Dense Self-Supervised Learning for Medical Image Segmentation

Maxime Seince[1,2] MAXIME.SEINCE22@IMPERIAL.AC.UK
Loïc Le Folgoc[1] LOIC.LEFOLGOC@TELECOM-PARIS.FR
Luiz Augusto Facury de Souza[1] LUIZ.FACURYDESOUZA@TELECOM-PARIS.FR
Elsa Angelini[1] ELSA.ANGELINI@TELECOM-PARIS.FR
[1] *LTCI, Telecom Paris, Institut Polytechnique de Paris, France*
[2] *Imperial College London, UK*

Abstract

Deep learning has revolutionized medical image segmentation, but it relies heavily on high-quality annotations. The time, cost and expertise required to label images at the pixel-level for each new task has slowed down widespread adoption of the paradigm. We propose `Pix2Rep`, a self-supervised learning (SSL) approach for few-shot segmentation, that reduces the manual annotation burden by learning powerful pixel-level representations directly from unlabeled images. `Pix2Rep` is a novel pixel-level loss and pre-training paradigm for contrastive SSL on whole images. It is applied to generic encoder-decoder deep learning backbones (*e.g.*, U-Net). Whereas most SSL methods enforce invariance of the learned *image-level* representations under intensity and spatial image augmentations, `Pix2Rep` enforces *equi*variance of the *pixel-level* representations. We demonstrate the framework on a task of cardiac MRI segmentation. Results show improved performance compared to existing semi- and self-supervised approaches; and a 5-fold reduction in the annotation burden for equivalent performance versus a fully supervised U-Net baseline. This includes a 30% (resp. 31%) DICE improvement for one-shot segmentation under linear-probing (resp. fine-tuning). Finally, we also integrate the novel `Pix2Rep` concept with the Barlow Twins non-contrastive SSL, which leads to even better segmentation performance.

Keywords: Deep Learning, Segmentation, Self-Supervised Learning, Representation Learning, Cardiac MRI

1. Introduction

Medical image segmentation has seen tremendous progress with the advent of deep learning (Ronneberger et al., 2015; Milletari et al., 2016; Kamnitsas et al., 2017). The drawback of this paradigm is its reliance on large quantities of data, annotated at the pixel-level, to train strong segmentation models. These pixel-level annotations are costly to obtain, and take precious time from medical experts.

To circumvent this burden, techniques have emerged in recent years that better exploit more widely available *un*labeled data. Semi-supervised approaches *e.g.*, pseudo-labels (Lee, 2013; Bai et al., 2017; Tran et al., 2022) and mean teacher (Tarvainen and Valpola, 2017; Yu et al., 2019), balance a supervised segmentation loss on a small labeled dataset with a consistency loss on the larger unlabeled dataset, yielding improved segmentation. Other semi-supervised approaches include Bayesian deep learning *e.g.*, Dalca et al. (2018) introduce anatomical priors that can be learnt using unlabeled or unpaired data.

Self-Supervised Learning (SSL) follows an alternative route whereby deep representations are directly learned from unlabeled data. Early methods trained these representations by solving pretext tasks *e.g.*, relative position prediction (Doersch et al., 2015), image recolorization (Zhang et al., 2016), jigsaw puzzles (Noroozi and Favaro, 2016) or rotation prediction (Gidaris et al., 2018). These methods are designed for image classification as a primary downstream task, thus an image is encoded to an image-level vector representation. Many recent methods for image-level representation learning coexist in the state-of-the-art, based on contrastive learning (Chen et al., 2020; He et al., 2020), on redundancy-reduction (Zbontar et al., 2021), on self-distillation (Grill et al., 2020; Caron et al., 2021), on Masked Image Modeling (He et al., 2022) and many more.

We propose instead **a framework for pixel-level (dense) representation learning**, dubbed `Pix2Rep`, which can be used to **pretrain encoder-decoder architectures**, such as U-Nets. Whereas most aforementioned methods rely on invariance under certain intensity-based augmentations (brightness & contrast, Gaussian noise, etc.) and geometric augmentations (crops), `Pix2Rep` is based on *equi*variance under geometric transformations. For the task of cardiac MRI segmentation, **we propose rotations and intensity reversals as additional augmentations** that further improve results. Finally, **we investigate the performance** of `Pix2Rep` **in various data regimes** (one-shot, few-shot segmentation, or large annotated data), both under linear probing and fine-tuning.

2. Related Work

Comparatively, fewer pixel-level representation learning methods have been proposed so far. Kalapos and Gyires-Tóth (2023); Punn and Agarwal (2022) propose to pretrain a U-Net encoder (a.k.a. its downsampling branch) using image-level SSL (BYOL/Barlow Twins). The U-Net decoder however is randomly initialized before fine-tuning on the downstream segmentation task. Tang et al. (2022) pretrain a Swin UNETR encoder using a combination of image-level contrastive learning, pretext task and masked image modeling. Zeng et al. (2021) exploit the positional information of slices within stacks for contrastive pretraining of a U-Net encoder.

Chaitanya et al. (2020) manage pretraining of the first decoder layers by introducing a local contrastive loss that relies on rough alignment of subject volumes. For contrastive pretraining of the whole decoder (Xie et al., 2021), positive pairs of *pixels* need to be defined. Zhong et al. (2021) constrain the two augmented views to differ only up to intensity transformations, so as to form positive pairs from pixels at identical locations in the two views. Hu et al. (2021); Zhao et al. (2021) regard as positive all pixels sharing the same label, at the cost of pretraining only on the (potentially smaller amount of) labeled data. Wang et al. (2021); Bardes et al. (2022) define pixels with highly similar features as positives. The closest related work to our proposed approach are those of O. Pinheiro et al. (2020); Yan et al. (2022); Goncharov et al. (2023), which define positive pairs to be pixels that describe the same physical location in the scene on different augmented patches, that differ up to intensity-based and geometric augmentations. Our framework targets equivariance rather than equivalence to random spatial transforms and works at whole image level for augmentations. O. Pinheiro et al. (2020) experiment on natural –not medical– images. Goncharov et al. (2023); Yan et al. (2022) focus less on few-shot segmentation.

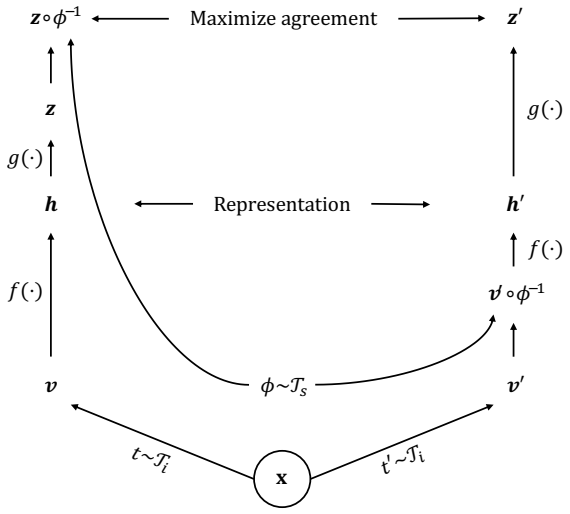

Figure 1: Pretraining of arbitrary encoder-decoder architectures f (*e.g.*, U-Net). $\mathbf{x}$ an unlabeled training image; $\phi \sim \mathcal{T}_s$ a random spatial transformation; $t, t' \sim \mathcal{T}_i$ two random intensity transformations; g a projection head. We train pixel representation maps output by f to be equivariant under ϕ and invariant to t, t' by maximizing agreement between the outputs of the two branches, via a pixel-level contrastive loss.

3. Methods

We consider an arbitrary (trainable) neural network backbone $f : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{H \times W \times D}$ that maps input images to pixel representation maps. f can be any encoder-decoder network and we opt for simplicity for a U-Net (Ronneberger et al., 2015), without its final segmentation head ($\equiv$ typically 1×1 conv + softmax)[1]. f can be interpreted as embedding C-dimensional input pixels to D-dimensional vector representations, accounting for local and global context. At the core of the proposed self-supervised pretraining of f is a contrastive loss which forces the pixel representations derived from f to be invariant under the action of intensity augmentations and equivariant under the action of spatial transformations. The method is summarized in Fig. 1.

`Pix2Rep` pretraining relies exclusively on an unlabeled dataset $\mathcal{D} \triangleq \{\mathbf{x} \in \mathbb{R}^{H \times W \times C}\}$. Apart from the encoder-decoder $f(\cdot)$ which extracts pixel-level representation maps, the framework consists of three major components.

(1) For any input image $\mathbf{x}$, a stochastic data augmentation module generates two random intensity-based transformations $t, t' \sim \mathcal{T}_i$ (incl. brightness & contrast augmentation, Gaussian noise, bias field, and intensity reversal), resulting in two views $\mathbf{v} \triangleq t(\mathbf{x}), \mathbf{v}' \triangleq t'(\mathbf{x})$. In addition, the stochastic data augmentation module generates a single random spatial

1. For downstream segmentation tasks, the segmentation head $\mathbb{R}^{H \times W \times D} \mapsto \mathbb{R}^{H \times W \times K}$, with K the number of classes, is plugged back at the end of the pretrained backbone

transformation $\phi \sim \mathcal{T}_s$ (incl. zooms, flips, rotations), which is applied asymmetrically to $\mathbf{v}, \mathbf{v}'$.

(2) A small projection head $g : \mathbb{R}^{H\times W\times D} \mapsto \mathbb{R}^{H\times W\times d}$ transfers pixel representation maps to the space where the contrastive loss is applied. The projection head $g(\cdot) \triangleq W^{(2)} * \sigma(W^{(1)} * \cdot)$ consists in a MLP with one hidden layer (where σ is a ReLu non-linearity), implemented using 1×1 convolutions. We have experimented with deeper projection heads (up to 3 hidden layers) without noticing significant benefits. Then:

- For the view $\mathbf{v}'$: we first transform it to the new spatial viewpoint by applying ϕ, yielding[2] $\phi \cdot \mathbf{v}' \triangleq \mathbf{v}' \circ \phi^{-1} \in \mathbb{R}^{H\times W\times C}$, then pass the spatially-transformed image through $g \circ f$, yielding $\mathbf{z}' \triangleq (g \circ f)(\phi \cdot \mathbf{v}')$.
- For the view $\mathbf{v}$: we first apply $g\circ f$, yielding the pixel representation map $\mathbf{z} = (g\circ f)(\mathbf{v})$, before transporting $\mathbf{z}$ to the new viewpoint: $\phi \cdot \mathbf{z} = \phi \cdot (g \circ f)(\mathbf{v})$.

Note that ϕ is applied exactly once along the two computational branches, hence $\phi \cdot \mathbf{z} = \phi \cdot (g \circ f)(t(\mathbf{x}))$ and $\mathbf{z}' = (g \circ f)(\phi \cdot t'(\mathbf{x}))$ share the same viewpoint.

(3) Thirdly, a contrastive loss $\mathcal{L}$ is defined for a pixel-level contrastive prediction task. Given any pixel coordinate $s \in \mathbb{R}^2$, features $(\phi \cdot \mathbf{z})(s) \in \mathbb{R}^d$ and $\mathbf{z}'(s) \in \mathbb{R}^d$ correspond to the same anatomical point in the two views, thus they form natural positive pairs. Natural negative pairs are obtained from samples from the same view at all other pixel coordinates, or from other images in the minibatch at any pixel coordinate. However, creating one positive pair per pixel coordinate would yield hundreds of thousands of positive pairs and potentially billions of negative pairs.

Instead, we sample M random pixel coordinates $\{s_m^{(n)}\}_{m=1\ldots M}$ independently for each image $1 \leq n \leq N_b$ in the minibatch. For a given image $\mathbf{x}$ and a given coordinate s, we obtain a positive pair of examples $\mathbf{u}_i \triangleq (\phi\cdot\mathbf{z})(s)$, $\mathbf{u}_j \triangleq \mathbf{z}'(s)$. The negative examples for this positive pair $(\mathbf{u}_i, \mathbf{u}_j)$ are obtained from the other $N_bM - 1$ pairs of positive examples across all sampling coordinates and images in the minibatch. In total, this generates $N \triangleq N_bM$ positive pairs and $2(N_bM - 1)$ negative examples for each positive pair. By analogy to SimCLR (Chen et al., 2020), we use the InfoNCE loss $l_{i,j}$ of Eq. (1) for each positive pair $(\mathbf{u}_i, \mathbf{u}_j)$:

$$l_{i,j} \triangleq -\log \frac{\exp\left(\mathrm{sim}(\mathbf{u}_i, \mathbf{u}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}[k \neq i] \exp\left(\mathrm{sim}(\mathbf{u}_i, \mathbf{u}_k)/\tau\right)} \ , \tag{1}$$

where τ denotes a temperature parameter, and $\mathrm{sim}(\mathbf{u}, \mathbf{u}') \triangleq \mathbf{u}^T\mathbf{u}'/\|\mathbf{u}\|_2\|\mathbf{u}'\|_2$ denotes the cosine similarity. The total loss is aggregated by summing over all $l_{i,j}$, including symmetrizing the roles of i, j. Assuming without loss of generality that positive pairs have consecutive indices $i = 2k - 1$ and $j = 2k$ in the list of sampled pixels, this yields Eq. (2):

$$\mathcal{L} \triangleq \frac{1}{2N} \sum_{k=1}^{N} \left(l_{2k-1,2k} + l_{2k,2k-1}\right) \ . \tag{2}$$

2. The action $\phi\cdot\mathbf{v}'$ of a spatial transformation ϕ on an image $\mathbf{v}'$ is $\mathbf{v}'\circ\phi^{-1}$. This is well-known in the image registration literature. Following standard practice, we randomly sample ϕ^{-1} directly (rather than ϕ) to circumvent the numerical inversion.

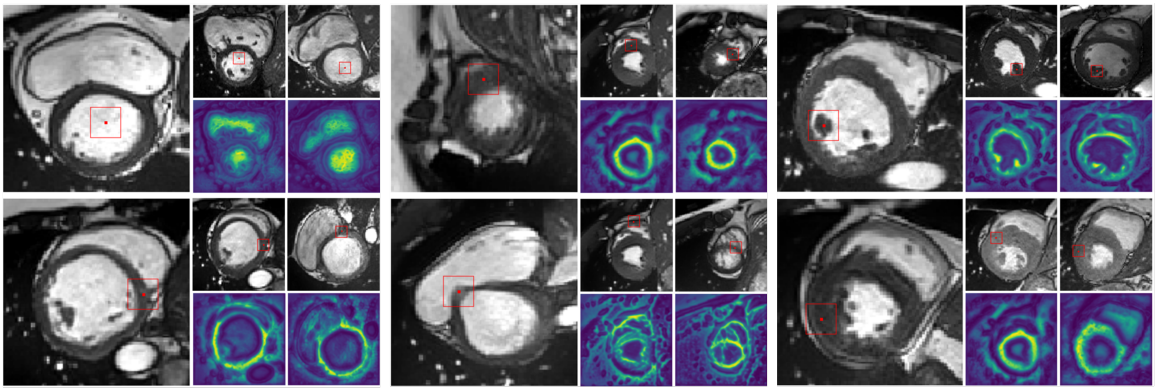

Figure 2: Pixel embedding similarity maps. Large images: query images in which we select a query pixel (highlighted in red). For each query, we display two test images, with the pixel closest (in embedding space) to the query pixel highlighted in red. Similarity maps (cosine similarity between pixel embeddings) are also shown.

Pix2Rep-v2. The computational complexity of contrasting positive and negative pairs of pixels limits us to sample M coordinates for the InfoNCE contrastive loss. Alternatively, we propose to replace this contrastive loss with a loss based on `Barlow Twins` (Zbontar et al., 2021): we call this variant `Pix2Rep-v2`. It minimizes the `Barlow Twins` loss defined from the cross-correlation matrix $\mathcal{C} \in \mathbb{R}^{d \times d}$ between twin pixel embeddings $(\phi \cdot \mathbf{z})(s)$ and $\mathbf{z}'(s) \in \mathbb{R}^d$, aggregated over all pixel coordinates s and the whole minibatch. Although it aggregates information from the whole pixel representation maps (rather than samples), `Pix2Rep-v2` has a reduced memory footprint compared to `Pix2Rep`'s contrastive loss.

Downstream segmentation. For a given segmentation task, we initialize the encoder-decoder f with the pretrained weights (discarding the projection head g), and add a task-specific, learnable segmentation head (1×1 conv + softmax), projecting pixel representations to class probabilities. We keep f frozen, and only train the segmentation head (called **linear probing**), or allow f to be fine-tuned from the supervised data (called **fine-tuning**).

4. Experiments

We demonstrate the self-supervised pretraining on a downstream task of cardiac MRI segmentation.

Data. The ACDC dataset (Bernard et al., 2018) consists of $3D$ short-axis cardiac cine MR images of 150 subjects, including expert annotations at End-Systole and End-Diastole for the left ventricle, right ventricle and myocardium. It is split into a training-validation set (100 images) and a test set (50 images). Slices are intensity-normalized using min-max normalization (using the 1st and 99th percentiles), cropped and resized to 128×128.

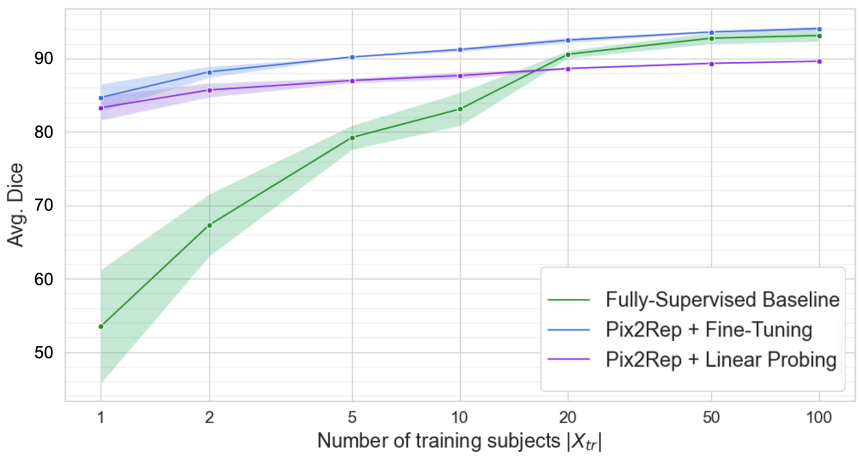

Figure 3: Proposed pretraining vs. fully-supervised baseline (same U-Net architecture).

Experimental setup and Evaluation. We use the provided split between training data and test data. The test data ($|X_{ts}| = 50$) is only used for the final evaluation. For pretraining (via `Pix2Rep` or other approaches), we use the entirety of the raw training data ($|X_{pre}| = 100$), without segmentation labels. For linear probing, fine-tuning, or fully-supervised training, we reuse a smaller number of training images with their segmentation labels ($|X_{tr}| \in \{1, 2, 5, 10, 20, 50, 100\}$). The slices in the stacks of X_{tr} are divided between training data (90%) and validation data (10%).

We quantify the contribution of the `Pix2Rep` pretraining on the performance on a downstream segmentation task. The $3D$ Dice similarity coefficient is used as the evaluation metric. We report the average Dice over the test set over the segmented structures. For each evaluated method, each reported score is an average over five runs (training+test).

Comparison to other methods. A natural baseline is obtained by keeping the same backbone U-Net, initialized with random (rather than pretrained) weights, then trained using varying amounts of labeled data X_{tr} ($|X_{tr}| \in \{1, 2, 5, 10, 20, 50, 100\}$). We refer to it as the fully-supervised baseline.

In addition, we compare with several methods in the state-of-the-art. We implement a mean teacher (Tarvainen and Valpola, 2017) semi-supervised model (using the same U-Net backbone), where the supervised loss applies to the available labeled slices, and the unsupervised consistency loss uses all slices from the 100 raw training volumes. As for self-supervised models, we compare to (1) encoder-only pretraining of the same U-Net backbone, using image-level SimCLR, representative of Kalapos and Gyires-Tóth (2023); Punn and Agarwal (2022); (2) encoder-decoder (`Pix2Rep`) pretraining without rotations (crops only) and without intensity reversals as augmentations; and (3) the method of Chaitanya et al. (2020). Unless stating otherwise, all self-supervised approaches use fine-tuning in the second stage rather than linear probing, as this yields higher performance. For the proposed approach, we report numbers both for linear-probing and fine-tuning.

Backbone and Implementation details. We use a standard U-Net architecture consisting of five `DoubleConv` encoding blocks ({`Conv,BatchNorm,ReLu,Conv,BatchNorm,ReLu`}) with `MaxPooling` downsampling and five `DoubleConv` decoding blocks with `Trans-Conv` upsampling, starting with 128 feature maps and doubling after each block up to 2048 at the bottleneck. We experimented with changing the number of feature maps in the last layer (cf. ablation study), starting at 64, noting a monotonic improvement in performance up to the maximum tested value of 1024. We report the performance for $n_{ft} := 1024$. For the fully-supervised baseline instead, we noticed decreased performance above and below $n_{ft} := 128$ and report performance for this setting.

The framework is implemented in `PyTorch` (https://github.com/pix2rep/). The spatial transformation ϕ is applied in auto-differentiable manner via `affine_grid`, `grid_sample`. Intensity reversal augmentations correspond to $\tilde{\mathbf{x}} := 1 - \mathbf{x}$. We pretrain using the proposed approach for 200 epochs with the `Adam` optimizer. The batch size is set to 8, the number of sampled pixels to 1000. The learning rate for the backbone is set to $5 \cdot 10^{-4}$. For fine-tuning (100 epochs), the learning rate of the backbone is set to $5 \cdot 10^{-5}$ and of the classification head to 10^{-2}. For linear probing (100 epochs), the learning rate is set to 10^{-2}.

Results. To gain qualitative insight into the learnt pixel representations, Fig. 2 plots the cosine similarity between a query pixel embedding and other pixel embeddings in two other representative images. The semantics of the anatomical structures are captured to some extent, as pixels anatomically similar to the query pixel have the highest similarity with it.

	Avg. Dice (ACDC), for $\lvert X_{tr}\rvert$ set to:						
Method: / $\lvert X_{tr}\rvert :=$	1	2	5	10	20	50	100
Fully-supervised learning:							
Baseline (U-Net)	53.5	67.3	79.2	83.1	90.6	92.8	93.1
Semi-supervised learning:							
Mean Teacher	57.3	68.9	84.5	89.0	90.3	92.4	93.9
Self-supervised learning (+ Fine-tuning by default):							
SimCLR pretraining	63.3	77.8	85.0	88.9	90.2	92.1	92.6
Chaitanya et al. (2020)	76.7	78.0	85.9	88.7	90.8	92.2	92.7
Proposed (w/o rot. & int. reversal)	77.7	80.4	87.9	90.2	92.2	93.3	**94.3**
Proposed (only Linear-probing)	83.3	85.7	87.0	87.7	88.6	89.3	89.6
Proposed	<u>84.7</u>	<u>88.2</u>	<u>90.2</u>	<u>91.2</u>	<u>92.5</u>	<u>93.6</u>	94.1
Combined method:							
Proposed + Mean Teacher	**86.1**	**89.7**	**91.1**	**91.9**	**92.7**	**93.8**	<u>94.2</u>

Table 1: Comparison with the state-of-the-art on the test ACDC dataset (avg. Dice score), for various amounts of labeled training data $\lvert X_{tr}\rvert$. Best results in **bold** and second best <u>underlined</u>. Each score is averaged over five runs (see main text – Evaluation).

Table 1 summarizes the main results. The proposed dense contrastive pretraining yields higher Dice than the fully-supervised baseline for all data regimes, as also shown in Fig. 3.

The gap is largest for few-shot segmentation; there is a 29.8% (resp. 31.2%) improvement in Dice for one-shot segmentation in linear probing (resp. fine-tuning). Of note is the limited drop in performance with linear probing even in the large data regime, highlighting the quality of the unsupervised nonlinear pixel representations. With fine-tuning, there is roughly a 5× reduction in the annotation burden for equivalent performance vs. the fully supervised baseline. Pretraining also reduces the variability in performance resulting from the choice of training subject, as shown by lower standard deviations in Fig. 3.

Ablation study. Table 2 shows the incremental contributions of various modifications on the proposed framework, starting from a suboptimal setup, in the few-shot segmentation setting (for $|X_{tr}| = 5$). The inclusion of image intensity reversals (*i.e.*, $\tilde{\mathbf{x}} := 1 - \mathbf{x}$) and rotations as augmentations has a significant impact on performance.

`Pix2Rep` **pretraining settings:**	**Avg. Dice**
Defaults w/ 10 epochs pretraining	39.7 ± 1.0
10 epochs → 100 epochs	51.2 ± 2.7
64 feature maps → 1024 feature maps	73.4 ± 0.9
Add 'intensity reversal' augmentation	82.5 ± 3.0
Add rotation augmentation	84.5 ± 1.0
Linear-probing → Fine-tuning	89.2 ± 1.3

Table 2: Ablation study ($|X_{tr}| = 5$). Self-supervised pretraining benefits from the increased number of feature maps as well as from the additional augmentations.

Pix2Rep-v2. Table 3 reports results obtained with the non-contrastive variant `Pix2Rep-v2`, which generally improves slightly over `Pix2Rep`.

Method: / $\lvert X_{tr}\rvert :=$	1	2	5	10	20	50	100
w/ Linear probing	82.6	85.0	**88.1**	**89.1**	**89.8**	**90.4**	**90.7**
w/ Fine-tuning	**85.6**	**88.4**	**91.1**	**92.0**	**92.6**	93.4	94.0

Table 3: Avg. Dice score (ACDC dataset), for the proposed `Pix2Rep-v2` approach. Numbers in **bold** when above their `Pix2Rep` counterpart.

5. Discussion & Conclusion

We have introduced `Pix2Rep`, a novel framework for pixel-level (dense) self-supervised representation learning, that allows to pretrain encoder-decoder architectures such as U-Nets directly from unlabeled images. We have shown performance gains on a downstream cardiac MRI segmentation task. Especially in the few-shot segmentation regime for the most challenging structure, we got 83% Dice with 5 training subjects vs. 70% for the fully-supervised baseline, and closest SOTA method 3% below. As future work, in addition to comparing to Yan et al. (2022), we plan to evaluate the framework on various segmentation tasks and assess the generalizability of learned representations for novel tasks (foundation model).

References

Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, pages 253–260. Springer, 2017.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8799–8810. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39cee562b91611c16ac0b100f0bc1ea1-Paper-Conference.pdf.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. doi: 10.1109/TMI.2018.2837502.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/949686ecef4ee20a62d16b4a2d7ccca3-Paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Adrian V. Dalca, John Guttag, and Mert R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018.

Mikhail Goncharov, Vera Soboleva, Anvar Kurmukov, Maxim Pisov, and Mikhail Belyaev. vox2vec: A framework for self-supervised contrastive learning of voxel-level representations in medical images. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 605–614, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43907-0.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

Xinrong Hu, Dewen Zeng, Xiaowei Xu, and Yiyu Shi. Semi-supervised contrastive learning for label-efficient medical image segmentation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 481–490, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87196-3.

András Kalapos and Bálint Gyires-Tóth. Self-supervised pretraining for 2d medical image segmentation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 472–484, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-25082-8.

Konstantinos Kamnitsas, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.10.004. URL https://www.sciencedirect.com/science/article/pii/S1361841516301839.

Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 7 2013.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016. doi: 10.1109/3DV.2016.79.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46466-4.

Pedro O O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4489–4500. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3000311ca56a1cb93397bc676c0b7fff-Paper.pdf.

Narinder Singh Punn and Sonali Agarwal. Bt-unet: A self-supervised learning framework for biomedical image segmentation using barlow twins with u-net models. *Machine Learning*, 111:4585–4600, 2022. ISSN 1573-0565. doi: 10.1007/s10994-022-06219-3. URL https://doi.org/10.1007/s10994-022-06219-3.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20730–20740, June 2022.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

Manuel Tran, Sophia J. Wagner, Melanie Boxberg, and Tingying Peng. S5cl: Unifying fully-supervised, self-supervised, and semi-supervised learning through hierarchical contrastive learning. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 99–108, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16434-7.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3024–3033, June 2021.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16684–16693, June 2021.

Ke Yan, Jinzheng Cai, Dakai Jin, Shun Miao, Dazhou Guo, Adam P. Harrison, Youbao Tang, Jing Xiao, Jingjing Lu, and Le Lu. Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Transactions on Medical Imaging*, 41(10):2658–2669, 2022. doi: 10.1109/TMI.2022.3169003.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 605–613, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32245-8.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhuang, Jingtong Hu, and Yiyu Shi. Positional contrastive learning for volumetric medical image segmentation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 221–230, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87196-3.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46487-9.

Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10623–10633, October 2021.

Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7273–7282, October 2021.

Appendix A. Additional Results

Pix2Rep-v2 + Mean Teacher. As a combined method, we have also tested the combination of pretraining via `Pix2Rep-v2` with semi-supervised fine-tuning (Mean Teacher). Results are provided in Table 4:

Method: / $\lvert X_{tr}\rvert :=$	1	2	5	10	20	50	100
`Pix2Rep-v2` + Mean Teacher	**87.9**	**90.0**	**91.8**	**92.4**	**92.7**	93.4	93.9

Table 4: Performance on the ACDC dataset of `Pix2Rep-v2` followed by (Mean Teacher) semi-supervised fine-tuning. Results are highlighted in **bold** when above the results reported in Table 1.

Appendix B. Dice Scores per Anatomical Structures

Table 1 reports the test Dice scores averaged over the four segmented structures: Left Ventricle, Right Ventricle, Myocardium and Background. We report the corresponding Dice scores over individual anatomical structures in Tables 5, 6, 7, 8.

	LV Dice (ACDC), for $\lvert X_{tr}\rvert$ set to:						
Method: / $\lvert X_{tr}\rvert :=$	1	2	5	10	20	50	100
Fully-supervised learning:							
Baseline (U-Net)	49.2	70.2	82.7	87.4	93.7	95.0	95.6
Semi-supervised learning:							
Mean Teacher	59.7	75.1	89.9	93.1	94.1	95.5	**96.4**
Self-supervised learning (+ Fine-tuning by default):							
SimCLR pretraining	66.9	83.9	90.5	92.8	93.8	95.2	95.3
Chaitanya et al. (2020)	82.0	80.2	89.5	91.6	94.8	95.7	96.1
Proposed (w/o rot. & int. reversal)	84.1	84.2	92.2	93.9	95.0	95.8	96.2
Proposed (only Linear-probing)	88.6	90.2	91.3	92.0	92.5	93.1	93.4
Proposed	89.9	92.7	93.9	94.3	95.3	95.9	96.3
Combined method:							
Proposed + Mean Teacher	**91.2**	**93.2**	**94.3**	**95.3**	**95.7**	**96.3**	96.1

Table 5: Comparison with the state-of-the-art on the test ACDC dataset, for various amounts of labeled training data $\lvert X_{tr}\rvert$, in terms of Left Ventricle (LV) Dice overlap. Each score is averaged over five runs (see main text – Evaluation).

	RV Dice (ACDC), for $\lvert X_{tr}\rvert$ set to:						
Method: / $\lvert X_{tr}\rvert :=$	1	2	5	10	20	50	100
Fully-supervised learning:							
Baseline (U-Net)	41.1	54.9	69.8	74.5	84.8	88.5	88.8
Semi-supervised learning:							
Mean Teacher	40.9	53.1	74.1	81.2	82.8	86.7	89.3
Self-supervised learning (+ Fine-tuning by default):							
SimCLR pretraining	43.4	68.0	75.2	81.2	83.1	86.2	87.2
Chaitanya et al. (2020)	60.4	61.4	79.7	82.1	83.9	86.6	87.3
Proposed (w/o rot. & int. reversal)	57.9	69.6	80.3	83.8	87.2	89.2	**91.5**
Proposed (only Linear-probing)	71.0	74.7	77.2	78.4	80.1	81.2	81.8
Proposed	74.0	79.1	83.0	85.2	**87.7**	**89.7**	90.5
Combined method:							
Proposed + Mean Teacher	**75.2**	**81.9**	**84.6**	**85.9**	87.5	89.4	90.4

Table 6: Comparison with the state-of-the-art on the test ACDC dataset, for various amounts of labeled training data $\lvert X_{tr}\rvert$, in terms of Right Ventricle (RV) Dice overlap. Each score is averaged over five runs (see main text – Evaluation).

	MYO Dice (ACDC), for $\lvert X_{tr}\rvert$ set to:						
Method: / $\lvert X_{tr}\rvert :=$	1	2	5	10	20	50	100
Fully-supervised learning:							
Baseline (U-Net)	34.7	52.1	69.2	74.8	86.3	89.3	89.9
Semi-supervised learning:							
Mean Teacher	46.0	59.0	78.1	84.3	86.2	89.4	91.3
Self-supervised learning (+ Fine-tuning by default):							
SimCLR pretraining	53.1	64.5	77.8	84.4	86.5	88.9	89.8
Chaitanya et al. (2020)	65.4	71.0	74.8	81.7	84.7	87.1	87.8
Proposed (w/o rot. & int. reversal)	74.7	73.0	82.1	85.6	88.6	90.1	90.9
Proposed (only Linear-probing)	77.4	81.2	82.5	83.2	84.6	85.6	85.9
Proposed	78.3	83.7	86.3	87.5	88.9	90.4	91.0
Combined method:							
Proposed + Mean Teacher	**81.4**	**87.0**	**87.9**	**88.8**	**89.4**	**91.1**	**91.9**

Table 7: Comparison with the state-of-the-art on the test ACDC dataset, for various amounts of labeled training data $\lvert X_{tr}\rvert$, in terms of Myocardium (MYO) Dice overlap. Each score is averaged over five runs (see main text – Evaluation).

	BG Dice (ACDC), for $\lvert X_{tr}\rvert$ set to:						
Method: / $\lvert X_{tr}\rvert :=$	1	2	5	10	20	50	100
Fully-supervised learning:							
Baseline (U-Net)	89.2	92.0	95.3	95.9	97.7	98.3	98.3
Semi-supervised learning:							
Mean Teacher	82.5	88.3	95.9	97.2	97.4	98.1	98.5
Self-supervised learning (+ Fine-tuning by default):							
SimCLR pretraining	89.7	94.9	96.4	97.3	97.5	98.0	98.2
Chaitanya et al. (2020)	**99.2**	**99.3**	**99.5**	**99.6**	**99.7**	**99.7**	**99.7**
Proposed (w/o rot. & int. reversal)	94.3	95.1	96.8	97.5	98.1	98.4	98.6
Proposed (only Linear-probing)	96.2	96.7	97.1	97.1	97.3	97.5	97.5
Proposed	96.6	97.2	97.6	97.9	98.2	98.5	98.6
Combined method:							
Proposed + Mean Teacher	96.6	96.9	97.5	97.8	98.1	98.3	98.3

Table 8: Comparison with the state-of-the-art on the test ACDC dataset, for various amounts of labeled training data $\lvert X_{tr}\rvert$, in terms of Dice overlap for the background (BG). Each score is averaged over five runs (see main text – Evaluation).

Appendix C. Visualization of the `Pix2Rep` Pixel Embeddings

To gain qualitative insights into the learned pixel representations, Fig. 4 graphically presents pixel embeddings returned by our pretrained `Pix2Rep` model, learned without supervision and *prior* to the fine-tuning stage of the downstream segmentation task.

Firstly, we show 2D `t-SNE` projections of all pixel embeddings for three test images, color-coded by their true class (background, left ventricle, right ventricle, myocardium). We can see that the four anatomical structures are well separated in the representation space, despite not using any label supervision.

Secondly, we assign to each 2D `t-SNE` coordinate positions a color, according to a reference colormap shown at the top right of Fig. 4. Then, we color-code each pixel of each test image by its color as obtained by this scheme (MRI image pixel $\mapsto$ pixel `Pix2Rep` embedding $\mapsto$ pixel embedding mapped to 2D coordinates after `t-SNE` projection $\mapsto$ pixel embedding mapped to an individual color value with a 2D reference colormap applied on the 2D `t-SNE` space $\mapsto$ colored pixel embedding displayed in original MRI image space). We remark that visually, our `Pix2Rep` embedding leads to coarse segmentations of the cardiac structures, again without involving any supervision with label annotations.

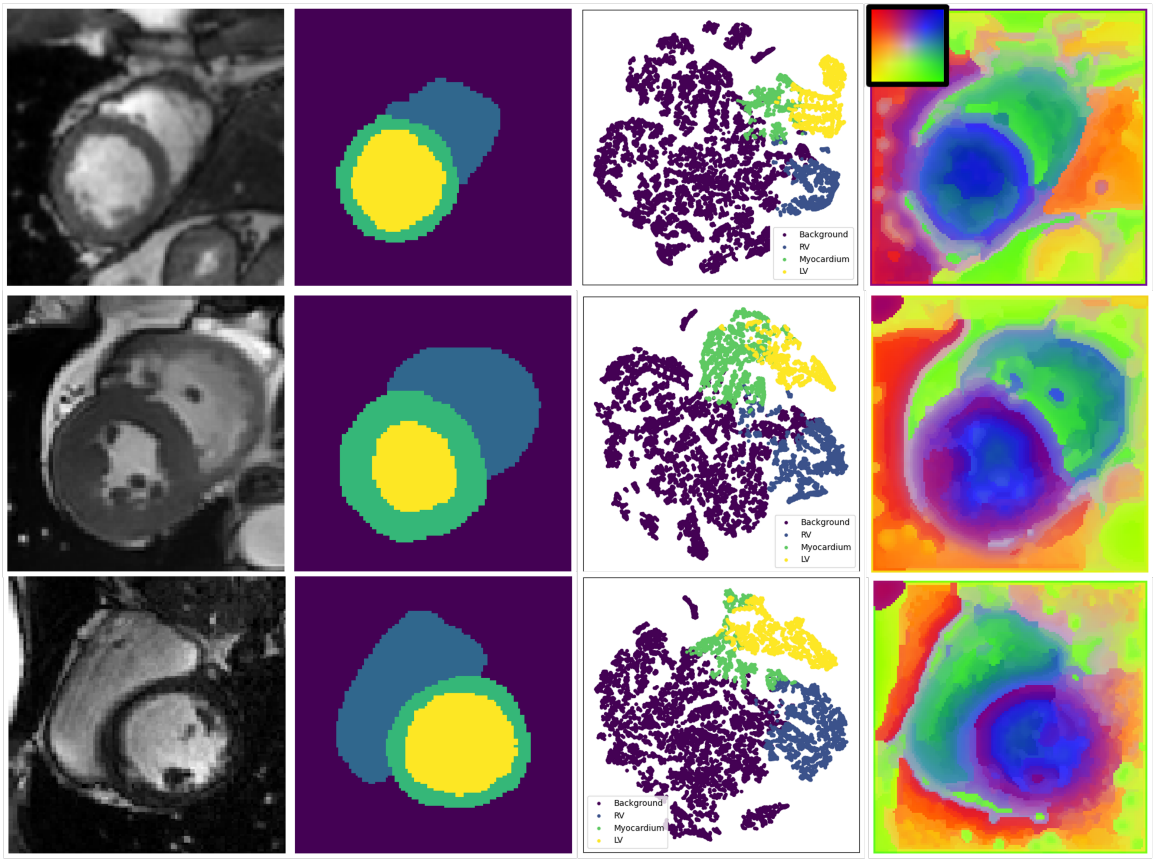

Figure 4: `Pix2Rep` pixel-level embeddings. First and Second columns: test cardiac MRI images and ground truth segmentations. Third column: 2D `t-SNE` coordinates of `Pix2Rep` pixel embeddings. Fourth column: colored pixel embedding displayed in original MRI image space. The reference colormap used to map 2D `t-SNE` coordinates with individual colors is shown in the vignette on the top row example.