Representation Recovering for Self-Supervised Pre-training on Medical Images

Xiangyi Yan Junayed Naushad Shanlin Sun Kun Han Hao Tang
Deying Kong Haoyu Ma Chenyu You Xiaohui Xie
University of California, Irvine Yale University
{xiangyy4, jnaushad, shanlins, khan7, htang6, deyingk, haoyum3, xhx}@uci.edu
chenyu.you@yale.edu

Abstract

Advances in self-supervised learning have drawn attention to developing techniques to extract effective visual representations from unlabeled images. ***Contrastive*** *learning (CL) trains a model to extract consistent features by generating different views. Recent success of Masked Autoencoders (MAE) highlights the benefit of* ***generative*** *modeling in self-supervised learning. The generative approaches encode the input into a compact embedding and empower the model's ability of recovering the original input. However, in our experiments, we found vanilla MAE mainly recovers coarse high level semantic information and is inadequate in recovering detailed low level information. We show that in dense downstream prediction tasks like multi-organ segmentation, directly applying MAE is not ideal. Here, we propose RepRec, a hybrid visual representation learning framework for self-supervised pre-training on large-scale unlabelled medical datasets, which takes advantage of both contrastive and generative modeling. To solve the aforementioned dilemma that MAE encounters, a convolutional encoder is pre-trained to provide low-level feature information, in a contrastive way; and a transformer encoder is pre-trained to produce high level semantic dependency, in a generative way – by recovering masked representations from the convolutional encoder. Extensive experiments on three multi-organ segmentation datasets demonstrate that our method outperforms current state-of-the-art methods.*

1. Introduction

Organ segmentation is an essential step used in many applications, such as diagnostic interventions, treatment planning and delivery. Usually, these image analyses are carried out by experienced doctors. However, it is time-consuming and labor-intensive, since a 3D CT volume can contain up to hundreds of 2D slices. Therefore, developing robust and accurate organ segmentation tools is a fundamental need in medical image analysis. There is a vast volume of work on organ segmentation using computed tomography (CT) [41, 27, 56, 55] or magnetic resonance (MR) [54, 31, 32] images. Traditional segmentation methods are mostly atlas-based. These methods rely on a set of accurate image templates with manual segmentation, and then use image registration to align the new image to the templates. However, these methods may not adequately account for the anatomical variance due to variations in organ shapes, removal of tissues, growth of tumor and differences in image acquisition.

Deep learning-based methods provide an alternative solution with substantial accuracy improvement and speedup, which has been shown effective in many applications, such as detection[40], segmentation [41, 22], registration [42, 24, 23], pose estimation [28, 13, 29], etc. With recent advances in deep learning, automatic segmentation using computer vision algorithms has shown great promise. Various applications have been deployed in clinical practice. However, to train deep learning-based organ segmentation models, large amount of densely annotated images are typically required yet preparing large-scale labeled datasets is expensive and time-consuming. This request becomes even more urgent with the raise of Transformers[17, 1, 34, 33].

A promising solution to the aforementioned issue is self-supervised representation learning, which has shown great success in the field of natural language processing and computer vision, arguably due to its potentials of extracting general and transferable features that apply to various downstream tasks. Compared to supervised learning where manual annotations are naturally used as learning objectives, the key of self-supervised learning is to design some type of pretext tasks so that extracted features satisfy annotation-free constraints. [45]

In computer vision, current self-supervised learning methods can be broadly divided into two main categories, generative modeling and discriminative modeling. In earlier times, discriminative self-supervised pretext tasks are designed as rotation prediction [19], jigsaw solving [37], and relative patch location prediction [16], etc. Recently,

contrastive learning, which belongs to the discriminative branch, achieves great success in self-supervised visual representation learning. The core idea of contrastive learning is to attract different augmented views of the same image and repulse augmented views of different images. Based on this core idea, MoCo [52] and SimCLR [52] are proposed, which greatly shrink the gap between self-supervised learning and fully-supervised learning. The success of MoCo [26] and SimCLR [5] highlights the benefits of contrastive learning. More advanced techniques have emerged recently [8, 12, 6, 21, 35, 57, 58]. However, the aforementioned pre-training strategy are mainly designed for image classification and object detection tasks. To close the gap between self-supervised pre-training and dense prediction tasks such as semantic segmentation and instance segmentation, Wang *et al*. [48] present dense contrastive learning (DenseCL), which implements self-supervised learning by optimizing a pairwise contrastive (dis)similarity loss at the pixel level between two views of input images. In [4], Chaitanya *et al*. propose domain-specific contrastive loss - a local version of the contrastive loss to learn distinctive representations of local regions that are useful for per-pixel segmentation.

Generative modeling also provides a feasible way for self-supervised pre-training [60, 39, 59]. Recently, He *et al*. propose MAE [52] and yield a nontrivial and meaningful generative self-supervisory task, by masking a high proportion of the input image. Transfer learning performance in downstream tasks outperforms supervised pre-training and shows promising scaling behavior. Influenced by the idea of MAE, Wei *et al*. propose MaskFeat [49], which regresses histograms of oriented gradients (HOG), a hand-crafted feature descriptor, of the masked content rather than raw pixels. Compared to pixel color targets, HOG come with less ambiguity in the experimental results. Normalizing gradients handles the color ambiguity and spatial binning of gradients texture ambiguity [49]. SaGe [45] combines both discrimination and generation approaches together by using an encoder to extract visual features into a compact vector, and a decoder to recover the original image based on the compact vector. However, in both [49] and our experimental results, regressing features instead of directly recovering raw pixels provides better representation on the down-stream tasks.

Inspired by the above discussion, we propose RepRec, a hybrid visual representation learning framework for self-supervised pre-training on large scale unlabelled medical datasets, which takes advantages of a contrastive stage to remedy MAE's shortage at fine level information learning. In our method, a convolutional encoder is pre-trained to provide fine level feature extraction, in a contrastive way. Afterwards, a transformer encoder is pre-trained to produce global level semantic dependency, in a generative way – by recovering masked feature maps from the convolutional encoder. Our major contributions are summarized as follows.

- We are the first to leverage both generative and discriminative modeling in large-scale self-supervised learning on medical image segmentation tasks.
- We propose RepRec, a novel generative mechanism for pre-training transformer encoder, by recovering representations from a parameterized network rather than raw images or hand-crafted feature descriptors.
- We conduct extensive experiments on three multi-organ segmentation benchmarks, and demonstrate superior performance of RepRec compared to current self-supervised pre-training approaches.

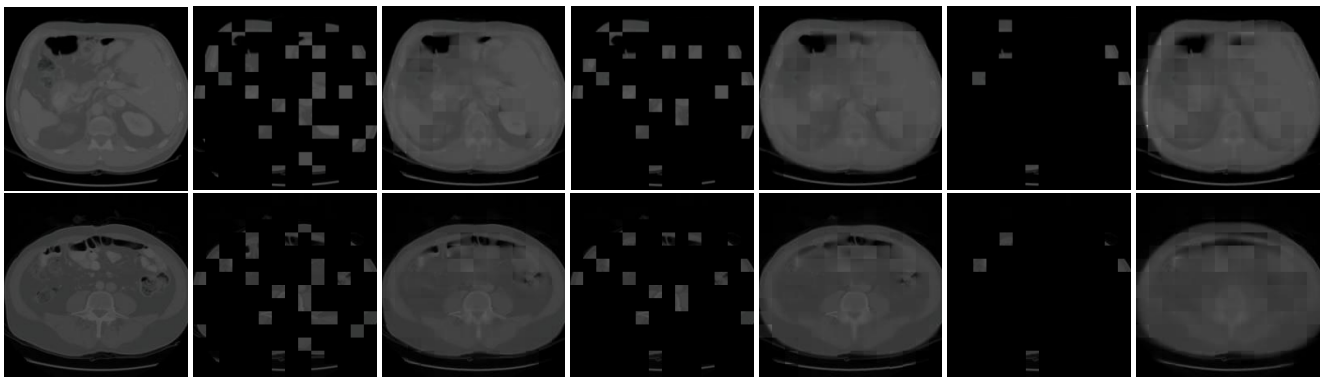
2. Related Work

Self-supervised pre-training approaches can be divided into two main categories: self-supervised discriminative learning and self-supervised generative learning.

2.1. Self-supervised Discriminative Learning

In earlier times, discriminative self-supervised pretext tasks such as rotation prediction [20], Jigsaw solving [38], and relative patch location prediction [15], are designed to learn high-level semantic features.

Recently, contrastive learning as a subcategory of self-supervised discriminative learning has shown its great promise to the community. Wu *et al*.[51] propose InstDisc and memory bank for large scale contrastive learning. The main idea is to use a contrastive loss to pull a query image and its similar samples (positive keys) together and push different ones (negative keys) away. However, in the InstDisc paper, features in memory bank are mostly inconsistent with each other due to the asynchronisation update of model weights and memory bank, which hurts the contrastive learning process [26]. An end-to-end contrastive learning framework was proposed by SimCLR [5], which solves the inconsistency among different key features. However, due to the end-to-end design of SimCLR and GPU memory limit, the model is not able to learn from a large number of negative samples in a single update, which also hurts the contrastive learning process [26]. He *et al*. [26] propose MoCo to use a queue structure to maintain the feature maps and use a momentum encoder to slowly update the key encoder, which largely remits the inconsistency among different key features, meanwhile the model can also learn from numerous negative keys in the queue. Approaches with new techniques, such as projectors[6, 8] are proposed in the following work. Researchers also propose contrastive learning without negative samples [21, 9, 2], with multi-modality [46] and multi-view [44], with ViT [17] and towards stable and larger scale representation learning [12, 3].



(a) Original (b) mask 75% (c) mask 85% (d) mask 95%

Figure 1. Recovered raw images by vanilla MAE, with 75%, 85%, 95% mask ratio respectively. We show that even given very few ratio (25%, 15%, and even 5%) of patches in the original image, MAE is capable of recovering high level semantic information such as the bounding contour of the abdomen, the location and a rough contour of the kidney, spleen, spinal cord, etc. However, when transfer to dense prediction tasks like multi-organ segmentation, such abundant high level information brings marginal benefits. This motivates us to make up the gap between pre-training and downstream tasks.

2.2. Self-supervised Generative Learning

In earlier times, generative pretext tasks like image inpainting[39] and image colorization[59] are proposed to train an auto-encoder for feature extraction. As BERT getting more popular in the domain of NLP, researchers extend the idea of BERT to the field of computer vision [1]. Recently, He *et al*.propose MAE [25] and yield a nontrivial and meaningful generative self-supervisory task, by masking a high proportion of the input image. Xie *et al*. propose SimMIM[53], which also use a similar self-supervisory task. Influenced by the idea of MAE, Wei *et al*.propose MaskFeat [49], which regresses histograms of oriented gradients (HOG), a hand-crafted feature descriptor, of the masked content rather than raw pixels. A series of work are proposed based on the general idea of MAE [7, 18]. However, these former work all perform the masking operation on either raw images or hand crafted features, e.g. histograms of oriented gradients (HOG). We show that the aforementioned approaches recover few detailed information, which is essential for on dense downstream tasks like multi-organ segmentation.

3. Motivation

To examine vanilla MAE's transfer-ability to multi-organ segmentation tasks, we pre-train it on a large scale abdomen dataset Abdomen-1K [36] with the original settings from [25]. Following the protocol in [25] with 75%, 85%, 95% masking ratio, Fig.1 shows the recovered raw images by vanilla MAE, respectively. We found that even given very few ratio (25%, 15%, and even 5%) of patches in the original image, MAE is capable of recovering high level semantic information such as the bounding contour of the whole abdomen region, the location and a rough contour of the kidney, spleen, spinal cord, etc. Although such abundant high level information brings benefits after transferring to downstream classification and detection tasks in the original paper [25], marginal benefit is provided when we evaluate it on downstream **dense** prediction tasks like multi-organ segmentation. See table 1. This motivates us to make up this gap between pre-training and dense downstream tasks.

4. Methodology

Figure 2 sketches the pipeline of RepRec, which includes three stages, contrastive pre-training stage, generative pre-training stage and the fine-tuning stage. We now elaborate the details of each stage in the following subsections.

4.1. Contrastive Pre-training Stage

In the contrastive pre-training stage, we follow the contrastive protocol in [26]. In one batch, an image $\mathbf{x}_q$ is randomly chosen from B images as a query sample, and the rest images $\mathbf{x}_n \in \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_B\}$ are considered as negative key samples, where $n \neq q$. To formulate a positive key sample $\mathbf{x}_p$, elastic transforms are performed on the query sample $\mathbf{x}_q$. Afterwards, three sets of feature maps $\mathbf{f}_q$, $\mathbf{f}_p$, $\mathbf{f}_n$ are extracted by a convolutional encoder $\mathcal{E}_c$ from $\mathbf{x}_q$, $\mathbf{x}_p$, $\mathbf{x}_n$ correspondingly. With similarity measured by dot product, a form of a contrastive loss function, called InfoNCE [47] is considered:

$$\mathcal{L}_c = -\log \frac{\exp\left(\mathbf{f}_q \cdot \mathbf{f}_p/\tau\right)}{\sum_{i=1}^{B} \exp\left(\mathbf{f}_q \cdot \mathbf{f}_i/\tau\right)},$$

where τ is a temperature hyper-parameter per [50].

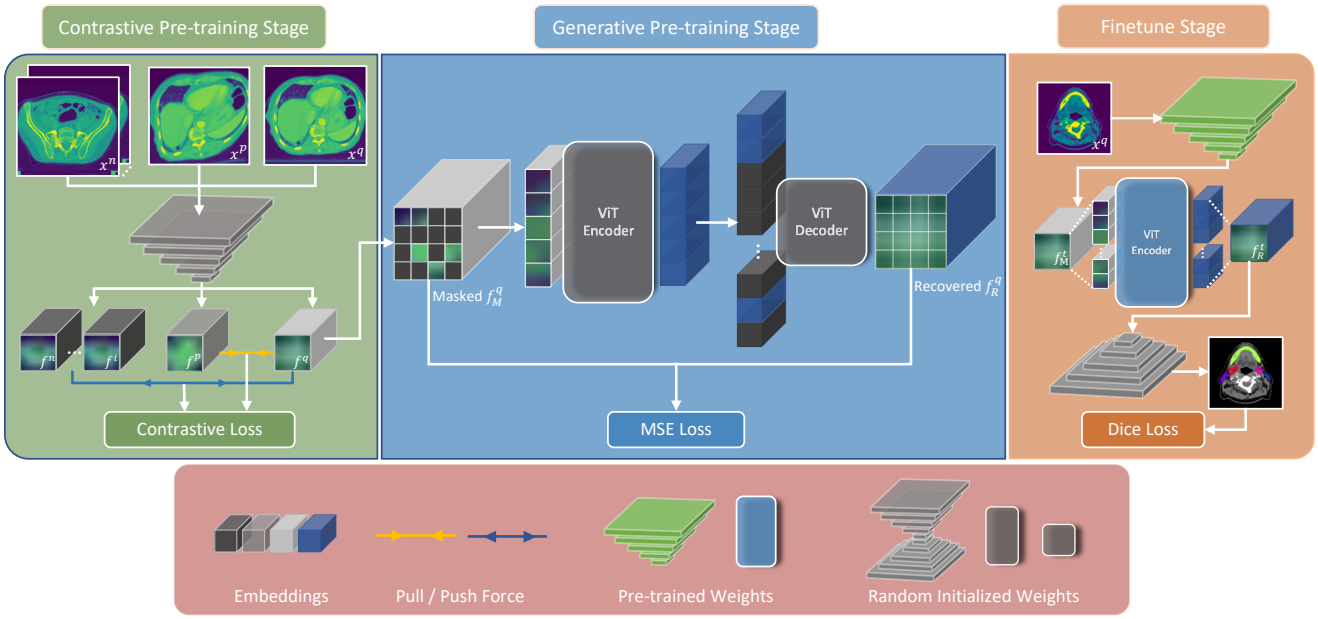


Figure 2. Overview of the RepRec framework. Three stages are included: the contrastive pre-training stage, the generative pre-training stage and the fine-tuning stage. In the contrastive pre-training stage, the weights of a CNN encoder is initialized to provide dense spacial latent information. Then these latent feature maps are split into patches, randomly sampled and forwarded to a ViT encoder. The output of the ViT encoder will be remapped to the original position in the feature map and forwarded to a ViT decoder. After pre-training, we finetune the full model on a small dataset with limited ground truth masks (bottom right) and evaluate the DSC score.

The contrastive pre-training stage is designed for two main purposes. One is to provide learnable feature maps, which will be utilized as inputs in the later generative pre-training stage. The other is to generate different levels of feature maps, which follows the common design of the U-Net [41] model family. These feature maps will be used through skip connections to provide alternative paths for the gradient in the later fine-tuning stage.

4.2. Generative Pre-training Stage

Unlike former work to perform generative pre-training by predicting masked raw images [25] or hand-crafted feature descriptors [49], RepRec directly recovers learnable representation extracted from the convolutional encoder $\mathcal{E}_c$.

Compared to MAE [25], instead of dividing an image into patches, we divide feature maps into patches. Then we randomly sample a subset of patches from the query embedding $\mathbf{f}^q$, following a uniform distribution and mask the remaining ones to form the masked query embedding $\mathbf{f}_M^q$. In order to formulate a non-trivial pre-training task, we perform random sample with a high masking ratio, *i.e.*, over 75%, to eliminate redundancy. The remaining masked patches are then embed by a ViT encoder $\mathcal{E}_t$, which includes N_t Transformer blocks. Afterwards, the encoded visible patches and mask tokens are regrouped and decoded by a ViT decoder $\mathcal{D}_t$ to output the recovered embedding $\mathbf{f}_R^p$. Each mask token [14] is a shared, learned vector that indicates the presence of a missing patch to be predicted. The ViT decoder is only used in the generative pre-training stage to perform the image reconstruction task (only the encoders will be used to produce image representations for further segmentation task). The loss function in this stage $\mathcal{L}_g$ computes the mean squared error (MSE) between the recovered representation $\mathbf{f}^q$ and original representation $\mathbf{f}_R^q$.

The RepRec mechanism follows the design of MAE, which means masked patches are removed and no mask tokens are used. Compared to [1], it allows us to train large encoders with only a fraction of compute and memory. Furthermore, instead of directly recovering the masked raw images, RepRec recovers the feature maps which takes a even lower usage of compute and memory.

4.3. Fine-tuning Stage

In the former pre-training stages, an convolutional encoder $\mathcal{E}_c$ and a ViT encoder $\mathcal{E}_t$ are pre-trained with a large number of unlabelled images. In the fine-tuning stage, we only fine-tune the model with a limited number of labelled images $\mathbf{x}^t \in \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^T\}$, where T is the size of the fine-tuning target dataset. Besides $\mathcal{E}_c$ and $\mathcal{E}_t$, a randomly initialized convolutional decoder $\mathcal{D}_c$ is added to predict segmentation masks from the recovered representation. Different levels of feature maps from $\mathcal{E}_c$ are concatenated with corresponding layers of $\mathcal{D}_c$ through skip connections, to provide alternative paths for the gradient. Dice loss

$$\mathcal{L}_{dice} = \sum_{c}^{C} \sum_{i}^{N} \frac{\mathbf{p}_{ic}\mathbf{g}_{ic}}{\mathbf{p}_{ic}\mathbf{g}_{ic} + (1-\mathbf{p}_{ic})\mathbf{g}_{ic} + \mathbf{p}_{ic}(1-\mathbf{g}_{ic})}$$

is applied as in usual multi-organ segmentation tasks. N is the total number of pixel in each mini-batch and i is the index of each individual pixel. C denotes the total number of classes. $\mathbf{p}_{ic}$ is the predicted probability that i-th pixel is class c and $\mathbf{g}_{ic}$ is 1 if i-th pixel is class c and 0 otherwise. The entire model is trained in an end-to-end fashion.

5. Experiments

5.1. Setup

5.1.1 Pre-training Dataset

During both contrastive and generative pre-training stages, we pre-train the encoders on the Abdomen-1K [36] dataset. It contains over 1,112 CT scans, which contains over 240K 2D slices. The CT scans are from 12 medical centers, including multi-phase, multi-vendor, and multi-disease cases. Although segmentation masks for liver, kidney, spleen, and pancreas are provided in this dataset, we ignore these labels during pre-training, since we are following the self-supervised protocol.

5.1.2 Fine-tuning Dataset

During the fine-tuning stage, we perform extensive experiments on three datasets with respect to different regions of human body, to evaluate the transfer-ability of the pre-trained models.

ABD-110 is an abdomen dataset from [43] that contains 110 CT scans from patients with various abdomen tumors and these CT scans were taken during the treatment planning stage. We report the average DSC on 11 abdomen organs (large bowel, duodenum, spinal cord, liver, spleen, small bowel, pancreas, left kidney, right kidney, stomach and gallbladder), with a random split of 1, 10, 50 training cases and 25 test cases.

Thorax-85 is a thorax dataset from [10] that contains 85 thorax CT scans. We report the average DSC on 6 thorax organs (eso, trachea, spinal cord, left lung, right lung, and heart), with a random split of 1, 10, 50 training cases and 25 test cases.

HaN is from [11] and contains 120 CT scans covering the the region of head and neck. We report the average DSC on 28 head and neck organs (brachial plexus, brainstem, constrictor naris, left ear, right ear, left eye, right eye, hypophysis, larynx, left lens, right lens, mandible, optical chiasm, left optical nerve, right optical nerve, oral cavity, left parotid, right parotid, left submandibular gland, right submandibular gland, spinal cord, sublingual gland, left temporal lobe, right temporal lobe, thyroid, left TMJ, right TMJ and trachea), also with a random split of 1, 10, 50 training cases and 25 test cases.

5.1.3 Evaluation metric

We use the same evaluation metric Sørensen–Dice coefficient (DSC) as in previous work [41]. DSC measures the overlap of the prediction mask $\mathbf{m}_p$ and ground truth mask $\mathbf{m}_g$ and is defined as

$$\mathrm{DSC}(\mathbf{m}_p, \mathbf{m}_g) = \frac{2|\mathbf{m}_p \cup \mathbf{m}_g|}{|\mathbf{m}_p| + |\mathbf{m}_g|}$$

.

5.1.4 Implementation Details

All images are re-sampled to have spacing of 2.5mm $\times$ 1.0mm $\times$ 1.0mm, with respect to the depth, height, and width of the 3D volume.

In the contrastive pre-training stage, we apply random resized crop with size of 224 and scale between 0.2 and 1.0; color jittering with brightness of 0.4, contrast of 0.4, saturation of 0.4 and hue of 0.4; and random horizontal flip to formulate positive samples. All data augmentation techniques are available in PyTorch's torchvision package. We use the SGD optimizer with momentum of 0.9 and weight decay of 10^{-4} to train a U-Net [41] encoder $\mathcal{E}_c$ for 200 epochs.

In the generative pre-training stage, we use the AdamW [30] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$ to train $\mathcal{E}_t$ with 12 ViT-base [17] and $\mathcal{D}_t$ with 4 ViT-base for 1600 epochs. Note that we are able to train such a large number of epochs due to two reasons. The first is that we take advantage of the strategy in [25], which mask out most part of the input and only perform training on the rest parts. The second is that due to the efficient design of RepRec, we only recovers feature maps from a limited latent space rather than from raw pixels space, which reduces both training time and memory space. In the fine-tuning stage, we use the Adam optimizer with momentum of 0.9 and weight decay of 10^{-4} to train the whole framework end to end.

5.2. Quantitative Results

5.2.1 Results on ABD-110

Table 1 shows the performance comparison of RepRec with previous work on ABD-110. We ran the following contrastive self-supervised pre-training algorithms: MoCo [26], DenseCL [48], Domain-Specific [4]; generative self-supervised pre-training algorithms: MAE [25], MaskFeat [49]; and combination of both contrastive and generative pre-training: SaGe [45]. We also compare with random initialization and ImageNet pre-trained (fully supervised) initialization.

By comparing the DSC scores on ABD-110 dataset, we demonstrate RepRec's scalability over 3 different training set sizes, $|T| = 1$, 10, and 50. RepRec provides Dice score

Method	ABD-110			Thorax-85			HaN		
	$\|T\|$=1	$\|T\|$=10	$\|T\|$=50	$\|T\|$=1	$\|T\|$=10	$\|T\|$=50	$\|T\|$=1	$\|T\|$=10	$\|T\|$=50
Baseline									
Random init.	47.08	74.32	79.64	50.75	84.73	87.66	37.16	55.94	75.45
ImageNet	50.03	80.47	83.39	53.77	85.74	89.47	40.74	69.56	76.84
Contrastive loss pre-training									
He *et al*. [26]	50.02	81.25	83.20	52.90	86.34	89.42	40.25	67.07	76.68
Wang *et al*. [48]	49.23	81.03	83.86	52.46	86.41	89.12	40.92	59.94	75.12
Chaitanya *et al*. [4]	49.60	81.43	84.23	53.04	**87.04**	89.61	41.12	65.24	76.75
Generative loss pre-training									
He *et al*. [25]	47.84	77.61	80.70	50.91	84.87	88.78	37.54	64.10	75.04
Wei *et al*. [49]	47.17	76.34	80.94	51.63	84.83	88.99	37.15	67.82	75.70
Combination of contrastive and generative methods									
Tian *et al*. [45]	49.90	81.45	84.16	52.22	86.86	89.74	40.78	70.24	**77.92**
Ours	**50.31**	**81.89**	**84.67**	**53.97**	87.01	**90.37**	**41.99**	**71.71**	77.31

Table 1. Comparison of the proposed method with other pre-training methods including the contrastive ones, the generative ones and the combination of the two. After extensive experiments on three datasets for different parts of human body, among different sizes of target training size $|T|$, we show that RepRec presents its effectiveness compared to other methods.

of 50.31%, 81.89%, and 84.67% on the ABD-110 dataset. Comparing MAE with random initialization, marginal benefit is gained after pre-training: only 0.76% improvement when only given 1 labelled CT scan, 3.29% improvement when only given 10 labelled CT scans, and 1.06% improvement when only given 50 labelled CT scans. This supports our arguments in the Motivation section. Our RepRec approach gains 3.23%, 7.57%, 4.52% improvement with different fine-tuning set size of 1, 10, 50 respectively. Compared to ImageNet pre-training, which is fully-supervised, RepRec provides 0.28%, 1.42%, and 1.28% improvement respectively.

5.2.2 Results on Thorax-85

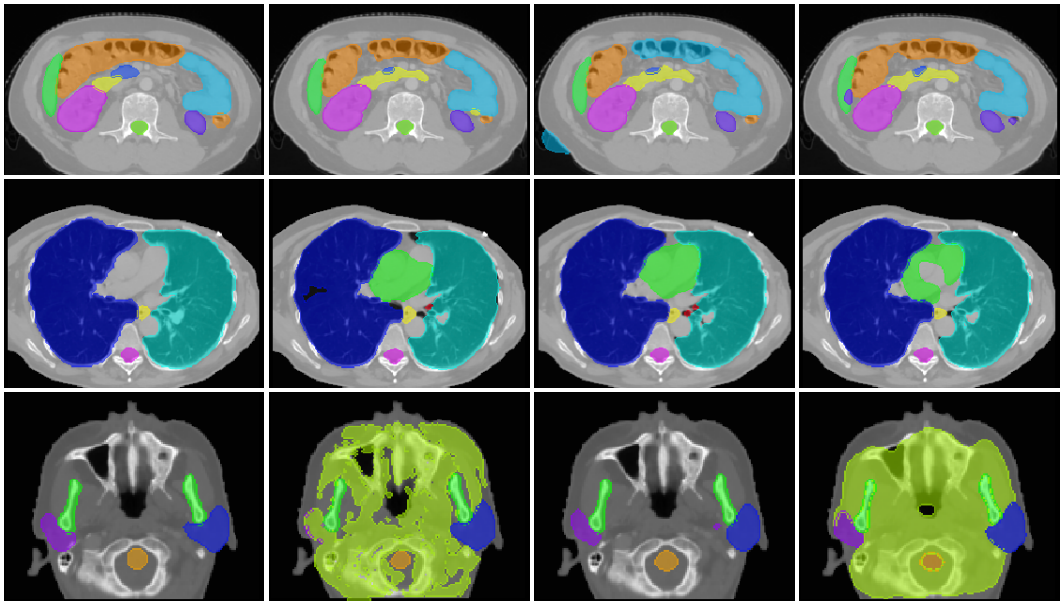
Comparing the DSC scores on Thorax-85 dataset with other SOTA pre-training algorithms, we demonstrate RepRec's superior performance, while $|T| = 1$, 10, and 50. RepRec provides Dice score of 53.97%, 87.01%, and 90.37% on the ABD-110 dataset. Comparing MAE with random initialization, marginal benefit is gained after pre-training: only 0.16% improvement when only given 1 labelled CT scan, 1.61% improvement when only given 10 labelled CT scans, and 1.12% improvement when only given 50 labelled CT scans. This supports our arguments in the Motivation section again. Our RepRec approach gains 3.22%, 2.28%, 2.71% improvement with different fine-tuning set size of 1, 10, 50 respectively. Comparing with Chaitanya *et al*.[4], our method outperforms theirs 0.93% when $|T| = 1$ and 0.76% when $|T| = 50$, while $|T| = 10$, the DSC score of [4] is only 0.03% higher than ours. Compared to ImageNet pre-training, which is fully-supervised, RepRec provides 0.20%, 1.27%, and 0.9% improvement respectively. Experiments on Thorax-85 shows that even though RepRec is pre-trained on abdomen dataset, its highly flexible transferability allows it to compete with other SOTA approaches.

5.2.3 Results on HaN

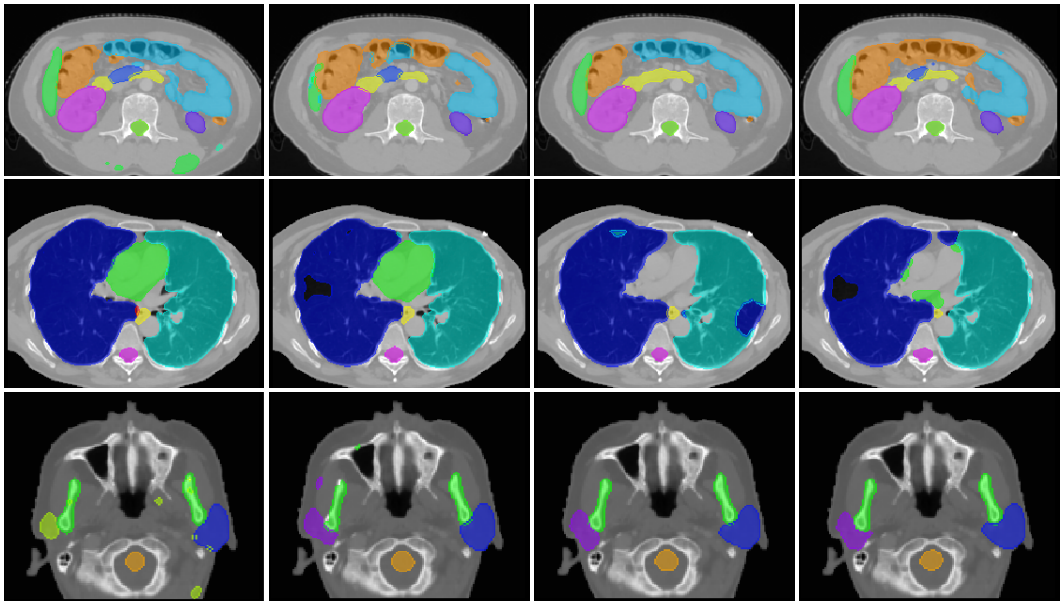
By fine-tuning on HaN dataset, RepRec provides Dice score of 41.99%, 71.71%, and 77.92% when $|T| = 1$, 10, and 50. Comparing MAE with random initialization, marginal benefit is gained after pre-training: only 3.09% improvement when only given 1 labelled CT scan, 11.13% improvement when only given 10 labelled CT scans, and 1.23% improvement when only given 50 labelled CT scans. This supports our arguments in the Motivation section once again. Our RepRec approach gains 4.83%, 15.77%, 1.86% improvement with different fine-tuning set size of 1, 10, 50 respectively. Comparing with Tian *et al*.[45], our method outperforms theirs 0.93% when $|T| = 1$ and 1.47% when $|T| = 10$, while $|T| = 50$, the DSC score of [45] is only 0.61% higher than ours. Compared to ImageNet pre-training, which is fully-supervised, RepRec provides 1.25%, 2.15%, and 0.47% improvement respectively. Experiments on both HaN and Thorax-85 verifies that even though RepRec is pre-trained on abdomen dataset, it can be transferred to datasets of other locations on human body.

5.3. Qualitative Results

In Figure 3, we present visualized segmentation results on ABD-110 (row 1 and row 4), Thorax-85 (row 2 and row 5) and HaN (row 3 and row 6) datasets respectively. All the results are provided by the models trained with target dataset size $|T| = 10$. Thanks to the representation recovering mechanism, RepRec presents its effectiveness com-



(a) Ground Truth (b) Random (c) ImageNet (d) MoCo



(e) Domain-Specific (f) SaGe (g) MAE (h) Ours

Figure 3. Qualitative results provided by different models on ABD-110 (row 1 and row 4), Thorax-85 (row 2 and row 5) and HaN (row 3 and row 6) datasets. All the results are provided by the models trained with target dataset size $|T|$ =10. Thanks to the representation recovering mechanism, RepRec presents its effectiveness compared to other methods (better view in color).

pared to other methods.

5.4. Ablation Study

For example, in our example of ABD-110 dataset, ImageNet, Domain-Specific [4], SaGe [45] and [25] pretrained models make false prediction masks for Small Bowel. In SaGe [45], liver is covered by Small Bowel. MoCo [26] is making additional false predictions on the left kidney.

On Thorax-85, every model predicts reasonable masks for left lung, right lung, spinal cord and eso. Random initialized, ImageNet, MoCo, Domain-Specific, SaGe and RepRec pretrained models all make false prediction on heart. However, in terms of the area of false positive, our approach produces the smallest error.

On HaN, every model predicts reasonable masks for

Methods	Random	MoCo [26]	SaGe [45]	Ours
$\mathcal{E}_c$ + $\mathcal{D}_c$	74.32	81.25	81.45	N/A
$\mathcal{E}_c$ + $\mathcal{E}_t$+ $\mathcal{D}_c$	75.46	81.29	81.57	**81.89**

Table 2. DSC scores provided by RepRec with different decoder on ABD-110 dataset. All the results are provided by the models trained with target dataset size $|T|$ =10. We compare our method with previous SOTA methods under the same setting of parameter numbers. We show that with the same number of additional parameters from the ViT decoder D_t, RepRec still achieves the state of the art results.

Methods	4×ViT + 1×Conv	PUP [61]	U-Net (w.o. skip connections)	U-Net
Random	73.59	73.78	73.99	**75.46**
RepRec	77.97	79.03	80.07	**81.89**

Table 3. DSC scores of different methods with the same number of additional parameters from the ViT decoder D_t, on ABD-110 dataset. All the results are provided by the models trained with target dataset size $|T|$ =10. We show that increasing scales of decoder $\mathcal{D}_c$, the fine-tuning results can be improved clearly, which is not the case argued in the MAE paper. This is because of the special dense prediction property for multi-organ segmentation tasks compared to natural image classification and detection tasks.

brain stem ■ and mandible ■. However, random initialized model and MoCo [26] make large false positive mask predictions. ImageNet and SaGe [45] pretrained models make false positive masks for parotid ■.

5.4.1 Effect of additional parameters in ViT encoder $\mathcal{E}_t$

In table 2, we pre-train our whole model (including convolutional encoder $\mathcal{E}_c$, vision transformer encoder $\mathcal{E}_t$ and convolutional decoder $\mathcal{D}_c$) in an end to end fashion, using MoCo [26] and SaGe [45] pre-training strategy. With an additional vision transformer encoder $\mathcal{E}_t$, MoCo [26] and SaGe [45] only benefit 0.04% and 0.12% from it. We compare our method with previous SOTA methods under the same setting of parameter numbers. We show that with the same number of additional parameters from the ViT decoder D_t, RepRec still achieves the state of the art results. We verify the superiority of RepRec is from the way it process the global and local contextual information by utilizing both contrastive and generative learning rather than by adding additional parameters.

5.4.2 Choices of decoder $\mathcal{D}_c$

In table 3, we show that with increasing scales of decoder $\mathcal{D}_c$, the fine-tuning results can be improved clearly. By comparing U-Net decoder and PUP decoder mentioned in [61] with a simple 4 x ViT + 1 x Conv decoder, RepRec gains 2.1% and 1.06% improvement on downstream segmentation tasks. This is not the case in MAE [25]. This shows for dense down-stream tasks like multi-organ segmentation, decoder still plays an important role compared to classification and object detection tasks.

By adding skip connections in U-Net [41] from convolutional encoder $\mathcal{E}_c$ to convolutional decoder $\mathcal{D}_c$, DSC score of both random initialized method and our RepRec get improvements by 1.47% and 1.82% respectively. This shows skip connections from encoder to decoder are also essential in segmentation tasks. However, no such structure can be applied to pure transformer based model such as MAE [25], which hurts the performance of MAE on multi-organ segmentation tasks.

5.4.3 Limitations of Vanilla MAE

We demonstrate the recovering results with different mask ratio by a **vanilla** MAE model in figure 1. The vanilla MAE model provides reasonable recovering results given different mask ratio. However, in the fine-tuning stage, the experimental results in table 1 shows that vanilla MAE doesn't provide potential transferable capability for dense modeling tasks such as multi-organ segmentation, even though the model is able to recover the original image with reasonable quality.

6. Conclusion

In this paper, we propose RepRec, a hybrid visual representation learning framework for self-supervised pre-training on large scale unlabelled medical datasets. RepRec utilizes the advantage of both contrastive and generative modeling. Both quantitative and qualitative studies, validate the favorable ability of RepRec in down-stream multi-organ segmentation tasks compared to former state of the art models. Overall, we believe that the proposed RepRec algorithm is a feasible way of unifying existing self-supervised generative approaches and discriminative approaches. We hope that RepRec inspire future study and will be incorporated with other pre-training strategies as well.

Beyond the currently proposed framework, it is possible to merge the contrastive branch and the discriminative branch in a more unified way, and train the whole frame work in an end-to-end fashion, which we will study in the future.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers, 2021.

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *ArXiv*, abs/2006.09882, 2020.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

[4] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33, 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[7] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning, 2022.

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021.

[10] Xuming Chen, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Huang, Guoqi Zhao, Yi Xu, Tingfeng Chen, Xiaohui Xie, and Yong Liu. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, July 2021.

[11] Xuming Chen, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Huang, Guoqi Zhao, Yi Xu, Tingfeng Chen, Xiaohui Xie, and Yong Liu. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, 2021.

[12] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[13] Yifei Chen, Haoyu Ma, Deying Kong, Xiangyi Yan, Jianbao Wu, Wei Fan, and Xiaohui Xie. Nonparametric structure regularization machine for 2d hand pose estimation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 381–390, 2020.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] 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*, pages 1422–1430, 2015.

[16] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction, 2016.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[18] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners, 2022.

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

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

[21] 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, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[22] Indranil Guha, Syed Ahmed Nadeem, Chenyu You, Xiaoliu Zhang, Steven M Levy, Ge Wang, James C Torner, and Punam K Saha. Deep learning based high-resolution reconstruction of trabecular bone microstructures from low-resolution ct scans using gan-circle. In *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 11317, page 113170U. International Society for Optics and Photonics, 2020.

[23] Kun Han, Shanlin Sun, Xiangyi Yan, Chenyu You, Hao Tang, Junayed Naushad, Haoyu Ma, Deying Kong, and Xiaohui Xie. Diffeomorphic image registration with neural velocity field. In *WACV*, 2023.

[24] Kun Han, Shanlin Sun, Chenyu You, Hao Tang, Deying Kong, Xiangyi Yan, and Xiaohui Xie. Diffeomorphic image registration with neural velocity field, 2022.

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[27] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Dec. 2020.

[28] Deying Kong, Yifei Chen, Haoyu Ma, Xiangyi Yan, and Xiaohui Xie. Adaptive graphical model network for 2d handpose estimation. *arXiv preprint arXiv:1909.08205*, 2019.

[29] Deying Kong, Linguang Zhang, Liangjian Chen, Haoyu Ma, Xiangyi Yan, Shanlin Sun, Xingwei Liu, Kun Han, and Xiaohui Xie. Identity-aware hand mesh estimation and personalization from rgb images. *arXiv preprint arXiv:2209.10840*, 2022.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[31] Qing Lyu, Chenyu You, Hongming Shan, and Ge Wang. Super-resolution mri through deep learning. *arXiv preprint arXiv:1810.06776*, 2018.

[32] Qing Lyu, Chenyu You, Hongming Shan, Yi Zhang, and Ge Wang. Super-resolution mri and ct through gan-circle. In *Developments in X-ray tomography XII*, volume 11113, page 111130X. International Society for Optics and Photonics, 2019.

[33] Haoyu Ma, Liangjian Chen, Deying Kong, Zhe Wang, Xingwei Liu, Hao Tang, Xiangyi Yan, Yusheng Xie, Shi yao Lin, and Xiaohui Xie. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. In *BMVC*, 2021.

[34] Haoyu Ma, Zhe Wang, Yifei Chen, Deying Kong, Liangjian Chen, Xingwei Liu, Xiangyi Yan, Hao Tang, and Xiaohui Xie. Ppt: token-pruned pose transformer for monocular and multi-view human pose estimation. In *ECCV*, 2022.

[35] Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18051–18061, 2022.

[36] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[37] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[38] 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.

[39] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[41] 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.

[42] Shanlin Sun, Kun Han, Deying Kong, Hao Tang, Xiangyi Yan, and Xiaohui Xie. Topology-preserving shape reconstruction and registration via neural diffeomorphic flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20845–20855, June 2022.

[43] Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, and Xiaohui Xie. Recurrent mask refinement for few-shot medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3918–3928, October 2021.

[44] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020.

[45] Yunjie Tian, Lingxi Xie, Xiaopeng Zhang, Jiemin Fang, Haohang Xu, Wei Huang, Jianbin Jiao, Qi Tian, and Qixiang Ye. Semantic-aware generation for self-supervised visual representation learning, 2021.

[46] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

[47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[48] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021.

[49] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training, 2021.

[50] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[51] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[52] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better, 2021.

[53] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling, 2021.

[54] Fan Xu, Haoyu Ma, Junxiao Sun, Rui Wu, Xu Liu, and Youyong Kong. Lstm multi-modal unet for brain tumor segmentation. In *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pages 236–240. IEEE, 2019.

[55] Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *WACV*, 2022.

[56] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE Transactions on Medical Imaging*, 39(1):188–203, 2019.

[57] Chenyu You, Ruihan Zhao, Lawrence Staib, and James S Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. *arXiv preprint arXiv:2105.07059*, 2021.

[58] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging, 2022*, 2022.

[59] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[60] Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction, 2017.

[61] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.