IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 April 2024, accepted 6 May 2024, date of publication 9 May 2024, date of current version 17 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3399628

RESEARCH ARTICLE

PatchLoc: Embedded Patch Localization Pretext Task for Tumor Segmentation in Medical Images

RAMCHANDRA CHEKE[1,2,3], CIARÁN EISING[1,2,3], (Senior Member, IEEE), PATRICK DENNY[1,3,4], (Member, IEEE), AND PEPIJN VAN DE VEN[1,2,3]

[1]Department of Electronic and Computer Engineering, University of Limerick, Limerick, V94 T9PX Ireland
[2]SFI CRT Foundations in Data Science, University of Limerick, Limerick, V94 T9PX Ireland
[3]Data Driven Computer Engineering Research Group, University of Limerick, Limerick, V94 T9PX Ireland
[4]Department of Computer Science and Information Systems (CSIS), University of Limerick, Limerick, V94 T9PX Ireland

Corresponding author: Ramchandra Cheke (cheke.ramchandra@ul.ie)

This work was supported by the Science Foundation Ireland under Grant 18/CRT/6049.

ABSTRACT Supervised deep learning methods have produced state-of-the-art results with large labeled datasets. However, accessing large labeled datasets is difficult in medical image analysis because of a shortage of medical experts, expensive annotations, and privacy constraints in the healthcare domain. Self-supervised learning is a branch of machine learning that exploits unlabeled data to encourage network weights toward a valid latent representation of the data during a so-called pretext task. The features learned by the model while solving pretext tasks are transferred to a downstream task where limited annotations are available. In this work, we propose PatchLoc, a novel pretext task whose objective is to find the location of a given patch from an image as a source of supervision. We validated the effectiveness of PatchLoc on a downstream segmentation task using three different medical datasets. PatchLoc yields substantial improvements compared to U-Net trained from scratch and other pretext task-based approaches in a low data regime.

INDEX TERMS Medical imaging, pretext tasks, self-supervised learning, limited annotations.

I. INTRODUCTION

Supervised deep learning methods have achieved state-of-the-art results in medical image segmentation tasks [1], [2], [3], [4], [5]. However, the success of supervised deep learning methods is heavily dependent on the availability of large amounts of labeled data. Medical expertise is generally required to annotate regions of interest in medical images. As a result, creating new large labeled datasets is an expensive process. Moreover, such labeling is time-consuming and does not represent the best use of already scarce medical resources [6].

Various methods have been proposed to address the problem of scarcity of labeled datasets in the medical domain. One of the proposed solutions is transfer learning, which is a well-studied method in machine learning that allows us to re-utilize features learned by a neural network from a previous task to improve the performance of the new target task [7]. In practice, models pre-trained on a large labeled dataset (e.g. ImageNet [8]) are fine-tuned on a smaller labeled dataset originating from the specific application of interest. Although there have been some encouraging results of transfer learning in medical image analysis [9], [10], this strategy has certain drawbacks. Medical images (CT scans and Magnetic Resonance Imaging) fundamentally differ from natural images. In the medical imaging context, CT scans represent intensity values using Hounsfield Units (HU) and use single-channel images to store these values [11]. Natural images consist of three separate channels to store the intensity values of an image's red, green, and blue components. To address this problem, researchers from IBM [12] pre-trained an image classification network on a grayscale ImageNet dataset and fine-tuned this pre-trained model for an X-ray disease classification task. While, the authors reported an increase in X-ray disease classification AUC from 0.7498 to 0.7706, recent studies [13], [14] have shown that

The associate editor coordinating the review of this manuscript and approving it for publication was Tai Fei.

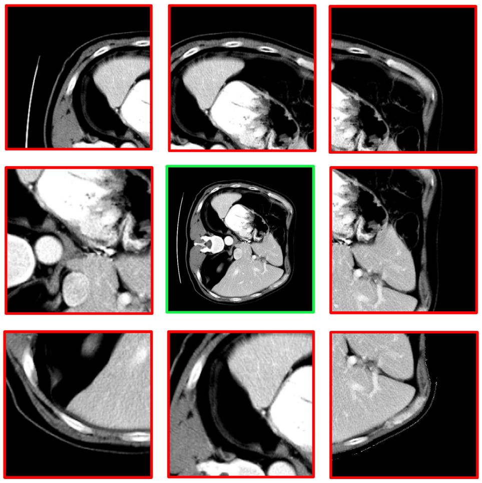

FIGURE 1. Medical image (center with green outline) and 8 examples of scaled patches used by PatchLoc (with red outline).

pre-training a model on a medical dataset can achieve better results compared to natural images. However, a significant amount of effort is needed to annotate large medical datasets. A further issue with the use of ImageNet for pre-training, is high-level features acquired by models are biased toward the original datasets. Due to these reasons, transfer learning from ImageNet pre-trained models to the medical datasets is a sub-optimal solution.

To circumvent the reliance of supervised learning algorithms on large amounts of annotated data, self-supervised learning was introduced. This approach involves automatically deriving labels from the data, allowing models to be trained, at least to some extent, without the requirement for manual data labeling. There are two steps in the pipeline of self-supervised learning. The first step is to pre-train a model to solve a pretext task using an unlabeled dataset. These pretext tasks are carefully designed so that target labels are automatically generated by applying appropriate geometric transformations on images. By solving a pretext task, the network is forced to learn efficient representations. This representation is then used in the second step to initialize a model on the main downstream task of image classification or segmentation. Transferring the features acquired during the pretext task to new target tasks with limited labeled data is advantageous because it reduces the issues of dataset bias and distribution shift since the pretext tasks and downstream tasks are trained on comparable data sets.

Numerous pretext task techniques have been suggested to advance the field of self-supervised learning in the domains of natural images [15], [16], [17], [18], [19], [20] and self-supervised learning methods have been used in a wide range of applications [21], [22], [23], [24], [25], [26], [27]. A common theme in the majority of these initiatives is that the use of the pretext results is limited to use on downstream tasks using the same dataset. So, pre-training a model on a pretext task using natural images is generally only used for downstream tasks on natural images. Similarly, training models to process CT scans would require a CT scan dataset for the pretext task. Being able to increase the dataset size used in such pretext tasks with natural images to learn a good representation is advantageous in medical imaging applications where the limited size of the dataset is generally of significant concern.

In this work, we aim to bridge these gaps by introducing PatchLoc, an innovative pretext task of patch localization that leverages large unlabeled datasets for the pre-training of neural networks to address the challenges of scarce label datasets. As shown in Figure 1, PatchLoc extracts a patch from the given image and scales this patch to the size of the original image. The objective of the pretext task is to locate this patch correctly within the original image. To solve this task, a network requires an understanding of objects present in the image and patch. As the patch is a zoomed version of a part of the image, the model also learns multi-scale mapping from patch to image. The contributions of this paper are as follows:

1) We propose "PatchLoc", a novel pretext task, which focuses on determining the location of a given patch from a whole image. This helps the network learn local features from a patch and global features from a complete image.
2) We successfully adopt this pre-training technique in the medical domain by training a network on CT images and we evaluate the benefits of PatchLoc using a downstream segmentation task on three distinct CT datasets.
3) We compare the results obtained with our pretext task to those achieved with no pre-training and pre-training with other pretext tasks. Our experimental findings show that using our suggested strategy in settings with minimal labeled data significantly improves downstream segmentation accuracy when compared to scenarios with no pre-training and alternative pretext tasks.
4) Finally, we also show that adding unlabeled grayscale natural images in pre-training improves the performance on the downstream task.

The rest of this paper is structured as follows: Section II offers a concise overview of prior research on pretext tasks designed for natural and medical images. Section III provides a detailed explanation of the proposed patch localization method. Section IV introduces the datasets used in the study and outlines the experimental settings. We analyze the models' performance across the three datasets in Section V. Lastly, Section VI summarizes the conclusions drawn from this work and discusses possible directions for further study.

II. PRIOR ART

A. PRETEXT TASKS IN NATURAL IMAGES

Inspired by the idea of context prediction [28] in natural language processing, Doersch et al. [15] developed one of the early works in patch-based pretext tasks for natural images. The objective of this task was to learn visual representations of data by correctly identifying the position of neighboring patches with respect to the central patch.

Noroozi and Favaro [18] show that predicting relative patch location tasks can be ambiguous when two non-central tiles have similar patterns and further argue that solving a jigsaw puzzle task that focuses on all tiles together leads to learning better representation. PatchLoc focuses on determining the location of a given patch from a whole image, helping to learn local features from a patch and global features from a complete image. Secondly, the ambiguity of predicting two non-central tiles having similar patterns in RPL is overcome by PatchLoc because it is easier to find the location of a patch from a whole image. Another pretext task aims to learn visual representations by predicting colors from its grayscale images [16]. This pretext task is not appropriate for CT datasets as CT contains only grayscale images.

Gidaris et al. [17] proposed a rotation prediction pretext task to learn semantic features from unlabeled natural images. In this task, a network is trained to predict one of four possible angles of rotation applied to an input image. This pretext task assumes that objects in the natural images are captured from canonical viewpoints. Therefore, to predict the rotations the network should understand the orientations of different objects in the image. However, medical image datasets do not characteristically contain a large number of distinctive objects or shapes to get maximum benefits from this method.

Recently, contrastive learning-based models [19], [20], [29], [30], [31] have achieved state-of-the-art performance on natural images [8]. In contrastive learning, the objective is to minimize the distance between positive pairs (different views of the same image) and maximize the distance between negative pairs (different images) in the hidden latent space using the contrastive loss function [32], [33]. The selection of positive and negative pairs plays a vital role in the success of contrastive learning models. A batch should contain a large number of negatives to avoid the mode collapse problem; the situation in which all inputs map to the single trivial solution. SimCLR [20] uses batch sizes of up to 8192, utilizing 128 TPU cores to achieve state-of-the-art performance. SwAV [31] also uses a batch size of 4096 distributed across 64 GPUs. Therefore, the computational cost to train these models is very high compared to that of their standard supervised counterparts. Data augmentation is another crucial factor in designing pretext tasks based on a contrastive learning framework. The augmentations applied for natural images may not be appropriate for medical image datasets because of the following reasons: 1) The difference between normal and abnormal conditions in medical images is defined by inspecting a small number of pixels. Therefore, two views from the same image should contain the region that defines abnormality. Using random crop and blurring may obscure the important ROI from an augmented image. 2) Color jittering and random grayscale cannot apply to already grayscale medical images. Therefore, the direct application of these methods for medical image analysis is a major challenge due to the high computing power and the need for careful selection of negative pairs.

B. PRETEXT TASKS IN MEDICAL IMAGES

Several pretext tasks have been proposed specifically for the medical domain that exploit the intrinsic properties of medical images. Jamaludin et al. [21] define a pretext task using contrastive loss to distinguish vertebral bodies of different patients and then use a classification loss to predict disc degeneration after 10 years. Whilst the authors reported a gain from 74.5% to 76% in classification accuracy, this approach is useful only when multiple scans of the same patients are available. In a pretext task proposed by Zhang et al. [34], two slices are randomly chosen from a CT volume, and the network is trained to determine the correct order of selected slices. The pre-trained model is fine-tuned on the downstream task of body parts recognition in CT and MR images. In the analysis of various pretext tasks for medical images, Tajbakhsh et al. [23] used colorization as a pretext task for skin cancer segmentation. However, the gain provided by the pretext task is smaller than that obtained from an ImageNet pre-trained model. The majority of these proposed techniques are constrained to the particular downstream task because of the assumptions upon which these pretext tasks are designed.

Sowrirajan et al. [24] used the MoCo [29] framework to pre-train the MoCo-CXR model on chest X-ray images and then used the pre-trained model for the classification of abnormalities from X-rays. The authors used pre-trained weights from ImageNet to initialize before pre-training MoCo-CXR. Chaitanya et al. [35] attempted to utilize contrastive learning for medical images by dividing the 3D volume into different partitions and identifying the slices' corresponding partitions in various volumes as positive pairs, and those of different partitions as negative pairs. However, the difference between the slices at the end of one partition and the starting slices of the adjacent partition is small. Therefore, Zeng et al. [36] proposed Positional Contrastive Learning for Volumetric Medical Image Segmentation. Whilst results reported in the latter two papers were good, the methods proposed by Chaitanya et al. and Zeng et al. require 3D datasets (CT or MRI) to construct a pretext task. Another work of Chaitanya et al. [37] used pseudo-labels and contrastive learning framework using self-training strategy and Wang et al. [38] used multi-task learning with a student-teacher model for medical image segmentation. Zhou et al. [39] proposed model genesis which employs a variety of self-supervised techniques, all of which are defined as an image restoration task, to generate the self-supervision signal. This approach uses an encoder-decoder architecture to restore the original image from the perturbed image to learn information about the appearance, context, and texture. Zhang et al. [40] utilized Multimodality from the medical datasets and proposed contrastive domain-sharing generative adversarial networks. Although these methods

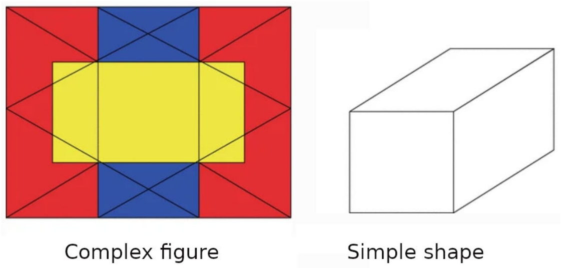

FIGURE 2. Example question used in embedded figure test (EFT) adopted from [41]. In this test, the participant's task is to search for obscured simple geometrical shapes (right) in a more complex color diagram (left).

achieved promising results, pixel-based pretext tasks are computationally very expensive.

C. *PatchLoc* ADVANTAGES

The disadvantages of existing methods as discussed in the previous sections, can be summarized as follows: 1) various pretext tasks are not suitable for medical images due to the absence of distinctive geometries in medical images; 2) most pretext tasks require the use of the same dataset for pretext task and downstream task; 3) many pretext tasks, such as those using contrastive loss and pixel-bases tasks, are computationally expensive. Our method is particularly suited to medical data and learns global-level features from images and local-level features from patches while solving pretext tasks. Moreover, the proposed method does not make any assumptions about the dataset and it leverages medical images as well as natural images in pre-training to learn representations from data in a computationally efficient way. Finally, Our approach is simple to implement and it eliminates the need for large batch sizes and the meticulous selection of negative pairs, which are essential requirements in contrastive learning frameworks.

III. METHODS

The motivation for our PatchLoc method comes from the Embedded Figures Test (EFT). The EFT was introduced by Gottschaldt (1926 [42], 1929 [43]) as a useful tool for evaluating a person's capacity to distinguish a figure from its surroundings. Later, Herman Witkin designed a more complex version of EFT [44] and used it to study field-dependent and field-independent cognitive styles [45]. A typical example question from EFT is shown in Figure 2. In this embedded figure test, the participant is shown the complex figure (left) for 15 seconds after which it is withdrawn. Then a simple shape (right) is displayed for 10 seconds. The complex figure is then again presented to participants with the objective of locating the simple figure in it. The EFT has applications in many areas, including psychology, education, and even certain occupational examinations, and has been important in understanding cognitive styles.

Inspired by EFT, we propose a novel pretext task to learn generic latent representations with application in medical images. We define a pretext task similar to the question in Embedded Figures Test (EFT). Instead of drawing a simple shape and embedding it into a complex figure, we extract a smaller patch from the image and train a neural network to determine its size and position in the original image. The patch is obtained as a random crop of a predefined size from a particular image scaled to the size of the original image (Figure 3). We use a Siamese-like architecture to solve the patch localization problem. Siamese networks were first used to verify signatures [46]. Later they became a popular choice in many applications [47] because of their ability to process two images/signals and compare/contrast between them using the appropriate loss function. They consist of two identical branches that share weight parameters and are connected at the end to calculate the similarity between the two different inputs. In our work, the image and patch are propagated through a Siamese network to extract features of the image and patch. The correct location of the patch is identified using the categorical loss function. Finally, the pre-trained encoder is used as the encoder in a U-Net model which is trained further using labeled data on the downstream task of segmentation. Figure 4 illustrates the entire workflow of our approach.

TABLE 1. Class index used in categorical cross-entropy loss function to pre-train a network using our patch localization method.

Class index	Y_{start}	X_{start}	Patch size	Class index	Y_{start}	X_{start}	Patch size
0	0	0	256	17	85	256	170
1	0	128	256	18	85	340	170
2	0	256	256	19	170	0	170
3	128	0	256	20	170	85	170
4	128	128	256	21	170	170	170
5	128	256	256	22	170	256	170
6	256	0	256	23	170	340	170
7	256	128	256	24	256	0	170
8	256	256	256	25	256	85	170
9	0	0	170	26	256	170	170
10	0	85	170	27	256	256	170
11	0	170	170	28	256	340	170
12	0	256	170	29	340	0	170
13	0	340	170	30	340	85	170
14	85	0	170	31	340	170	170
15	85	85	170	32	340	256	170
16	85	170	170	33	340	340	170

A. EXTRACTION OF PATCHES

Firstly, all images are resized to a dimension of 512 × 512 pixels, and then a random patch is extracted from an image. Next, an image and a patch both are resized to a uniform size of 256 × 256 pixels and provided as input to the network to maintain consistency. The location of the patch embedded within an image and its size are chosen from a pool of 34 potential configurations as shown in Table 1. In Table 1, X_{start} and Y_{start} define a starting coordinate of an image from where a square patch is extracted. There is a total of nine possible locations from where a square patch of size 256 × 256 can be extracted. To improve the discriminative power of the task, two adjacent patches have 50% overlap

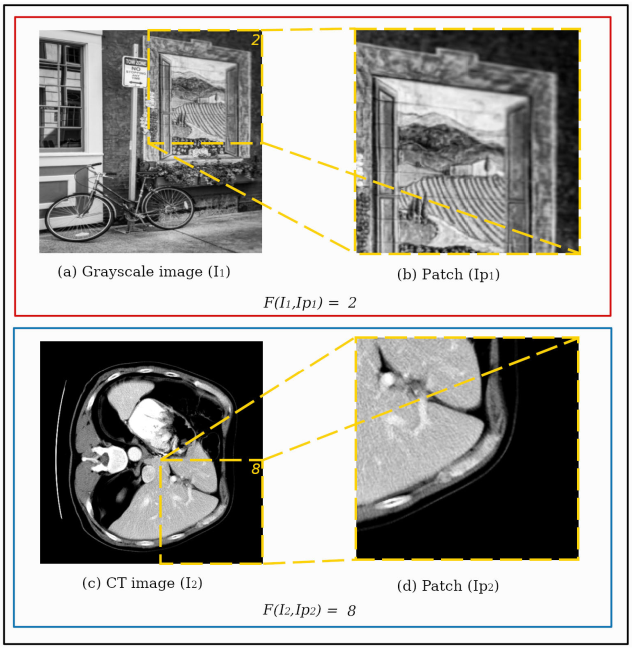

FIGURE 3. Example patch (I_p) extraction illustrated for a natural image (I_1) and a medical image (I_2).

between them. Furthermore, the height and width of the patch are reduced to $1/3^{rd}$ of the original image and this creates an additional 25 locations from where a patch can be extracted.

B. NETWORK DETAILS

We adopted the U-Net [1] based encoder as a feature extractor, as U-Net is one of the most successful architectures for segmentation in the medical domain [35], [36], [48], [49].

1) PRETEXT TASK

As depicted in Figure 5, we employ a Siamese neural network architecture to process the image denoted as I and its corresponding patch, referred to as I_p. Both branches of the Siamese encoder have the same weights. Our encoder consists of six convolution blocks starting from Conv1 to Conv6. Each convolution block has two 3 × 3 convolutions followed by ReLU activation units. In the initial five convolution blocks (Conv1 to Conv5), the first convolution operation is performed with a stride of 2 and the second convolution with a stride of 1. Stride 2 is used to reduce spatial dimension by half without using maxpooling. In the Conv6 block, both 3 × 3 convolutions have a stride of 1. The number of feature maps doubles from Conv1 to Conv4 and remains the same until Conv6. We use instance normalization [50] after every convolution operation. Residual connections are employed [51] for faster training. The output of Conv6 feature maps from both branches are combined and further processed through the fully connected layers. To classify the patch according to its position and size outlined in Table 1, we employ a classifier with a softmax activation.

We adopt a categorical cross-entropy loss function to pre-train the model on our patch localization task:

$$\mathcal{L}_{pretext}(I_{p,c}, \hat{I}_{p,c}) = -\frac{1}{M}\sum_{c=1}^{M} I_{p,c}\log(\hat{I}_{p,c}) \tag{1}$$

where M is the number of samples in the mini-batch. $I_{p,c}$ denotes true class c of that patch according to Table 1. $\hat{I}_{p,c}$ is the soft-max probability predicted by the model for the c^{th} class.

2) DOWNSTREAM TASK

For the segmentation task, we use a Residual U-Net implemented as shown in Figure 6. It is an encoder-decoder architecture with skip connections between the encoding layers and corresponding decoding layers. The details of the encoder are already discussed in the previous section. The decoder uses strided transpose convolution to up-sample the data from the previous layers. Each block in the decoding path has a transpose convolution followed by concatenation from the encoding block and then a 3 × 3 convolution operation to reduce the output feature maps. Unlike the original U-Net implementations, these down or up-sampling operations are implemented at the beginning of each block.

To train the network on the downstream task, we use a loss function based on the Dice Similarity Coefficient (DSC), also known as the Sørensen-Dice coefficient [52]. The DSC is a metric to measure the similarity or overlap between two sets. In the segmentation task, the predicted output is the segmentation mask and it is compared against ground truth. The DSC is widely used in medical image analysis to quantitatively evaluate the performance of different models on the segmentation task by measuring the degree of overlap between the predicted regions and the ground truth [2], [4]. The DSC is defined as:

$$DSC(R_p, R_g) = \frac{2\left|R_p \cap R_g\right|}{\left|R_p\right| + \left|R_g\right|} \tag{2}$$

where R_p denotes the predicted segmentation mask and R_g denotes the ground truth for a particular class. The DSC ranges from 0 to 1, where a higher value indicates a greater degree of similarity between the predicted segmentation map and ground truth or more accurate segmentation results. The downstream loss function is, therefore, given as:

$$\mathcal{L}_{downstream}(R_p, R_g) = 1 - DSC(R_p, R_g) \tag{3}$$

IV. EXPERIMENTAL SETUP

To investigate the effect of pre-training a model using our patch localization pretext task, we use the features learned by the pre-trained encoder to initialize the U-Net encoder on the segmentation task. The decoder of U-Net is randomly initialized. Following that, we train U-Net on the segmentation task on three different datasets using different fractions of labeled training data.

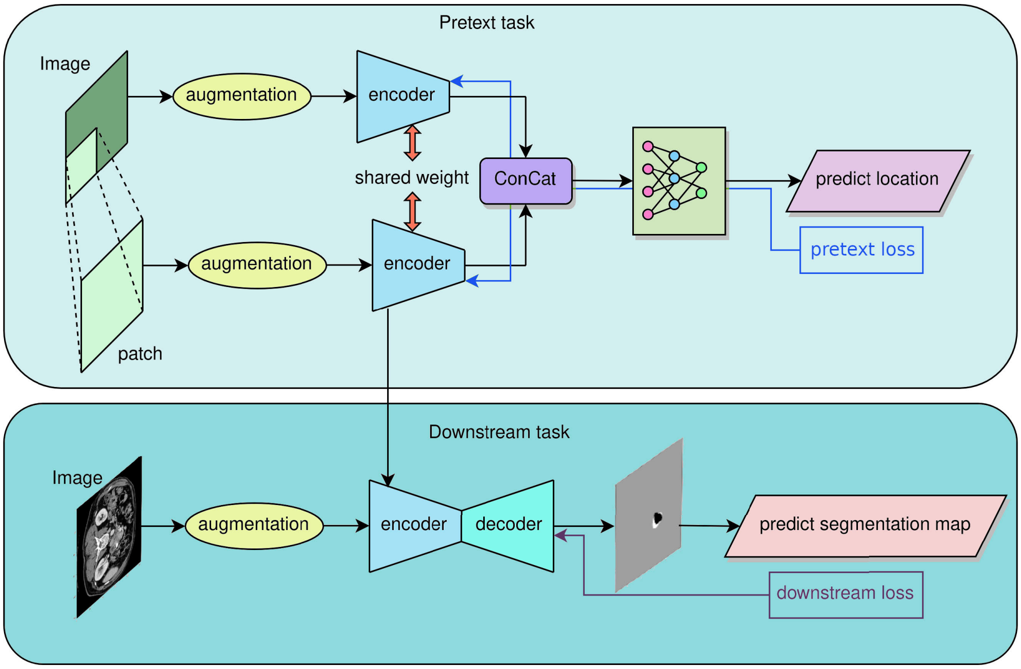

FIGURE 4. Conceptual diagram of our self-supervised patch localization method.

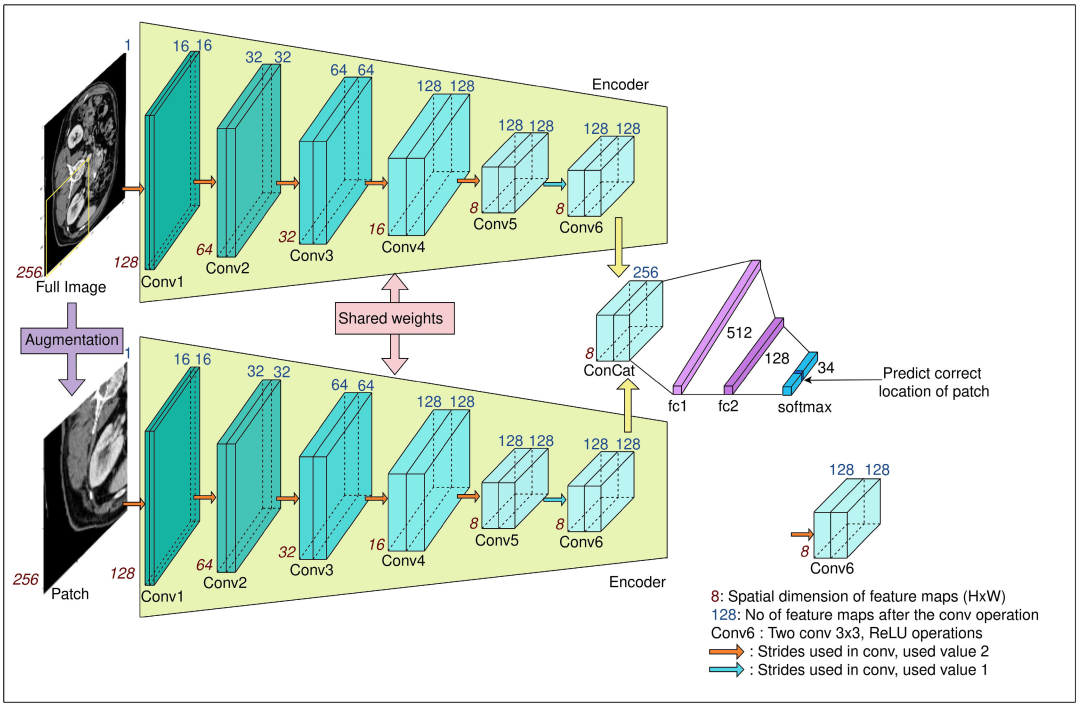

FIGURE 5. The Siamese network architecture to predict the location of the patch from the original image.

A. PRETEXT TASK DATASET

We have used 110k images from the COCO dataset [53] without using labels. These images have been converted into a single-channel grayscale image. Additionally, we have added 37k CT scan images from the pancreas dataset, resulting in a total of 147k images. The CT images have undergone preprocessing steps outlined in Section IV-C. Finally, the CT images were normalized to [0,255] to ensure that both natural images and CT images have the same range of grayscale values, so they can be processed and analyzed together effectively.

B. DOWNSTREAM TASK DATASETS AND SPLITS

We use three CT scan datasets from the MICCAI'18 medical segmentation decathlon challenge [54] to investigate the benefits of our proposed method. Examples of the CT scan

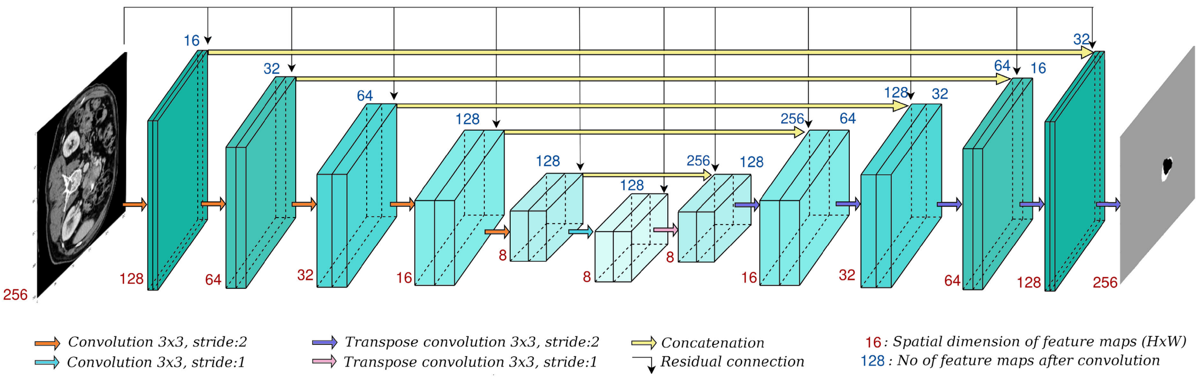
16
32
64
128
128
128
256
128
256
64
128
32
64
16
32
256
128
64
32
16
8
8
8
16
32
64
128
256
Convolution 3x3, stride:2
Convolution 3x3, stride:1
Transpose convolution 3x3, stride:2
Transpose convolution 3x3, stride:1
Concatenation
Residual connection
16 : Spatial dimension of feature maps (HxW)
128 : No of feature maps after convolution

FIGURE 6. U-Net with residual connections and transpose convolution for up-sampling.

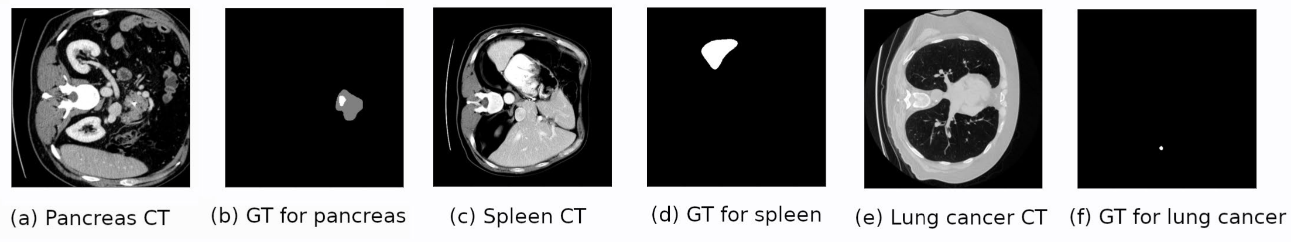
(a) Pancreas CT (b) GT for pancreas (c) Spleen CT (d) GT for spleen (e) Lung cancer CT (f) GT for lung cancer

FIGURE 7. Sample CT scans and corresponding ground truth (GT) for pancreas, spleen, and lung cancer dataset.

images and corresponding ground truth segmentation masks are shown in Figure 7.

1) PANCREAS DATASET

The pancreas dataset consists of CT scans of 281 subjects suffering from pancreatic cancer. Segmentation masks consist of three different classes: background, pancreas, and tumor. In the segmentation mask (Figure 7(b)), black denotes the background, the pancreas class is shown in gray, and the tumor is shown in white. We randomly selected 240 samples for training, reserving the remaining 41 samples for validation. A total of 40k images were extracted from these 240 volumes equivalent to 100% of the data size for training. Following that, we split this training dataset into five different data splits: 5%(∼2k), 10%(∼4k), 25%(∼10k), 50% (∼20k). We created 5 folds for each of these data splits. Finally, we reported validation scores on 41 samples which corresponds to 4.2k images. It is important to note that our validation set is larger than the 10% split.

2) SPLEEN DATASET

The spleen dataset is composed of 41 studies, which we divided into 32 for training and 9 for validation. The training data was further partitioned into five different subsets by the number of studies and images they contain as follows: (1) 2 studies (400 images), (2) 6 studies (1.2k images), (3) 11 studies (2.3k images), (4) 17 studies (3.9k images), and (5) the complete training set, consisting of 6800 images.

3) LUNG CANCER DATASET

This dataset includes 63 CT scans of patients with non-small cell lung cancer. As shown in Figure 7(f), detecting lung cancer from a CT scan image is a difficult problem compared to segmentation of the pancreas and spleen. We randomly chose 50 samples for the training dataset. The remaining 13 samples, which amount to 3.4k images, were retained for validation purposes. From the training dataset, a total of 13k images were extracted. Segmentation of lung cancer was the most challenging task among the three tasks used in our experiments due to the very small size of the cancer nodule. We created only three smaller data splits from the training dataset with 1.2k, 2.3k images, and 4.7k images because reducing the dataset size can lead to a significantly lower segmentation score and the model may start overfitting. It is important to note that the validation data is not used in the pre-training of pretext tasks.

C. PREPROCESSING

The in-plane resolution of these datasets varied from 0.6 × 0.6 mm to 0.97 × 0.97 mm and through-plane resolution varied from 0.7 mm to 7.5 mm. We have used the bi-linear and nearest-neighbor interpolation methods to re-sample 2D image slices and segmentation masks from 3D volumes. All images were re-sampled to the fixed resolution of 0.8 × 0.8 mm. For the pancreas dataset, the CT scan images were clipped to a range of [−96.0, 215.0] HU values as outlined in [4]. Subsequently, all images were normalized using a mean value of 77.99 HU and a standard deviation of 75.40 HU. In the case of spleen CT images, clipping was performed within the range of [−41, 176] HU. These images were then normalized by subtracting the mean value of 99.29 HU and dividing by the standard deviation of 39.47. Lung CT images were clipped to a range of [−1024, 325] HU, followed by adding the mean value of 158.58 and dividing by the standard deviation of 324.70 HU. All images and segmentation masks were resized to 256 × 256 pixels.

D. DATA AUGMENTATION

For both the pretext and downstream task, a data augmentation pipeline is used. In our pipeline, we employ a range of techniques to enhance input data variability using the Monai [55] framework, consisting of the following steps:

1) **Intensity scaling**: we scale initial intensity values from [0, 255] to $[-1, 1]$ to speed up convergence.
2) We **randomly rotate** patches by an angle θ ranging from $-10°$ to $10°$.
3) **Gaussian noise** is applied to the data with a mean of 0.5 and a standard deviation of 0.5 to enhance the model's generalizability.
4) **Coarse dropout** is applied by creating 5 holes of 10 × 10 dimensions in the input data. This dropout technique helps to learn spatial invariance, making it more robust to missing or occluded regions.
5) **Gaussian smoothing** is applied with a sigma that ranges between 0.25 and 1.2.
6) **Shift intensity**: finally, we include intensity variation by randomly adjusting pixel values within ± 0.2 of its original value, allowing the model to learn to handle differences in brightness and contrast.

These probabilistic augmentations are added to the data augmentation pipeline with probability p varies from 0.1 to 0.3.

E. TRAINING DETAILS

1) PRETEXT TASK

We have employed a PyTorch-based implementation for the pre-training of the Siamese model using both natural images and medical image datasets. The pretext model was trained over 150 epochs using a batch size of 512 on an NVIDIA A100 40GB GPU. To ensure a smooth start to the training process we used a warm-up of 5 epochs. We adopted the Layer-wise Adaptive Rate Scaling (LARS) [56] optimizer with a base learning rate of 0.1 and the weight decay was set to 1e-05. We reduced the learning rate by half at 50, 90, and 120 epochs. The network was trained until the validation loss reached convergence (requiring a total training duration of 10 hours).

2) DOWNSTREAM TASK

The downstream segmentation model was trained with a batch size of 64 over 300 epochs using an NVIDIA A100 40GB GPU. We used the Adam optimizer [57] with learning rate 1e-4; weight decay 1e-5; and default beta values ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The data augmentations specified in section IV-D were again used. As discussed previously, the Dice loss function (3), was used to train the model on the downstream task. For the pancreas and spleen datasets, the Dice loss was calculated by considering the background class. However, the Dice loss function was employed for the lung cancer dataset without considering the background, based on the observation that the ratio of lung cancer (positive) to the background (negative) was very small. The models' performance on the validation dataset was monitored during training, and the epoch where the model yielded the best performance on the validation set was reported. No post-processing was applied.

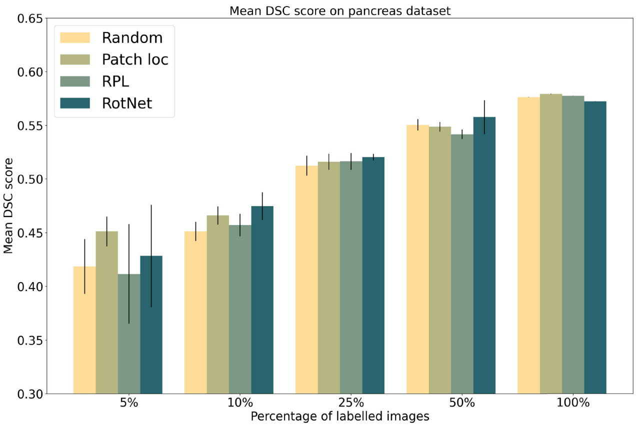

FIGURE 8. Comparison of our method with randomly initialized encoder and other pretext task-based methods: relative patch location (RPL), rotation prediction network (RotNet) within a standard deviation.

TABLE 2. Transfer learning on spleen and lung cancer dataset. Rotnet, RPL and PatchLoc methods are pre-trained using COCO and pancreas CT datasets.

Dataset	%Data	Random	PatchLoc	RPL	RotNet
Lung cancer	5%	0.2234	**0.2651**	0.2042	0.2041
Lung cancer	10%	0.4038	0.4071	0.3828	**0.4356**
Lung cancer	25%	0.5513	0.5755	0.5454	**0.5866**
Lung cancer	50%	0.6218	**0.6552**	0.6301	0.6391
Lung cancer	100%	0.6555	**0.6960**	0.6666	0.6867
Spleen	5%	0.5445	**0.6622**	0.5352	0.5655
Spleen	10%	0.7463	**0.8084**	0.7736	0.7708
Spleen	25%	0.8520	**0.8717**	0.8677	0.8532
Spleen	50%	0.8775	0.8889	0.8892	**0.8976**
Spleen	100%	0.9281	**0.9316**	0.9296	0.9301
Pancreas	5%	0.4185	**0.4512**	0.4116	0.4283
Pancreas	10%	0.4512	0.4661	0.4572	**0.4747**
Pancreas	25%	0.5124	0.5162	0.5165	**0.5205**
Pancreas	50%	0.5505	0.5487	0.5417	**0.5578**
Pancreas	100%	0.5764	**0.5794**	0.5775	0.5725

F. EVALUATION

The Dice score (2), was used for the evaluation of the downstream tasks by considering the foreground class. The spleen and lung datasets each contain a single foreground class, while the pancreas dataset comprises two classes: pancreas and tumor. As a result, for the pancreas dataset, the mean DSC was computed for the pancreas and tumor classes.

V. RESULTS AND DISCUSSION

We evaluated our results by comparing them with the performance of an encoder initialized with random weights and an encoder pre-trained using the rotation prediction task [17] and the relative patch location pretext task [15]. We implemented these tasks to process grayscale natural images and medical images.

TABLE 3. Statistical t-test results. Positive differences indicate that the PatchLoc dice score is higher than other comparing methods and p-value ≤ 0.05 indicates that the results are statistically significant.

	D5		D10		D25		D50	
	diff	p-value	diff	p-value	diff	p-value	diff	p-value
Lung dataset								
PatchLoc vs Random	**0.042**	**0.050**	0.003	0.422	0.024	0.194	**0.033**	**0.016**
PatchLoc vs RPL	**0.061**	**0.049**	0.024	0.120	**0.030**	**0.036**	0.025	0.187
PatchLoc vs RotNet	0.061	0.132	-0.028	0.075	-0.011	0.231	0.016	0.131
Spleen dataset								
PatchLoc vs Random	**0.118**	**0.015**	**0.062**	**0.038**	0.020	0.079	0.003	0.078
PatchLoc vs RPL	0.097	0.077	0.038	0.097	**0.018**	**0.014**	0.002	0.074
PatchLoc vs RotNet	**0.127**	**0.001**	0.035	0.084	0.004	0.360	0.001	0.351
Pancreas dataset								
PatchLoc vs Random	**0.033**	**0.003**	**0.015**	**0.046**	0.004	0.074	-0.002	0.042
PatchLoc vs RPL	**0.040**	**0.050**	**0.009**	**0.014**	0.000	0.102	**0.007**	**0.008**
PatchLoc vs RotNet	0.023	0.175	-0.009	0.073	-0.004	0.114	-0.009	0.152
Combined datasets								
PatchLoc vs Random	**0.067**	**0.001**	**0.029**	**0.020**	**0.016**	**0.047**	**0.012**	**0.039**
PatchLoc vs RPL	**0.068**	**0.006**	**0.039**	**0.016**	**0.018**	**0.002**	0.011	0.079
PatchLoc vs RotNet	**0.074**	**0.001**	0.001	0.451	-0.004	0.254	0.003	0.301

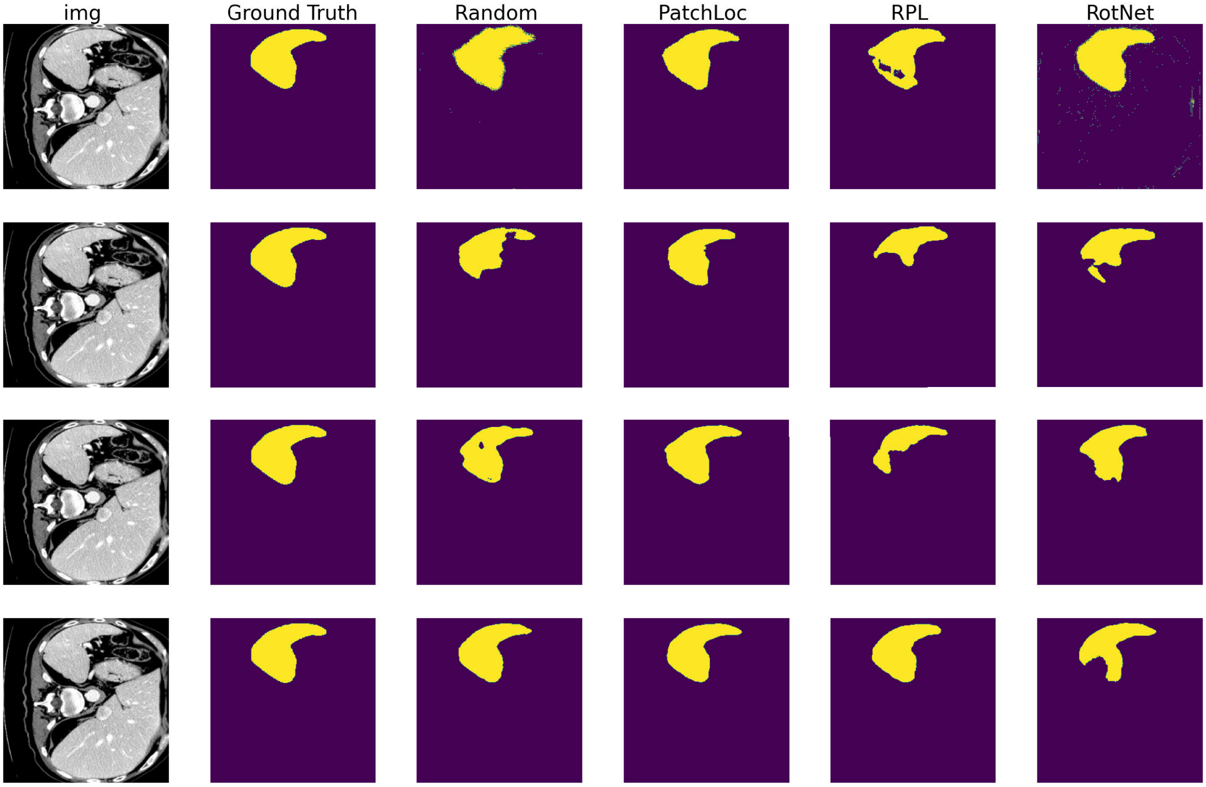

FIGURE 9. Visual segmentation outcomes for the Spleen dataset are illustrated. The initial row signifies a 5% data split, followed by rows indicating 10%, 25%, and 50% data splits sequentially.

Rotation prediction is originally proposed by Gidaris et al. as a simple yet effective pretext task for natural images. In this task, an input image is rotated by four angles (0, 90, 180, or 270 degrees) before propagating to the network, and the network is trained to predict the correct rotation. This four-class classification problem is solved by using the categorical cross-entropy.

In **Relative patch location** [15] task, a network is trained to predict the relative location between a central patch and a second patch chosen at random from one of its eight neighboring locations, which are sequentially numbered from 1 to 8. The Siamese network is used to process two patches and the categorical cross-entropy is used as a loss function.

Figure 8 shows the segmentation results from the pancreas dataset. The model pre-trained using our patch localization task outperformed the randomly initialized model with a good margin at small label fractions (5% and 10%). When only 5% data is available for training, our method achieves a mean dice score of 0.4512 which is higher by 7.81% than the model trained with random weights. Furthermore, in the case of 5% labeled data, our approach yields notable improvements when compared to a network pre-trained on predicting relative patch location (mean dice score: 0.4116) and rotation

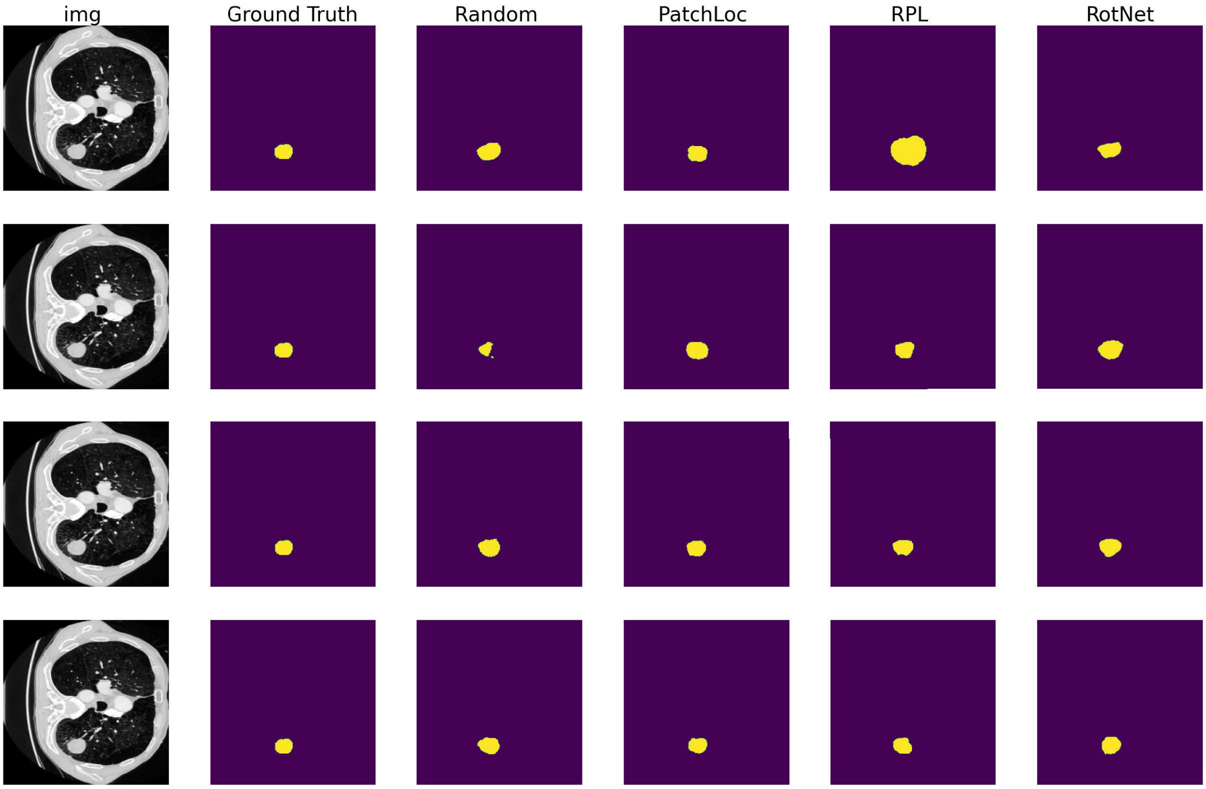

FIGURE 10. Visual segmentation outcomes for the Lung dataset are illustrated. The initial row signifies a 5% data split, followed by rows indicating 10%, 25%, and 50% data splits sequentially.

prediction task (mean dice score: 0.4283). We believe that our pretext task is difficult compared to the relative patch location task because it consists of patches with multiple scales and more patches than the RPL task. However, we are providing a complete image as opposed to the center patch (RPL) and it helps to map features from a complete image and multi-scaled patch information in the hidden latent space. Also, it is important to note the initialization produced by our methods is more robust than RPL and Rotnet because the standard deviation is smaller for our approach as compared to the rest of the approaches. Finally, the benefits produced by the pretext task-based approaches become negligible for large-label datasets.

We also examined the impact of transferring the features discovered through the pretext task on two additional CT datasets: 1) spleen and 2) lung cancer dataset, to illustrate the generalizability of our method by demonstrating the consistency across different datasets. Mean dice scores are reported in Table 2. Similar to the pancreas dataset, PatchLoc significantly enhances segmentation results on the spleen and lung cancer datasets, particularly in situations where there is a shortage of labeled data. Furthermore, our approach performs better than RPL, rotation prediction, and randomly initialized models in most cases.

The visual results for the spleen, lung, and pancreas dataset are shown in Figure 9, 10, and 11 respectively with each figure containing a matrix of images. The rows of these matrices of images represent a 5%, 10%, 25%, and 50% data split respectively. As the proportion of labeled data increases, the segmentation quality is enhanced. Whilst Figure 9 and 10 show that PatchLoc results in good segmentation on the spleen and lung dataset respectively, Figure 11 shows that results for the pancreas dataset demonstrate room for improvement in all tested methods. The segmentation problem in the pancreas dataset The performance on the pancreas dataset is limited because it involves addressing a multiclass segmentation challenge, characterized by high class imbalance, particularly for the tumor class. However, our results are better than one of the state-of-the-art methods nnUnet [4] which reported a mean dice score of 0.5619 for 2D network.

TABLE 3 quantifies the performance differences between Patchloc and the other methods using a statistical t-test. Particularly in the low data regime, PatchLoc provides a statistically significant gain over other initialization strategies with p-value $\leq$ 0.05. Available compute resources and data limit the number of train-test splits for use in the t-tests. As a result, statistical significance cannot be shown for all individual datasets, as indicated in table 3 by p-values exceeding 0.05. For this reason, we also performed the t-test by combining the mean dice score for each data split across three datasets. These results, shown under 'Combined datasets', show that our PatchLoc method outperforms comparator methods with statistically significant results. Even though RPL resembles PatchLoc closely, PatchLoc performs better than RPL. As shown in table 2, the mean dice

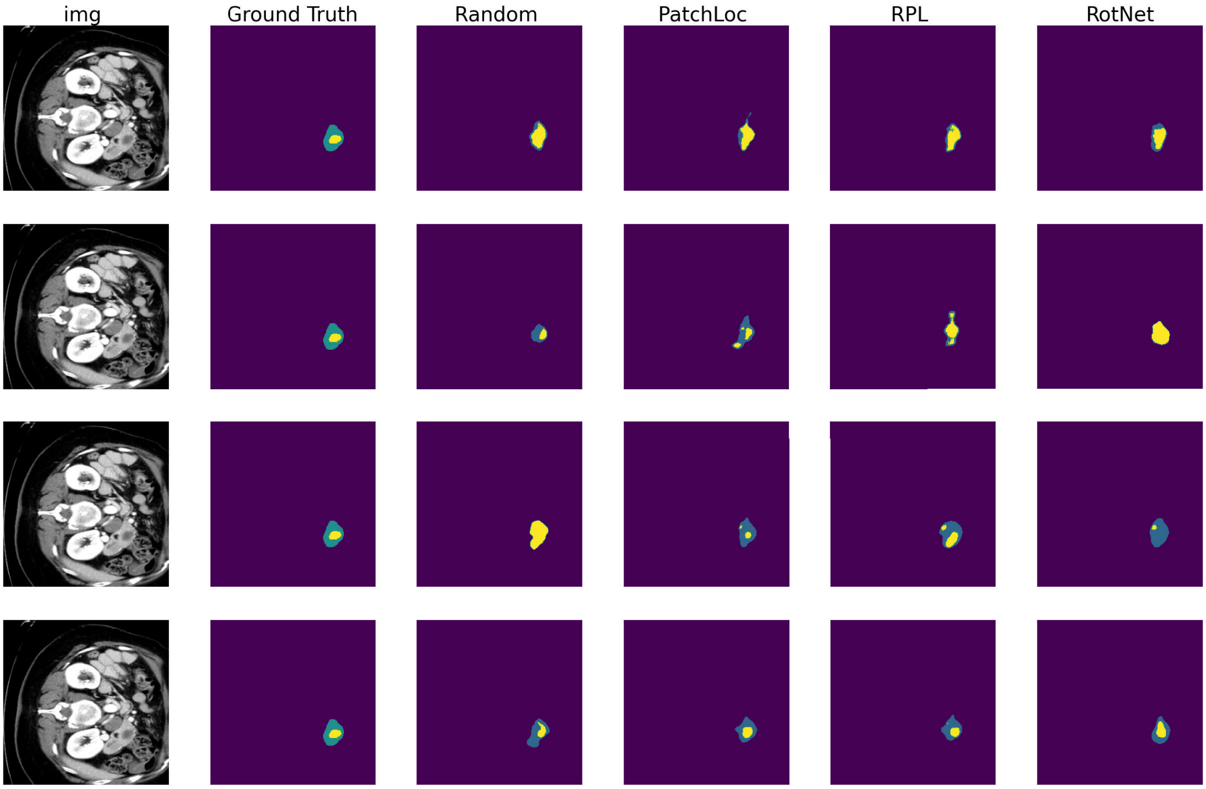

FIGURE 11. Visual segmentation outcomes for the Pancreas dataset are illustrated. The initial row signifies a 5% data split, followed by rows indicating 10%, 25%, and 50% data splits sequentially.

TABLE 4. Effect of using non-medical images in pre-training. pre-training a model using grayscale natural images and medical images improves results.

Method	Dataset used in pre-training	DSC ± std dev
Random weights	-	0.4185 ± 0.0255
Ours	(CT + COCO)	0.4512 ± 0.0139
RPL	(CT + COCO)	0.4116 ± 0.0463
RotNet	(CT + COCO)	0.4283 ± 0.0475
Ours	CT only	0.4449 ± 0.0158
RPL	CT only	0.4280 ± 0.0297
RotNet	CT only	0.4163 ± 0.0264

score for PatchLoc is higher than RPL in all cases and in most of the cases the results are statistically significant.

Finally, we investigated whether using grayscale natural images for pre-training a network helps to learn generic features as compared to using only CT scan images for pretext tasks. In this experiment, we focussed on only a limited data setting because, for the large data region, all initialization schemes have achieved comparative performance. Table 4 shows the mean DSC on the pancreas dataset using a 5% data split. Leveraging grayscale natural images along with CT images in pre-training helped to learn better feature representation using our task as compared to RPL and Rotnet. Rotnet, particularly, struggles when the data exhibits a mesh-like structure, and training it solely on CT scan images is suboptimal. These self-supervised tasks can easily overfit to data. This provides a potentially promising pathway for creating a pretext task that makes use of both real-world natural images and medical data to enhance feature representation.

One of the limitations of PatchLoc and other pretext tasks is that it trains only the encoder and the decoder is randomly initialized. Moreover, when the segmentation model is trained on a relatively larger labeled dataset, the original features learned by the pretext task change to a great extent. This leads to a decline in performance gain. This finding is consistent with [23], [24], and [35] which also found greater performance increases for models trained on smaller label fractions but with diminishing gains when trained on larger label fractions.

VI. CONCLUSION

We propose PatchLoc, a novel pretext task of patch localization and we adopt it for tumor segmentation in medical images. Specifically, we investigate the effectiveness of our proposed approach across three public CT datasets. The results from our experiments indicate that employing PatchLoc in scenarios with limited labeled data leads to a substantial improvement in segmentation accuracy as compared to training from scratch, or using other pretext tasks. Furthermore, we demonstrate that adding natural images in the pre-training of our patch localization task offers additional gains as compared to training only on medical images. Future research could involve applying this strategy to 3D data, where a 3D pre-trained model can be trained directly utilizing pretext tasks [49], [58].

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, 2015, pp. 234–241.

[2] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[3] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[4] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[5] P. Tang, P. Yang, D. Nie, X. Wu, J. Zhou, and Y. Wang, "Unified medical image segmentation by learning from uncertainty in an end-to-end manner," *Knowl.-Based Syst.*, vol. 241, Apr. 2022, Art. no. 108215. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705122000594

[6] A. Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college," *BMJ*, vol. 359, Oct. 2017. [Online]. Available: https://www.bmj.com/content/359/bmj.j4683

[7] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC Med. Imag.*, vol. 22, no. 1, p. 69, Dec. 2022.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[9] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[10] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[11] T. D. DenOtter and J. Schubert, *Hounsfield Unit*. Treasure Island, FL, USA: StatPearls, 2022. [Online]. Available: http://europepmc.org/books/NBK547721

[12] Y. Xie and D. Richmond, "Pre-training on grayscale ImageNet improves medical image classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany. Berlin, Germany: Springer, Sep. 2018, pp. 476–484, doi: 10.1007/978-3-030-11024-6_37.

[13] A. Parakh, H. Lee, J. H. Lee, B. H. Eisner, D. V. Sahani, and S. Do, "Urinary stone detection on CT images using deep convolutional neural networks: Evaluation of model performance and generalization," *Radiology: Artif. Intell.*, vol. 1, no. 4, Jul. 2019, Art. no. e180066.

[14] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang, "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artif. Intell.*, vol. 4, no. 5, Sep. 2022.

[15] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.

[16] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 649–666.

[17] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.

[18] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 69–84.

[19] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Researc, vol. 119, H. Daumé III and A. Singh, Eds., 2020, pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[21] A. Jamaludin, T. Kadir, and A. Zisserman, "Self-supervised learning for spinal MRIs," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. S. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis, and Z. Lu, Eds. Cham, Switzerland: Springer, 2017, pp. 294–302.

[22] W. Bai, C. Chen, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen, Y. Guo, P. M. Matthews, and D. Rueckert, "Self-supervised learning for cardiac MR image segmentation by anatomical position prediction," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Shenzhen, China. Germany: Springer-Verlag, Oct. 2019, pp. 541–549, doi: 10.1007/978-3-030-32245-8_60.

[23] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding, "Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1251–1255.

[24] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "MoCo pretraining improves representation and transferability of chest X-ray models," in *Proc. 4th Conf. Med. Imag. Deep Learn.*, 2021, pp. 728–744.

[25] X. Cao, H. Lin, S. Guo, T. Xiong, and L. Jiao, "Transformer-based masked autoencoder with contrastive loss for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023, Art. no. 5524312, doi: 10.1109/TGRS.2023.3315678.

[26] A. Konrad, C. Eising, G. Sistu, J. McDonald, R. Villing, and S. Yogamani, "FisheyeSuperPoint: Keypoint detection and description network for fisheye images," 2021, *arXiv:2103.00191*.

[27] R. Cheke, G. Sistu, C. Eising, P. van de Ven, V. Ravi Kumar, and S. Yogamani, "FisheyePixPro: Self-supervised pretraining using fisheye images for semantic segmentation," *Electron. Imag.*, vol. 34, no. 16, pp. 147-1–147-6, Jan. 2022.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[29] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[30] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent—A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 21271–21284. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf

[31] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 9912–9924.

[32] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.

[33] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[34] P. Zhang, F. Wang, and Y. Zheng, "Self supervised deep representation learning for fine-grained body part recognition," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 578–582.

[35] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2020, pp. 12546–12558.

[36] D. Zeng, Y. Wu, X. Hu, X. Xu, H. Yuan, M. Huang, J. Zhuang, J. Hu, and Y. Shi, "Positional contrastive learning for volumetric medical image segmentation," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Strasbourg, France. Berlin, Germany: Springer, 2021, p. 221230, doi: 10.1007/978-3-030-87196-3_21.

[37] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, ''Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation,'' *Med. Image Anal.*, vol. 87, Jul. 2023, Art. no. 102792. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523000531

[38] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, and Y. Wang, ''Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning,'' *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102447. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522000925

[39] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, ''Models genesis: Generic autodidactic models for 3D medical image analysis,'' in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11767. Cham, Switzerland: Springer, 2019, pp. 384–393, doi: 10.1007/978-3-030-32251-9_42.

[40] J. Zhang, S. Zhang, X. Shen, T. Lukasiewicz, and Z. Xu, ''Multi-ConDoS: Multimodal contrastive domain sharing generative adversarial networks for self-supervised medical image segmentation,'' *IEEE Trans. Med. Imag.*, vol. 43, no. 1, pp. 76–95, Jan. 2024, doi: 10.1109/TMI.2023.3290356.

[41] F. Happé, *Embedded Figures Test (EFT)*. New York, NY, USA: Springer, 2013, pp. 1077–1078, doi: 10.1007/978-1-4419-1698-3_1726.

[42] K. Gottschaldt, ''Über den einfluß der erfahrung auf die wahrnehmung von figuren: I. Über den einfluß gehäufter einprägung von figuren auf ihre sichtbarkeit in umfassenden konfigurationen,'' *Psychologische Forschung*, vol. 8, no. 1, pp. 261–317, 1926.

[43] K. Gottschaldt, ''Über den einfluß der erfahrung auf die wahrnehmung von figuren: II. Vergleichende untersuchungen über die wirkung figuraler einprägung und den einfluß spezifischer geschensverläfe auf die auffassung optischer komplexe,'' *Psychologische Forschung*, vol. 12, no. 1, pp. 1–87, 1929.

[44] H. A. Witkin, ''Individual differences in ease of perception of embedded figures,'' *J. Personality*, vol. 19, no. 1, pp. 1–15, Sep. 1950.

[45] H. A. Witkin and D. R. Goodenough, ''Cognitive styles: Essence and origins. field dependence and field independence,'' *Psychol. Issues*, vol. 51, no. 1, pp. 1–141, 1981. [Online]. Available: https://api.semanticscholar.org/CorpusID:40647079

[46] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, ''Signature verification using a 'Siamese' time delay neural network,'' in *Proc. 6th Int. Conf. Neural Inf. Process. Syst.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, pp. 737–744.

[47] Y. Li, C. L. P. Chen, and T. Zhang, ''A survey on Siamese network: Methodologies, applications, and opportunities,'' *IEEE Trans. Artif. Intell.*, vol. 3, no. 6, pp. 994–1014, Dec. 2022.

[48] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King, and J. A. Schnabel, ''Left-ventricle quantification using residual U-Net,'' in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, M. Pop, M. Sermesant, J. Zhao, S. Li, K. McLeod, A. Young, K. Rhode, and T. Mansi, Eds. Cham, Switzerland: Springer, 2019, pp. 371–380.

[49] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, ''3D self-supervised methods for medical imaging,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18158–18172.

[50] D. Ulyanov, A. Vedaldi, and V. Lempitsky, ''Instance normalization: The missing ingredient for fast stylization,'' 2016, *arXiv:1607.08022*.

[51] Z. Zhang, Q. Liu, and Y. Wang, ''Road extraction by deep residual U-Net,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[52] T. B. Srensen, T. Sorensen, T. Biering-Srensen, T. I. A. Srensen, and J. Sorensen, ''A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on Danish commons,'' in *Biologiske Skrifter*, vol. 5. Copenhagen, Denmark: The Royal Danish Academy of Sciences and Letters, Jan. 1948, pp. 1–34.

[53] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, ''Microsoft COCO: Common objects in context,'' in *Proc. 13th Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, 2014, pp. 740–755.

[54] M. Antonelli et al., ''The medical segmentation decathlon,'' *Nat. Commun.*, vol. 13, no. 1, p. 4128, 2022.

[55] MONAI Consortium, Nov. 2021, ''MONAI: Medical open network for AI,'' *Zenodo*, doi: 10.5281/zenodo.5728262.

[56] Y. You, I. Gitman, and B. Ginsburg, ''Large batch training of convolutional networks,'' 2017, *arXiv:1708.03888*.

[57] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[58] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, ''Self-supervised feature learning for 3D medical images by playing a Rubik's cube,'' in *Proc. MICCAI*, 2019, pp. 420–428.

RAMCHANDRA CHEKE received the B.E. degree in electronic and telecommunication engineering from the University of Mumbai, India, in 2014, and the M.Sc. degree in mathematical modeling from the University of Limerick, Ireland, in 2019. He is currently pursuing the Ph.D. degree in data science. From 2014 to 2018, he was a Software Engineer. His research interests include machine learning, computer vision, and medical image analysis.

CIARÁN EISING (Senior Member, IEEE) received the B.E. degree in electronic and computer engineering and the Ph.D. degree from the National University of Ireland, in 2003 and 2010, respectively. From 2009 to 2020, he was a Computer Vision Team Lead and an Architect with Valeo Vision Systems, where he also held the title of a Senior Expert. In 2016, he was awarded the position of an Adjunct Lecturer with the National University of Ireland, Galway. In 2020, he joined the University of Limerick as a Lecturer in artificial intelligence and computer vision.

PATRICK DENNY (Member, IEEE) received the B.Sc. degree in experimental physics and mathematics from NUI Maynooth, Ireland, in 1993, and the M.Sc. degree in mathematics and the Ph.D. degree in physics from the University of Galway, Ireland, in 1994 and 2000, respectively. He was with GFZ Potsdam, Germany. From 1999 to 2001, he was a RF Engineer with AVM GmbH, Germany, developing the RF hardware for the first integrated GSM/ISDN/USB modem. After working in supercomputing with Compaq-HP, from 2001 to 2002, he joined Connaught Electronics Ltd. (later Valeo), Galway, Ireland, as the Team Leader of RF Design. Over the next 20 years, he was a Lead Engineer developing novel RF and imaging systems and led the development of the first mass-production HDR automotive cameras for leading car companies, including Jaguar Land Rover, BMW, and Daimler. In 2010, he became an Adjunct Professor in engineering and informatics with the University of Galway. In 2022, he became a Lecturer in artificial intelligence with the Department of Electronic and Computer Engineering, University of Limerick, Ireland, where he became an Associate Professor in artificial intelligence and imaging with the Department of Computer Science and Information Systems (CSIS), in 2024. He is currently the Co-Founder and a Committee Member of the IEEE P2020 Automotive Imaging Standards Group, the AutoSens Conference on Automotive Imaging, and the IS&T Electronic Imaging Autonomous Vehicles and Machines (AVM) conference.

PEPIJN VAN DE VEN received the M.Sc. degree in electronic engineering from the Eindhoven University of Technology, The Netherlands, in 2000, and the Ph.D. degree in artificial intelligence for autonomous underwater vehicles from the University of Limerick, in 2005. He is currently a Professor in Artificial Intelligence at the University of Limerick. His research interests include artificial intelligence and machine learning, with a particular focus on medical applications.