with U-Net decoders, one with ConvTranspose layers and the other with Upsampling layers (bilinear interpolation). Adam optimizer is used throughout, with a learning rate of 1e-5.

**Hardware configuration:** NVIDIA Tesla K80 GPU and Linux OS.

**Evaluation metrics:** Dice Similarity Coefficient (DSC), Average Symmetric Distance (ASD), Hausdorff Distance (HD)

# Results


**Performance under limited annotations:**

Following existing works, we experimented with different label percentages (denoted by L in the table below) for the respective datasets, during the semi-supervised fine-tuning phase.

| Metric | ACDC | | | | MMWHS | | | |
|---|---|---|---|---|---|---|---|---|
| | L=1.25% | L=2.5% | L=10% | L=100% | L=10% | L=20% | L=40% | L=100% |
| ASD ↓ | 0.677 | 0.614 | 0.556 | 0.489 | 2.397 | 1.798 | 1.355 | 1.317 |
| HD ↓ | 2.409 | 2.209 | 2.094 | 1.999 | 5.001 | 3.499 | 2.221 | 2.183 |
| DSC ↑ | 0.738 | 0.846 | 0.879 | 0.882 | 0.626 | 0.791 | 0.815 | 0.826 |

Even with a fraction of labels available (10% for ACDC and 40% for MMWHS),the respective semi-supervised setups asymptotically approach the fully supervised setup (differs by 0.3% in ACDC and 1.1% in MMWHS).

# Results

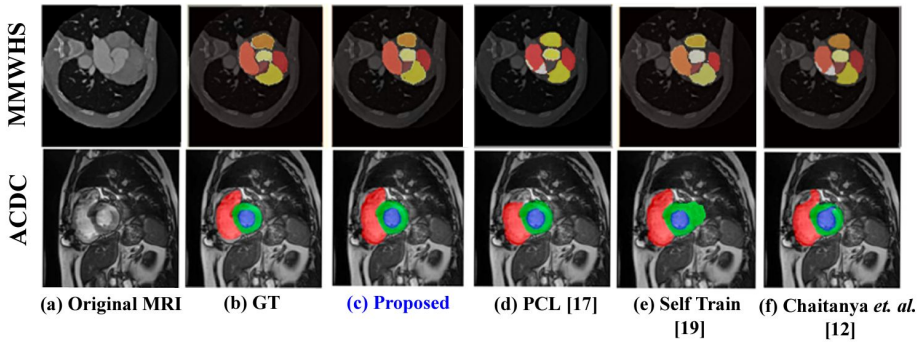

**Comparison to state-of-the-art:**

**ACDC:** IDEAL performs best in L=1.25% setting,while being highly competitive to PCL in L=2.5%(~0.004) and L=10%(~0.006).

**MMWHS:** IDEAL outperforms all previous frameworks by large margins, followed by Chaitanya *et al*.

Qualitative comparison of outputs in adjacent figure validates reliability of the proposed model.

| Method | Average DSC (ACDC) | | | Average DSC (MMWHS) | | |
|---|---|---|---|---|---|---|
| | L=1.25% | L=2.5% | L=10% | L=10% | L=20% | L=40% |
| Chaitanya *et. al.* [13] | 0.725 | 0.789 | 0.872 | 0.569 | 0.694 | 0.794 |
| Global CL [6] | 0.729 | – | 0.847 | 0.500 | 0.659 | 0.785 |
| PCL [18] | 0.671 | **0.850** | **0.885** | – | – | – |
| Context Restoration [10] | 0.625 | 0.714 | 0.851 | 0.482 | 0.654 | 0.783 |
| MC-Net [28] | 0.677 | 0.724 | 0.855 | 0.551 | 0.654 | 0.798 |
| Label Efficient [27] | – | – | – | 0.382 | 0.553 | 0.764 |
| Data Augmentation [21] | 0.731 | 0.786 | 0.865 | 0.529 | 0.661 | 0.785 |
| Self Train [20] | 0.690 | 0.749 | 0.860 | 0.563 | 0.691 | 0.801 |
| Ours | **0.738** | 0.846 | 0.879 | **0.626** | **0.791** | **0.815** |



# Conclusion & Future Work


1. IDEAL leverages contrastive learning with a novel projection head to capture dense local feature representation during pre-training, followed by employing a unique cross-consistency regularization scheme during fine-tuning for the downstream segmentation task.

2. Our framework outperforms several state-of-the-art methods on two widely used cardiac MRI datasets.

3. IDEAL can also be employed in cross-modal and cross-domain scenarios, as well as extended to natural images, something we intend to explore in near future.

# References


1. Chaitanya *et al.* Contrastive learning of global and local features for medical image segmentation with limited annotations, NeurIPS, 2020.
2. Chen *et al.* A simple framework for contrastive learning of visual representations, ICML, 2020.
3. He *et al.* Momentum contrast for unsupervised visual representation learning, CVPR, 2020.
4. Zeng *et al.* Positional contrastive learning for volumetric medical image segmentation, MICCAI, 2021.

Hritam Basak

Soumitri Chattopadhyay

Rohit Kundu

Sayan Nag

Rammohan Mallipeddi

# Thank You!