Self-Supervised Feature Learning for Medical Image Analysis

Liang Chen[a,b,*], Paul Bentley[b], Kensaku Mori[c], Kazunari Misawa[d], Michitaka Fujiwara[e], Daniel Rueckert[a]

[a]*BioMedIA Group, Department of Computing, Imperial College London, 180 Queen's Gate, London, SW7 2AZ, UK*
[b]*Division of Brain Sciences, Department of Medicine, Imperial College London*
[c]*Graduate School of Informatics, Nagoya University*
[d]*Aichi Cancer centre*
[e]*Nagoya University Hospital*

Abstract

Machine learning, particularly deep learning has boosted medical image analysis over the past years. Training a good model based on deep learning requires large amount of labelled data. However, it is often difficult to obtain a sufficient number of labelled images for training. In many scenarios the dataset in question consists of more unlabelled images than labelled ones. Therefore, boosting the performance of machine learning models by using unlabelled as well as labelled data is an important but challenging problem. Self-supervised learning presents one possible solution to this problem. However, existing self-supervised learning strategies applicable to medical images cannot result in significant performance improvement. Therefore, they often lead to only marginal improvements. In this paper, we propose a novel self-supervised learning strategy based on context restoration in order to better exploit unlabelled images. The context restoration strategy has three major features: 1) it learns meaningful image semantics; 2) it is useful for different types of subsequent image analysis tasks; and 3) its implementation is simple. We validate the context restoration strategy in three common problems in medical imaging: classification, localization, and segmentation. For classification, we apply and test it to scan plane detection in fetal 2D ultrasound images; to localise abdominal organs in CT images; and to segment brain tumours in multi-modal MR images. In all three cases, self-supervised learning based on context restoration learns meaningful semantic features and lead to improved machine learning models for the above tasks.

Keywords: Self-supervised learning, context restoration, medical image analysis

1. Introduction

Deep convolutional neural networks (CNNs) have achieved great success in computer vision, including image classification (Simonyan and Zisserman, 2014; Krizhevsky et al., 2012; Szegedy et al., 2015), object detection (Girshick, 2015; Ren et al., 2015) and semantic segmentation (Long et al., 2015; Chen et al., 2018). In medical image analysis, CNNs have also demonstrated significant improvement when applied to challenging tasks such as disease classification (Wang et al., 2017; Suk et al., 2014) and organ segmentation (Ronneberger et al., 2015; Çiçek et al., 2016; Kamnitsas et al., 2017). Large amounts of training data with manual labels have been crucial in many of these successes. In natural images, crowdsourcing can be used to obtain ground-truth labels for the images (Russakovsky et al., 2015). This is based on the fact that the annotation of natural images only requires simple human knowledge, e.g. most humans are able to recognize cars in natural images. However, crowdsourcing has limited applicability in medical imaging because annotation usually requires expert knowledge. This means it is usually easier to access a large number of unlabelled medical images rather than a large number of annotated images.

Training CNNs only using the small number of labelled images cannot always achieve satisfactory results and does not exploit the potentially large number of unlabelled images that may be available. The most straightforward method to make use of unlabelled data is to train an auto-encoder (Bengio et al., 2007) to initialise the task-specific CNN. However, the loss function used in auto-encoder is the L2 reconstruction loss which leads the auto-encoder to learn features that have limited value for discriminative tasks. The pretrained models from the natural image domain are not useful in the medical imaging domain since the intensity distribution of natural images is different from that of medical images.

Self-supervised learning is a type of machine learning strategy which has gained more and more popularity in recent years. It aims at supervised feature learning where the supervision tasks are generated from data itself. In this case, a very large number of training instances with supervision is available. Pretraining a CNN based on such self-supervision results in useful weights to initialise the subsequent CNN based on data with limited manual labels. Therefore, self-supervised learning is a good option to explore the unlabelled images to improve the CNN performance in case where only limited labelled data is available.

In this paper, we focus on self-supervision for medical images. Two existing self-supervised learning strategies are applicable in our cases, namely, the prediction of the relative posi-

*Corresponding author
Email address: liang.chen12@imperial.ac.uk (Liang Chen)

tions of image patches (Doersch et al., 2015) (the RP method) and local context prediction (Pathak et al., 2016) (the CP method). Figure 1 shows an example of these two methods. In the RP approach, a 3×3 patch grid is selected and the CNN learns the relative position between the central patch and one of its surrounding patches. For instance, a patch containing left cerebellum should locate at the bottom left corner of the patch of right cerebrum. In the CP method, a patch in the centre of image is selected and a CNN learns to predict its context using other image context.

We propose a novel self-supervised learning strategy for medical imaging. Our approach focuses on context restoration as a self-supervision task. Specifically, given an image, two small patches are randomly selected and swapped. Repeating this operation a number of times leads to a new image for which the intensity distribution is preserved but its spatial information is altered. A CNN is then trained to restore the altered image back to its original version. The proposed context restoration strategy has three advantages: 1) CNNs trained on this task focus on learning meaningful features; 2) CNN weights learned in this task are useful for different types of subsequent tasks including classification, localization, and segmentation; 3) implementation is simple and straightforward. We evaluate our novel self-supervised learning strategy in three different common problems in medical image analysis, namely classification, localization, and segmentation. Our evaluation uses different types of medical images: image classification is performed on 2D fetal ultrasound (US) images; organ localization is tested on abdominal computed tomography (CT) images; and segmentation is performed on brain magnetic resonance (MR) images. In all three tasks, the pretraining based on our context restoration strategy is superior to other self-supervised learning strategies, as well as no self-supervised training.

2. Related Work

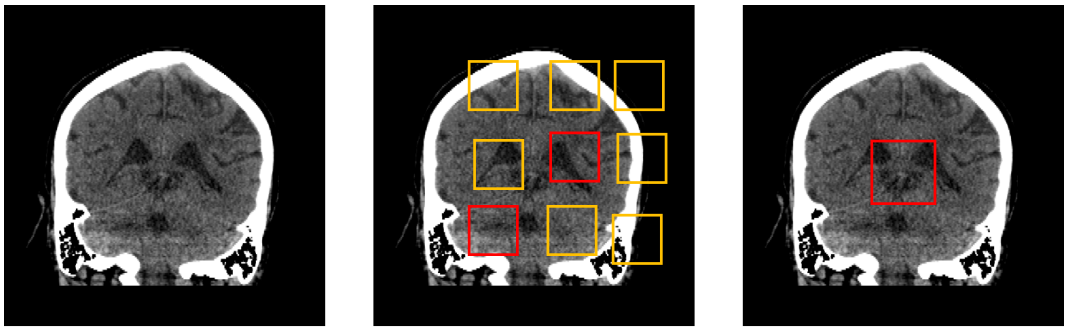
The key challenge for self-supervised learning is identifying a suitable self supervision task, i.e. generating input and output instance pairs from data. In computer vision, various types of self supervision have been proposed depending on data types, which is summarised in Table 1.

For static images, patch relative positions (Doersch et al., 2015; Noroozi and Favaro, 2016), local context (Pathak et al., 2016), and colour (Zhang et al., 2016, 2017) have been used in self-supervised learning. In the RP method, it was proposed to predict the relative positions between a central patch and its surrounding patches in a 3×3 patch grid (Doersch et al., 2015). The idea was that there are intrinsic position relations among divided parts of an object of interest. The RP method has three shortcomings: First, the relative position between two patches could have multiple correct answers, e.g. a patch of a car and a patch of a building. Second, it was reported that CNNs could complete the self-supervised learning tasks by learning trivial features, instead of meaningful features that are useful in other discriminative tasks such as classification and segmentation. Specifically, in the RP method, CNNs learns the shared edges or corners of two patches to predict their relative positions. Although techniques were proposed to address this effect, CNNs could still learn trivial features. For instance, it was proposed that patches are randomly jittered (Figure 1(b)) so that there is no shared information at edges or corners. However, the CNN may still learn patch positions from some background patterns. Third, the RP method is based on patches, which do not convey information about the global context of images. As a result, the RP method can only provide limited improvements for subsequent tasks requiring global context, such as classification. Later, a more complicated version of patch relative positions was proposed (Noroozi and Favaro, 2016), in which all 9 patches are input to CNNs in a random sequence. The CNNs were trained to find the correct sequence of the patches.

In terms of feature learning, learning to predict image context is more straightforward as proposed by Pathak et al. (Pathak et al., 2016). They proposed an idea which trains CNNs to learn how to inpaint missing information in images with patchy context removed. For the inpainting, an adversarial loss was proposed in addition to the L2 reconstruction loss while for feature learning only the L2 loss was used. They reported that if the removed patch is always in the centre of an image and in the square shape (Figure 1(c)), CNNs would only focus on the central context. As a result, patches with random shapes and in random locations were removed to improve the feature learning. However, the removal of context changes the image intensity distribution. Thus the resulting images belong to another domain and the learned features may not be useful for images in the original domain. Compared to the RP method, the CP method is more useful for the subsequent tasks. More precisely, the CNN weights learned in the CP method can be used to initialise subsequent CNNs for classification and segmentation; while CNN weights learned in the RP method can initialise subsequent classification CNNs and only the analysis part of the subsequent segmentation CNNs. This is because a CNN predicting the relative positions of patches is a classification model, which does not have layers to reconstruct image-level maps. Table 2 compares the RP method and the CP method in terms of subsequent task initialization.

Colour is one of the most important features in natural images. It was proposed that learning colours from greyscale images learns features that capture semantic information (Zhang et al., 2016), i.e. CNNs must implicitly perform object recognition in order to colour them appropriately. However, it is generally difficult to recognize if the weather is sunny or not in greyscale images. Therefore, learning semantics via colours is difficult to cover all aspects of stuff and things. In subsequent work, Zhang et al. (Zhang et al., 2017) proposed stronger supervision. Specifically, natural images were firstly converted into greyscale space and colour space. Then image representing each space was used to train a siamese CNN to predict the information in the other space. Combining the two outputs reconstructs the original image. This cross-supervision forces the CNNs to learn more meaningful semantics. In medical imaging, most images are in greyscale so that no colour information is available.

In addition, the exemplar learning has been proposed as a



(a) Original image (b) RP method (c) CP method

Figure 1: Demonstration of the RP and CP method on a brain CT image. (a) shows the original CT image in the coronal view. (b) shows the patch grid of the RP method and the red rectangles indicate patches of left cerebellum and right cerebrum. (c) shows the selected patch to be predicted.

Table 1: Summary of related literature. There are many self-supervision strategies have been proposed for natural images and videos while there is only one strategy relating to medical images.

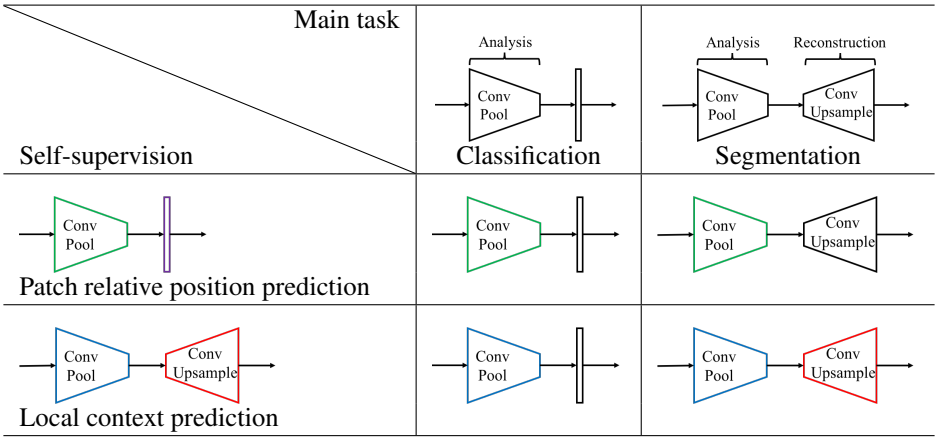
Data Type	Authors	Supervision
RGB images	Doersch et al. (Doersch et al., 2015) Noroozi et al. (Noroozi and Favaro, 2016)	patch relative position prediction
	Pathak et al. (Pathak et al., 2016)	local context prediction
	Zhang et al.(Zhang et al., 2016)	colourization
	Zhang et al.(Zhang et al., 2017)	colour-context cross prediction
	Dosovitskiy et al. (Dosovitskiy et al., 2016)	exemplar learning
Videos	Mobahi et al. (Mobahi et al., 2009) Jayaraman et al. (Jayaraman and Grauman, 2016)	temporal coherence
	Wang et al. (Wang and Gupta, 2015)	temporal continuous
	Walker et al. (Walker et al., 2015) Purushwalkam et al. (Purushwalkam and Gupta, 2016) Sermanet et al. (Sermanet et al., 2017)	object motion prediction
	Misra et al. (Misra et al., 2016) Fernando et al. (Fernando et al., 2017)	temporal order verification
Multi-modal data	Agrawal et al. (Agrawal et al., 2015) Jayaraman et al. (Jayaraman and Grauman, 2015)	ego-motion prediction
	Owens et al. (Owens et al., 2016) Chung et al. (Chung and Zisserman, 2017)	audio-video matching
MR images	Jamaludin et al. (Jamaludin et al., 2017)	follow-up scan recognition

self-supervised learning strategy (Dosovitskiy et al., 2016). In exemplar learning, the task is to classify each data instance into a unique class. In this case, heavy augmentation is required to generate training data. Since each data instance is regarded as one class, the exemplar learning method is difficult to apply to large datasets.

Image sequences (or videos) offer rich resources which could be used in self-supervised learning. First, neighbourhood frames should share similar features (Mobahi et al., 2009). Training CNNs to learn the similarities achieves the goal of learning contextual semantics. In addition, in events such as ball games, the features of frames representing a batting action should also be smooth, i.e. temporal continuous (Jayaraman and Grauman, 2016). Second, frames representing similar motions such as cycling should share similar visual features (Wang and Gupta, 2015). More generally, similar objects should share similar motions, which can be learned by CNNs (Walker et al., 2015). For instance, similar human poses should also share similar motions (Purushwalkam and Gupta, 2016; Sermanet et al., 2017). Third, frames representing actions should occur in a certain temporal order. This idea has led to the development of CNNs which learn whether a sequence of frames is in the correct order or not (Misra et al., 2016; Fernando et al., 2017).

Imaging data with multiple modalities can be easily used for self-supervised learning. The cross-supervision mentioned above is an obvious strategy to use for multi-modal imaging data. For instance, cameras at different angles offer different views. A siamese CNN could be trained to predict camera poses (Agrawal et al., 2015). More generally, images with the same ego-motion are likely to share similar features which can be learned by CNNs (Jayaraman and Grauman, 2015). For videos with audio, it is reasonable to assume similar events share sim-

Table 2: Comparison between the RP method and the CP method. Weights learned in both of them can initialise the subsequent classification CNN. Weights learned in the RP method can only initialise the analysis part of the subsequent segmentation CNN; while weights learning in the CP method can initialise analysis and reconstruction part of the subsequent segmentation CNN.



ilar audio sound (Owens et al., 2016). Exceptionally, in news broadcast videos, similar lip poses represent similar readings (Chung and Zisserman, 2017).

In medical imaging, patients often have follow-up scans. Recognizing scans of the same patient is a good method of self-supervised learning. Jamaludin et al. (Jamaludin et al., 2017) proposed a siamese CNN to recognize patients' MR scans and predict the level of vertebral bodies. A large number of scans was collected to train the CNN to recognize MR scans. Therefore, a small number of annotated scans is required for disease prediction. The above approach is one of the first works on self-supervised learning in medical imaging.

Our work also relates to the work of (Doersch and Zisserman, 2017), which proposed to combine multiple self-supervised learning tasks to improve the feature learning. In this work, patch relative position prediction (Doersch et al., 2015), colourization (Zhang et al., 2016), exemplar learning (Dosovitskiy et al., 2016), and motion segmentation (Pathak et al., 2017) were unified into one architecture. A novel input harmonization method was proposed to enable end-to-end training. Features learned in the individual tasks were then fused with an L1 penalty loss so that their combination could be sparse. The results showed that multi-task self-supervised learning improves subsequent tasks more than single-task self-supervised learning. The disadvantage of multi-task self-supervised learning is the training requires significant computational resources, i.e. 64 GPUs for approximately 16.8K GPU hours.

3. Self-supervision Based on Context Restoration

We propose a novel strategy for self-supervised learning which we term *context restoration*. We first introduce this concept before we provide further details of the training process.

3.1. Context Restoration

There are two steps in self-supervised learning based on context restoration: generating paired input/output images for training and learning a mapping between them. Given a dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ consisting of N images with no annotations, a new dataset

$$\tilde{\mathcal{X}} = f(\mathcal{X}) \tag{1}$$

is generated. Here $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_N\}$. $f(\cdot)$ is a function corrupting the context of original images. Subsequently, a CNN is learned to approximate the function $g(\cdot)$ which is designed to model the mapping $\tilde{\mathbf{x}}_i \mapsto \mathbf{x}_i$, i.e.

$$\mathbf{x}_i = g(\tilde{\mathbf{x}}_i) = f^{-1}(\tilde{\mathbf{x}}_i), \tag{2}$$

where $i = 1, 2, \ldots, N$.

Given an image $\mathbf{x}_i$, we randomly select two isolated small patches in $\mathbf{x}_i$ and swap their context. Repeating this process for T times results in $\tilde{\mathbf{x}}_i$. Figure 2 demonstrates this process on exemplar images and Algorithm 1 summarises the process in detail. Subsequently, $g(\cdot)$ aims to restore the context using CNN model by learning to approximate $f^{-1}(\cdot)$. This is illustrated in Figure 3.

Algorithm 1: Image context disordering

Input: original image $\mathbf{x_i}$
Output: image with disordered context $\tilde{\mathbf{x}}_\mathbf{i}$
for $iter = 1, 2, \ldots, T$ **do**
  randomly select a patch $p_1 \in \mathbf{x_i}$
  randomly select a patch $p_2 \in \mathbf{x_i}$
  $p_1 \cap p_2 = \emptyset$
  swap p_1 and p_2

Inspired by existing self-supervised learning strategies, a good self-supervised learning strategy should exhibit three key features: 1) features learned in the self-supervised training stage should be meaningful; 2) self-supervised pretraining is useful for different types of subsequent tasks; and 3) the implementation should be simple. Our proposed context restoration method

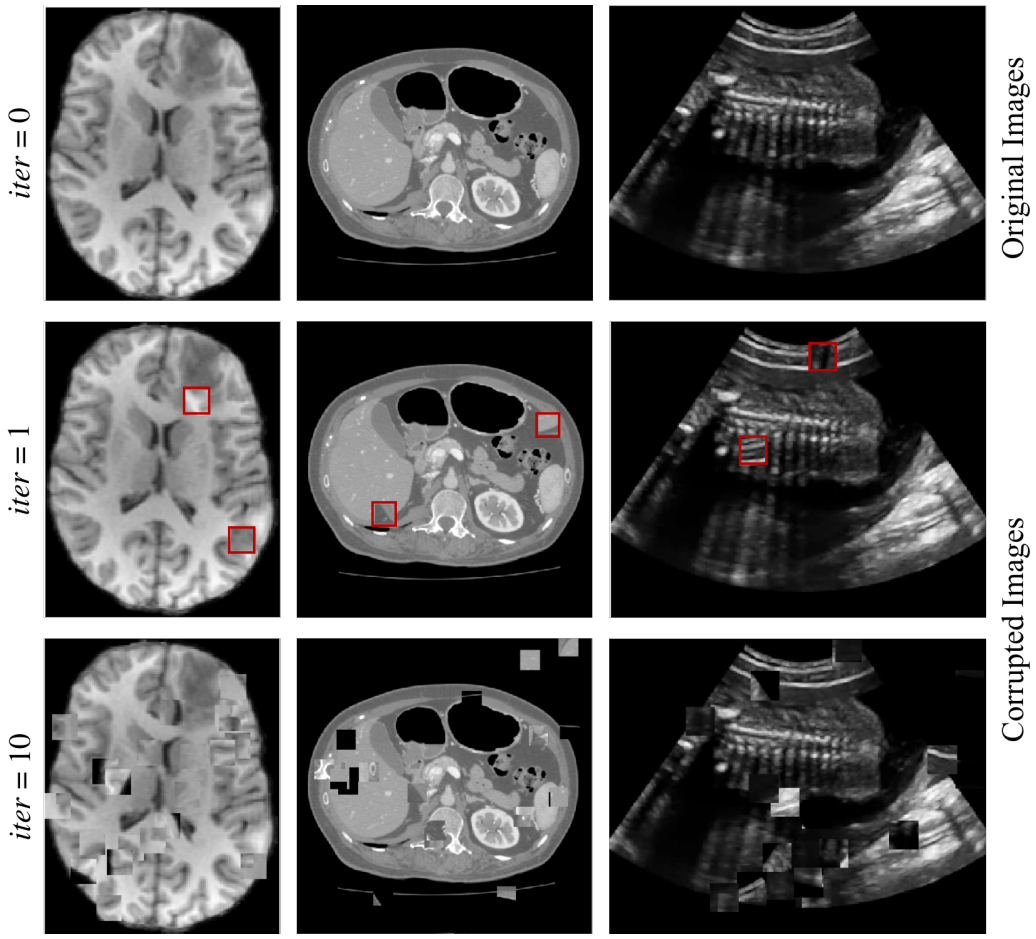


Figure 2: Generating training images for self-supervised context disordering: Brain T1 MR image, abdominal CT image, and 2D fetal ultrasound image, respectively. In figures in the second column, red boxes highlight the swapped patches after the first iteration.

features all these advantages. For many common problems in medical imaging such as classification, localization, and segmentation, learning image context is key. Therefore, learning the context of images in the self-supervised pretraining stage benefits the subsequent tasks. Restoring the image context can learn image context. Specifically, given the corrupted image $\tilde{\mathbf{x}}_i$, the $g(\cdot)$ function learns to restore it by solving two subtasks: 1) recognising which parts of the image contain corrupted context; 2) reconstructing the correct image context in these areas. Second, the proposed context restoration pretraining is applicable for different types of subsequent tasks by adjusting CNN architecture according to that of subsequent task. Finally, the implementation of the context restoration task is simple and straightforward.

3.2. Network Architectures

We model the proposed self-supervised learning strategy – context restoration – using CNNs. The CNNs can be implemented using various different architectures. Most of these networks are image-to-image networks consisting of two parts: an analysis part and a reconstruction part. Figure 3 shows an overview of the general architecture of feasible CNNs. The analysis part encodes input disordered images into feature maps and the reconstruction part uses these feature maps to produce output images in correct context.

Analysis Part: The analysis part consists of stacks of convolutional units and downsampling units, extracting feature maps from the input images. The convolutional units can be single convolution layers, residual convolution layers (He et al., 2016), inception layers (Szegedy et al., 2016), densely connected convolution layers (Huang et al., 2016) and so on. The downsampling units could be single pooling layers or inception pooling layers (Szegedy et al., 2016, 2017) and so on. The CNN weights learned in this part are then used to initialise the subsequent tasks.

Reconstruction Part: The reconstruction part consists of stacks of convolutional layers and upsampling layers, producing output images in which the context information has been restored. The upsampling layers can be deconvolution layers or other upsampling layers. Again, the CNN architectures used here are flexible. In subsequent classification tasks, the CNN weights learned in this part are not used. As suggested by (Doersch and Zisserman, 2017), simple CNN layers with a few deconvolution layers are sufficient (see Figure 3). In this condition, the analysis part makes most contributions to the context restoration. Therefore, the feature maps resulting from the anal-

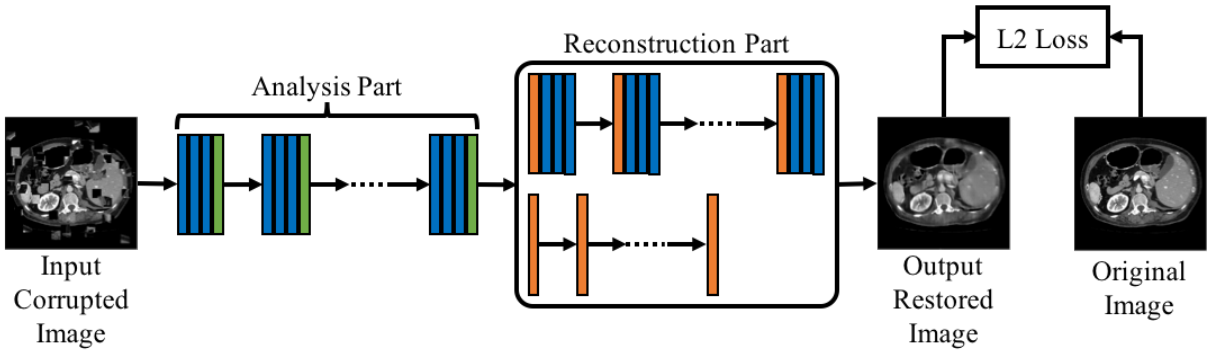


Figure 3: General CNN architecture for the context restoration self-supervised learning. In the figure, the blue, green, and orange strides represent convolutional units, downsampling units, and upsampling units, respectively. In the reconstruction part, CNN structures could vary depending on subsequent task type. For subsequent classification tasks, the simple structures such as a few deconvolution layers (2nd row) are preferred. For subsequent segmentation tasks, the complex structures (1st row) consistent with the segmentation CNNs are preferred.

ysis part are more useful. In subsequent segmentation tasks, the CNN weights learned in this part are then used. In this situation, the CNN architectures of the self-supervised learning and the subsequent main task learning can be consistent. As a result, almost all the weights of the subsequent segmentation CNN can be initialised using those learned in the self-supervised learning. This results in better segmentation results.

Loss Function: We propose to use the L2 loss for training the CNNs for the task of context restoration. As suggested by (Pathak et al., 2016), the L2 loss is sufficient for feature learning although the outputs from context restoration outputs may be blurry.

Implementation: In this work, the CNNs for context restoration employ single convolution layers as the convolutional units. In the analysis part, the architecture is similar to that of the VGG-Net (Simonyan and Zisserman, 2014), where there is a pooling layer following a few convolution layers. In the reconstruction part, if the subsequent task is a classification task, then there are only a few deconvolution layers; if the subsequent task is segmentation, then the reconstruction part is in symmetry with the analysis part with concatenation connections, which is similar to a U-Net architecture (Ronneberger et al., 2015). The loss function of CNNs in the subsequent tasks is the cross-entropy function.

All the CNNs use the Adam method (Kingma and Ba, 2015) for optimizing the loss function. We use $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$. The learning rates varies for the different problems. Batch normalization (Ioffe and Szegedy, 2015) is utilized in all CNNs. Random weights are used for initialization and sampled from a truncated normal distribution with standard deviation of 0.01. The kernel size of the convolution and deconvolution layers is 3×3. The stride size of the convolution layers is 1 and that of the deconvolution layers is 2.

The CNNs implemented in this paper use the Tensorflow[1] platform. Our experiments are performed on a desktop PC with an Core i7-3770 processor and 32GB RAM and with an NVIDIA TITAN XP GPU processor.

[1] https://www.tensorflow.org/

4. Experiments and Results

To evaluate the proposed self-supervision approach we have conducted four sets of experiments: First, we show the proposed self-supervision using context restoration task can be performed by CNNs on three different datasets, including brain MR images, abdominal CT images, and fetal US images. In addition, we use the pretrained CNNs for subsequent tasks such as classification, localization, and segmentation, respectively. For each of these problems, a different dataset is used. More importantly, we compare different self-supervised learning strategies, namely, training an auto-encoder (Bengio et al., 2007), self-supervision using patch relative position prediction (Doersch et al., 2015), self-supervision using local context prediction (Pathak et al., 2016), and the proposed context restoration. For each dataset, the self-supervised learning is based on the whole training set. The subsequent tasks are based on the whole, half, and quarter of the training set, respectively.

4.1. Context Restoration Results

We evaluate the CNNs employed for context restoration on three different datasets, including brain MR images, abdominal CT images, and fetal US images. Figure 4 shows examples of the three datasets. In all cases, the image context restoration achieve qualitatively good results. A shortcoming is that the L2 loss results in image blur.

4.2. Fetal Standard Scan Plane Classification

Overview: 2D US imaging is the most widely used medical imaging modality to assess the health of he fetus. In the UK, the fetal abnormality screening programme (FASP) handbook (NHS Screening Programmes, 2015) defines guidelines for selecting a number of standard scan planes, which are used to make biometric measurements and possible abnormalities. However, US images often have low quality because of noise, artefacts, shadows, etc. Therefore, interpreting fetal US images is challenging. Baumgartner et al. proposed a novel CNN-based approach (known as the SonoNet) to detect and localise the defined 13 different standard scan planes in real-time from US images (Baumgartner et al., 2017).

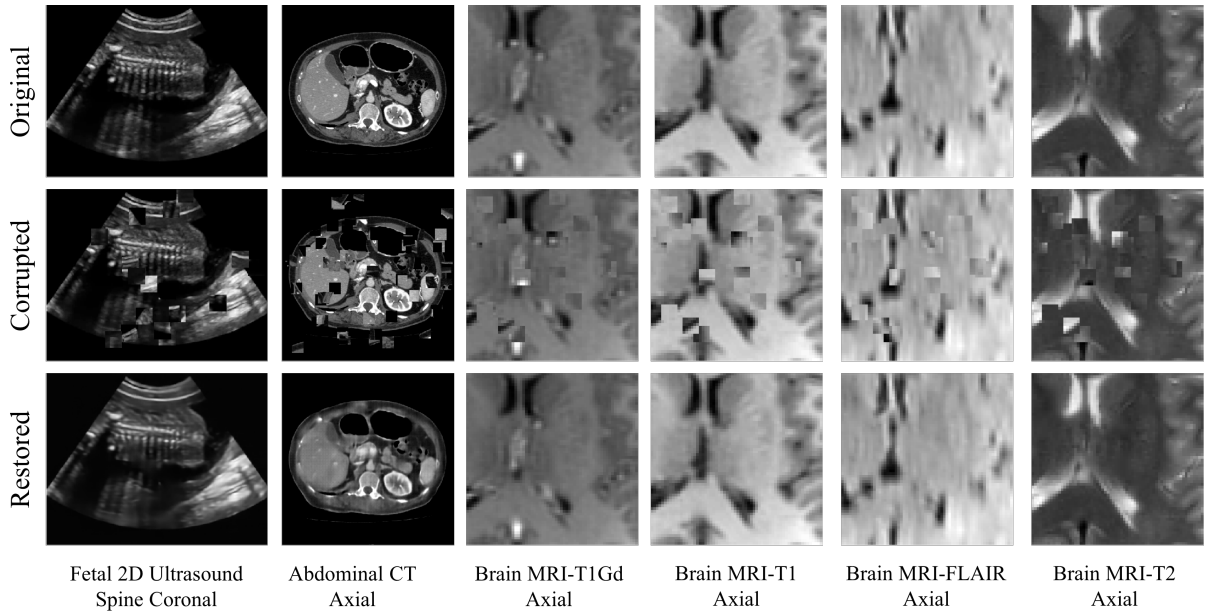


Figure 4: Self-srpervision using context restoration: For brain MR images, our training is on 2D image patch level. Therefore, the context restoration is also based on patches.

Dataset: We use the same dataset as used in (Baumgartner et al., 2017). Our dataset consists of 2694 2D ultrasound examinations of fetuses with gestational ages between 18 and 22 weeks. More details about the image acquisition protocol can be found in (Baumgartner et al., 2017). Figure 5 shows examples of each class of scan planes.

Implementation: The CNN for this classification problem is the SonoNet-64 which achieved the best performance in (Baumgartner et al., 2017). In terms of the training strategy, we use a fixed learning rate of 0.01. In the original training, each batch consists of 2 images from each of the standard scan plane categories and 26 images from background images. As a multi-class classification problem, the numbers of instances across classes are imbalanced. In our implementation, we sample the same number of background views with other classes.

Evaluation: As in (Baumgartner et al., 2017), we evaluate the performance of CNNs in this classification task using the precision, recall, and the F1-score.

Results: Table 3 displays the results of performance of the CNNs under different configurations. Balancing the numbers of instances in each class significantly improves the performance in all three metrics.

In training in random initialisation situations, it is not surprising that less training data leads to worse results. When the SonoNet is trained on half of the training data, the precision and recall both decrease, which lead to the decrease of the F1-score. Interestingly, when the SonoNet is trained on quarter of the training data, the precision decreases significantly while there is only slight decrease in terms of the recall. This suggests a large number of false positives (FPs) occur.

With the help of self-supervised pretraining, the performance of CNNs when using small training sets can be improved. Specifically, when learning on half of training images, the F1-scores keep stable in most cases except where the SonoNet is pretrained based on context restoration. In this scenario, the baseline (i.e. random initialisation) is not far away from the ceiling (i.e. SonoNet on the whole training set). Therefore, it is difficult to obtain improvements. The SonoNet pretrained using context restoration can only offers marginal improvement. When learning using only a quarter of training images, the SonoNet with feature initialisation from the auto-encoder pretraining still cannot improve the baseline; while SonoNets using other pretraining strategies perform better than the baseline. Our context restoration pretraining improves the SonoNet performance the most. This suggests that context restoration pretraining is more useful for image classification in this case.

4.3. Abdominal Multi-organ Localization

Overview: In many medical image analysis problems, localization anatomical structures is a prerequisite. For instance, in the liver segmentation challenge (Heimann et al., 2009) hosted in MICCAI 2007, the provided CT images were cropped such that the livers were roughly localized. This excludes irrelevant organs and tissue and benefits the segmentation. However, manual cropping requires expert knowledge and costly. de Vos et al. (de Vos et al., 2017) proposed a novel approach which can localize anatomical structures in 3D medical images. This approach defines the localization as discovering bounding boxes in 3D images so that regions within these bounding boxes contain target anatomical structures (see Figure 6). Following this idea, we localise multiple abominal organs in CT images. The organs of interest are pancreas, kidneys, liver, and spleen.

Dataset: A dataset of 3D abdominal CT image from 150 subjects is employed. The patient demographics and image acquisition details can be found in (Tong et al., 2015). We normalize the volume intensities in zero mean and unit deviation

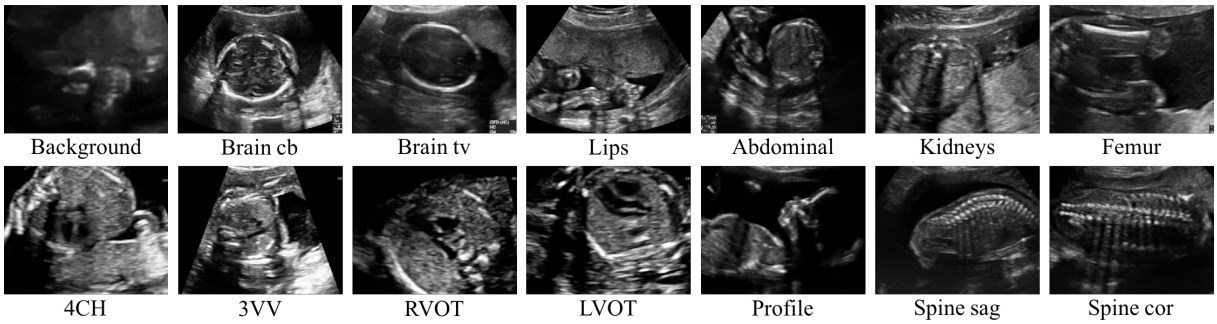


Figure 5: Examples of standard scan planes and background views of 2D fetal ultrasound images. The standard scan planes consist of brain view at the level of the cerebellum (Brain cb), brain view at posterior horn of the ventricle (Brain tv), coronal view of the lips and nose (Lips), standard abdominal view at stomach level (Abdominal), axial kidneys view (Kidneys), standard femur view (Femur), sagittal spine view (Spine sag), coronal spine view (Spine cor), four chamber view (4CH), three vessel view (3VV), right ventricular outflow tract (RVOT), left ventricular outflow tract (LVOT), and median facial profile (Profile).

Table 3: The classification of standard scan planes of fetal 2D ultrasound images. The entries in bold highlight the best comparable results.

Training data%	Initialisation	Precision (%)	Recall (%)	F1-score (%)
100% (Baumgartner et al., 2017)	Random	80.60	86.00	82.80
100%, Ours	Random	**89.39**	**89.66**	**89.42**
50%	Random	84.69	84.94	84.64
	Auto-encoder (Bengio et al., 2007)	84.63	86.09	84.50
	Relative positions (Doersch et al., 2015)	85.15	86.79	84.74
	Context prediction (Pathak et al., 2016)	84.43	85.27	84.43
	Context restoration	**85.52**	**87.56**	**85.94**
25%	Random	57.23	78.99	62.85
	Auto-encoder (Bengio et al., 2007)	55.54	82.87	62.32
	Relative positions (Doersch et al., 2015)	61.01	83.09	66.38
	Context prediction (Pathak et al., 2016)	57.73	81.58	63.10
	Context restoration	**65.69**	**85.25**	**69.93**

before analysis. The whole dataset is randomly divided into two equal halves. The first half is used for training and validation and the other half is used for testing. Images in this dataset were annotated at voxel level. We derive the reference bounding boxes and slice labels (organ presence) using these annotations.

Implementation: The CNN for multi-organ localization task is similar to the SonoNet (Baumgartner et al., 2017). It has one more stack of convolution and pooling layers than the SonoNet since the input images are 512×512 which is approximately twice larger than the processed 2D ultrasound frames in each side. The CNN for localization is also equipped with a global mean pooling layer. The output of this CNN is a prediction vector with K elements indicating the probabilities of presence of the K organs. The learning rate in this task is fixed as 0.001.

Evaluation: We follow (de Vos et al., 2017) that distances (in mm) from the reference bounding boxes to the predicted bounding boxes are used to evaluate the localization performance. Specifically, we compute the distances of the centroids and walls between bounding boxes.

Results: Table 4 displays localization performance of the CNN in different training strategies. Initialising by pretrained features, particularly those from context restoration tasks, improves the CNN performance.

Performance is compared among CNNs using different pretraining strategies. Training on incomplete training set using random initialization is used as baseline in each comparison group. Within each group, the CNN pretrained using the auto-encoder sometimes improves the performance upon the baseline. For instance, on half training data, it improves the centroid prediction of pancreas. However, it is worse than the baseline in terms of liver. In total, the results cannot verify auto-encoding pretraining improves the CNN performance. In contrast, pretraining based on relative position prediction and context prediction improves the CNN performance. Specifically, in most cases, pretraining of these two tasks decreases the errors on baselines in terms of both centroid and walls. Importantly, pretraining based on context restoration results in more localization improvements. In some cases, the CNN using context restoration pretraining is comparable to or even better than none pretraining on more annotated training data. For instance, in terms of left kidney, the CNN on half training data slightly outperforms that on all the training data; in terms of spleen, the CNN on a quarter training data performs better than the one on half training data. These improvements cannot be achieved by

Table 4: The performance of the CNN solving the multi-organ localization problem in different training settings. The entries in bold highlight the best comparable results. The RD, AE, RP, CP, CR are short for random, auto-encoder (Bengio et al., 2007), relative positions (Doersch et al., 2015), context prediction (Pathak et al., 2016), and our proposed context restoration. The numbers displayed are the mean ± std distances in mm.

Train data %	Init.	Left Kidney		Right Kidney	
		Centroid	Wall	Centroid	Wall
100%	RD	6.45 ± 8.47	3.68 ± 21.41	5.71 ± 10.17	2.79 ± 23.65
50%	RD	17.49 ± 49.67	9.36 ± 75.00	10.40 ± 30.37	5.89 ± 48.28
	AE	12.79 ± 38.67	6.84 ± 56.97	20.44 ± 41.48	11.52 ± 67.01
	RP	12.11 ± 39.01	6.75 ± 61.67	10.61 ± 30.41	5.77 ± 48.64
	CP	11.95 ± 38.97	6.82 ± 61.23	8.30 ± 11.92	4.47 ± 27.83
	CR	$\mathbf{5.99 \pm 9.83}$	$\mathbf{3.16 \pm 22.66}$	$\mathbf{5.83 \pm 10.10}$	$\mathbf{2.90 \pm 22.04}$
25%	RD	28.23 ± 71.95	15.87 ± 107.18	12.71 ± 30.39	6.77 ± 49.26
	AE	25.90 ± 65.64	14.40 ± 98.28	36.28 ± 73.65	19.55 ± 111.46
	RP	27.65 ± 75.31	15.41 ± 111.82	$\mathbf{8.34 \pm 11.22}$	$\mathbf{3.97 \pm 23.26}$
	CP	21.86 ± 60.28	13.03 ± 90.92	15.58 ± 35.3	8.42 ± 57.53
	CR	$\mathbf{7.63 \pm 9.02}$	$\mathbf{3.94 \pm 22.78}$	17.51 ± 52.67	9.8 ± 78.57

Train data %	Init.	Pancreas		Liver		Spleen	
		Centroid	Wall	Centroid	Wall	Centroid	Wall
100%	RD	13.39 ± 9.73	8.98 ± 23.27	7.50 ± 5.22	4.35 ± 14.07	6.63 ± 9.68	4.10 ± 23.02
50%	RD	16.45 ± 9.00	10.74 ± 26.77	12.79 ± 8.19	6.89 ± 22.6	13.24 ± 36.97	8.54 ± 56.87
	AE	15.59 ± 8.51	10.35 ± 24.35	14.07 ± 8.66	7.41 ± 24.39	12.36 ± 11.31	8.54 ± 31.16
	RP	15.54 ± 7.98	11.13 ± 23.50	10.12 ± 8.85	6.18 ± 22.31	7.64 ± 10.16	4.77 ± 24.41
	CP	14.76 ± 8.78	$\mathbf{10.07 \pm 26.26}$	9.91 ± 6.78	5.03 ± 15.39	7.79 ± 11.41	4.82 ± 25.98
	CR	$\mathbf{14.76 \pm 8.10}$	10.14 ± 24.86	$\mathbf{8.91 \pm 6.20}$	$\mathbf{4.67 \pm 16.83}$	$\mathbf{7.07 \pm 9.54}$	$\mathbf{4.05 \pm 22.17}$
25%	RD	22.09 ± 11.72	17.14 ± 39.23	12.02 ± 6.46	7.14 ± 20.27	24.86 ± 36.64	15.30 ± 61.38
	AE	17.67 ± 8.40	12.24 ± 25.54	16.79 ± 9.47	9.56 ± 28.30	22.65 ± 47.91	13.95 ± 73.05
	RP	17.84 ± 8.94	$\mathbf{11.74 \pm 25.06}$	15.59 ± 9.79	9.25 ± 29.74	14.51 ± 38.89	9.95 ± 62.12
	CP	21.81 ± 11.44	18.59 ± 41.57	11.40 ± 8.69	$\mathbf{6.18 \pm 22.50}$	10.34 ± 9.92	7.56 ± 27.58
	CR	$\mathbf{16.01 \pm 8.46}$	11.78 ± 28.79	$\mathbf{11.17 \pm 9.03}$	7.52 ± 25.68	$\mathbf{8.39 \pm 6.28}$	$\mathbf{5.82 \pm 19.50}$

CNNs using other pretraining strategies.

In terms of different organs, the distance variance of centroid and walls in kidneys is significantly larger than that of other organs. This is because not all patients have two kidneys. It is challenging for CNNs to distinguish two kidneys individually because of inter-subject variance. CNNs are more likely to make mistakes based on less training data. Although the CNN pretrained using the RP method on quarter training data outperforms that using context restoration pretraining, it performs much worse in left kidney. Regarding the pancreas, the performance of CNNs without pretraining decreases slightly when the training data halves. However, it decreases significantly when there is only quarter training data. In the opposite, in terms of the liver, the CNN performance decreases sharply with half training data; while it remains stable with quarter training data. On the spleen, the situation is different. The CNN performance keeps decreasing rapidly with less and less training data. It is noteworthy that if less training data leads to significant decrease of results, self-supervised learning is likely to improve the results significantly.

4.4. Brain Tumour Segmentation

Overview: Gliomas are the major brain tumours occurring in adults. They are routinely assessed using MR imaging (Bakas et al., 2017b). Accurate segmentation of gliomas on MR image is a key step for quantification. Our segmentation task is based on the Brain tumour segmentation (BraTS) chanllenge (Menze et al., 2015). The task is to segment the necrotic and non-enhancing tissues, the peritumoral edema, and gadolinium enhancing tissues of tumour (Bakas et al., 2017a) on multimodal MR images. Figure 7 shows such an example.

Dataset: We use the dataset of the BraTS 2017 challenge which consists of 285 subjects. Each subject has MR images in multiple modalities, namely, native T1 (T1), post-contrast T1-weighted (T1-Gd), T2-weighted (T2), T2 fluid attenuated inversion recovery (FLAIR). These images were preprocessed that images in different modalities are co-registered into the same anatomical template; skulls are removed; and voxels are resampled into the isotropic resolution ($1mm^3$) (Menze et al., 2015). Intensities are normalized to zero mean and unit variance. We use 142 out of the 285 images for training and validation and remaining 143 ones for testing.

Implementation: For the tumour segmentation in this work, we use a 2D patch-based CNN approach as suggested in (Kamnitsas et al., 2017; Chen et al., 2017) in medical image segmentation since medical images usually have large sizes while lesions of interest are small. Figure 4 shows an example of such

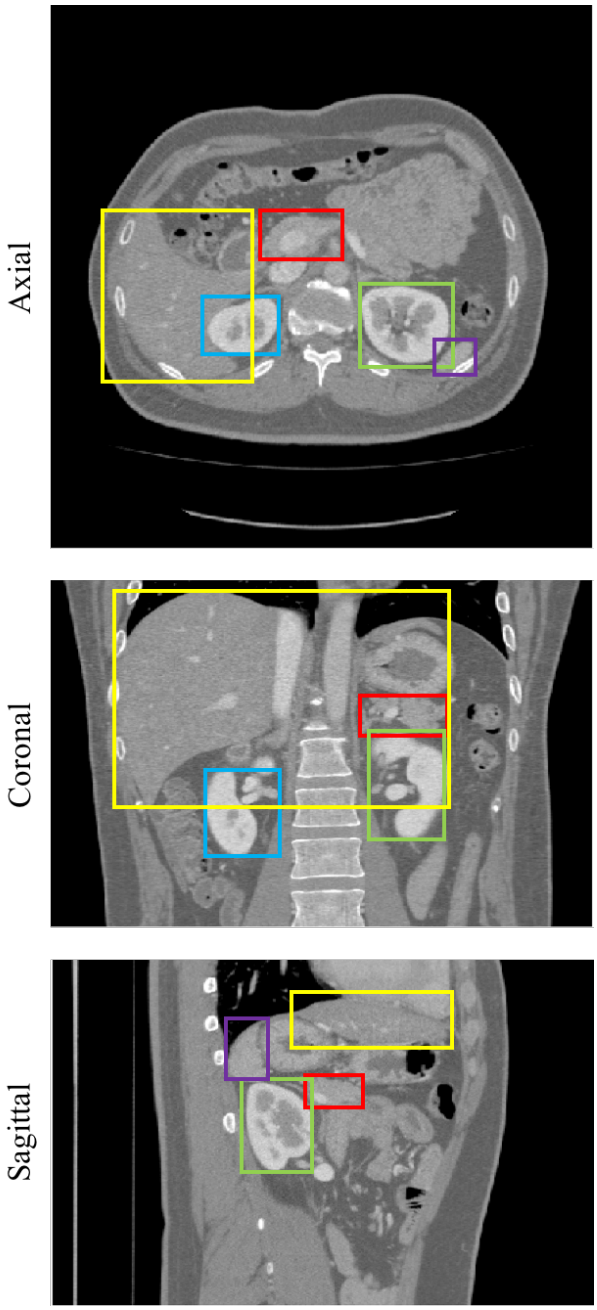


Figure 6: An example of abdominal CT image in axial, coronal, and sagittal views. The pancreas, left kidney, right kidney, liver, and spleen are colours in red, green, blue, yellow, and purple, respectively.

patches. The patch size used is 64×64. The CNN used in this experiment is a 2D U-Net (Ronneberger et al., 2015). The learning rate is fixed as 0.001. We follow the post-processing strategy proposed in (Kamnitsas et al., 2016): a 3D dense conditional random fields (CRFs) (Krähenbühl and Koltun, 2011) is used to refine the output of whole tumour structures; isolated voxel clusters of whole tumours less than 1000 voxel size are then removed based on the connected component analysis; the predicted voxels of tumour cores outside the regions of whole tumours are removed.

Evaluation: The evaluation is not based on three tumour classes individually. It is based on the following three classes: the whole tumour region which include all tumour structures, the tumour core region which include tumour structures except edema, and the enhancing tumour core region. We use the same evaluation metrics in the BraTS 2017 challenge: Dice score, sensitivity, specificity, and Hausdorff distance. Particularly, we use a robust version of the Hausdorff distance (Hausdorff95), which measures the 95% quantile, instead of the maximum distance between two surfaces.

Results: Table 5 shows the results on the BraTS problem. The general experiment settings are similar to the previous experiments. According to the results, U-Nets (Ronneberger et al., 2015) initialised by context restoration pretraining achieve the best performance in total.

In terms of different pretraining strategies, the auto-encoding pretraining does not improve CNN performance, which has been verified in previous experiments. This is also similar to the previous experiments that pretraining based on relative positions and context prediction tasks improves the segmentations but they are not as good as the pretraining based on the context restoration task. Again, self-supervision based on context restoration offers best pretraining startegy for the segmentation task.

The decrease in U-Net performance is not significant every time when the size of the training data halves. Therefore, the differences in performance among different self-supervision strategies are not significant. The performance using self-supervision based on context restoration approaches that of random initialisation on a larger dataset. For instance, using 50% of the training set, the proposed self-supervision strategy offers similar performance to using the whole training set. The Dice score in enhanced tumour core, the sensitivity in non-enhanced and enhanced tumour cores, and the Hausdorff distances in all aspects are even slightly better.

5. Discussion and Conclusion

In this paper, we proposed a novel self-supervised learning strategy based on context restoration. This enables CNNs to learn useful image semantics without any labels. The subsequent task-specific CNNs benefit from this pretraining. We conclude from the existing self-supervised feature learning literature that the ideal pretraining task should have similar goal to the subsequent task. Particularly, in medical image analysis, the image context is the common feature for classification, localization/detection, and segmentation tasks. Therefore, the context restoration learning contribute to learning features for these goals.

In addition, the CNNs for context restoration can be structured in flexible architectures depending on subsequent tasks. The idea is to ensure subsequent tasks can make full advantages of the weights from pretrained CNNs. Furthermore, the implementation of the context restoration task is simple and straightforward, meaning that it can be widely used. Compared with the existing strategies such as relative positions and context prediction, solving the context restoration task requires pattern

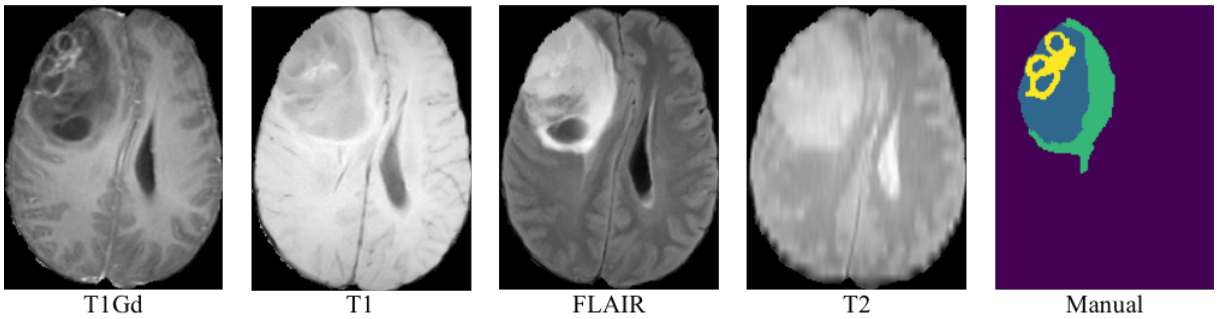


Figure 7: An example of MR image in multiple modalities with gliomas and the tumour structure annotations. In the manual annotation image, the background, edema, non-enhancing tumours, and enhancing tumours are coloured in purple, green, blue, and yellow, respectively.

Table 5: The segmentation results of the customised U-Nets (Ronneberger et al., 2015) in different training settings. The entries in bold highlight the best comparable results. The RD, AE, RP, CP, CR are short for random, auto-encoder (Bengio et al., 2007), relative positions (Doersch et al., 2015), context prediction (Pathak et al., 2016), and our proposed context restoration.

Train data %	Init.	Dice %			Sensitivity %			Specificity %			Hausdorff95		
		Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	Enh.
100%	RD	86.56	77.04	66.31	87.05	77.28	77.62	99.88	99.94	99.95	30.78	25.03	25.74
50%	RD	84.41	75.55	65.11	**84.75**	77.76	80.2	99.86	99.91	99.94	31.29	25.26	26.81
	AE	84.33	71.85	65.07	84.71	74.19	77.38	99.87	99.91	**99.95**	33.36	25.24	24.56
	RP	84.38	75.65	66.73	84.65	77.02	79.48	99.87	**99.92**	**99.95**	36.43	23.15	**20.69**
	CP	84.54	73.86	66.01	84.59	75.28	79.46	99.86	**99.92**	99.94	33.59	28.59	26.90
	CR	**85.57**	**76.2**	**68.24**	83.83	**78.17**	**80.53**	**99.89**	**99.92**	**99.95**	**26.41**	**20.34**	24.38
25%	RD	81.91	71.22	62.57	84.08	75.68	75.98	99.82	99.89	**99.94**	36.34	37.21	31.57
	AE	83.05	68.92	61.28	83.90	76.52	76.75	**99.85**	99.86	99.93	33.21	34.9	31.95
	RP	82.38	71.33	61.86	84.23	72.53	75.38	99.83	**99.92**	**99.94**	37.83	**31.81**	31.04
	CP	83.19	71.55	62.77	**85.75**	73.68	76.88	99.83	99.91	**99.94**	36.21	36.45	31.90
	CR	**84.27**	**73.43**	**64.12**	85.57	**78.79**	**79.14**	**99.85**	99.89	**99.94**	**33.15**	32.18	**30.61**

recognition and prediction, which ensures the context restoration task offers more efficient image semantics.

We have validated the proposed context restoration pretraining on three types of representative tasks in medical image analysis, which are classification, localization, and segmentation. Each of these tasks are based on a different type of medical images. The classification task is based on fetal 2D ultrasound images; the localization task is based on abdominal CT image; and the segmentation task is based on multi-modal brain MR images. In all three tasks, context restoration pretraining outperforms other pretraining methods. These results underlines the advantages of our context restoration strategy. In our experiments, we found that if the reduction of training data causes significant performance decrease, the context restoration pretraining can offer significant performance improvement over the baselines.

In computer vision, many CNNs are pretrained before the main task. For instance, the Faster R-CNN (Ren et al., 2015) is based on the pretraining of the VGG-Net (Simonyan and Zisserman, 2014). This type of pretraining leads to good detection results in the Faster R-CNN. However, it was reported that the self-supervised pretraining is not as good as the supervised pretraining (Larsson et al., 2017). This is not verified in this paper since in medical image analysis, it is difficult to conduct supervised pretraining, which requires a large number of annotations. However, it is noteworthy to exploring more powerful self-supervised learning method so that the self-supervised pretraining can be as good as supervised pretraining in the future.

Acknowledgement

This work was supported by the Wellcome Trust IEH Award under Grant 102431. We acknowledge the kind donation of the GPUs from the NVidia.

References

References

Agrawal, P., Carreira, J., Malik, J., 2015. Learning to see by moving, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 37–45.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C., 2017a. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive .

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017b. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data 4, 170117.

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D., 2017. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Transactions on Medical Imaging 36, 2204–2215.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks, in: Advances in Neural Information Processing Systems, pp. 153–160.

Chen, L., Bentley, P., Rueckert, D., 2017. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. NeuroImage: Clinical .

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence PP, 1–1.

Chung, J.S., Zisserman, A., 2017. Lip reading in profile, in: Proceedings of the British Machine Vision Conference, pp. 1–11.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-net: learning dense volumetric segmentation from sparse annotation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 424–432.

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1422–1430.

Doersch, C., Zisserman, A., 2017. Multi-task self-supervised visual learning, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2051–2060.

Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., Brox, T., 2016. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 1734–1747.

Fernando, B., Bilen, H., Gavves, E., Gould, S., 2017. Self-supervised video representation learning with odd-one-out networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5729–5738.

Girshick, R., 2015. Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heimann, T., Van Ginneken, B., Styner, M.A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., et al., 2009. Comparison and evaluation of methods for liver segmentation from ct datasets. IEEE Transactions on Medical Imaging 28, 1251–1265.

Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L., 2016. Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International Conference on Machine Learning, pp. 448–456.

Jamaludin, A., Kadir, T., Zisserman, A., 2017. Self-supervised learning for Spinal MRIs, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 294–302.

Jayaraman, D., Grauman, K., 2015. Learning image representations tied to egomotion, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1413–1421.

Jayaraman, D., Grauman, K., 2016. Slow and steady feature analysis: higher order temporal coherence in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3852–3861.

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A., Criminisi, A., Rueckert, D., Glocker, B., 2016. DeepMedic on brain tumor segmentation, in: Proceedings of the MICCAI-BRATS 2016 Multimodal Brain Tumor Image Segmentation Benchmark: "Change Detection", pp. 18–22.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis 36, 61–78.

Kingma, D., Ba, J., 2015. Adam: a method for stochastic optimization, in: Proceedings of the International Conference on Learning Representations.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected CRFs with Gaussian edge potentials, in: Advances in Neural Information Processing Systems, pp. 109–117.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Proceedings of the Conference on Neural Information Processing Systems, pp. 1097–1105.

Larsson, G., Maire, M., Shakhnarovich, G., 2017. Colorization as a proxy task for visual understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 6874–6883.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging 34, 1993–2024.

Misra, I., Zitnick, C.L., Hebert, M., 2016. Shuffle and learn: unsupervised learning using temporal order verification, in: Proceedings of the European Conference on Computer Vision, pp. 527–544.

Mobahi, H., Collobert, R., Weston, J., 2009. Deep learning from temporal coherence in video, in: Proceedings of the International Conference on Machine Learning, pp. 737–744.

NHS Screening Programmes, 2015. Fetal Anomaly Screen Programme Handbook. NHS.

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles, in: Proceedings of the European Conference on Computer Vision, pp. 69–84.

Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A., 2016. Ambient sound provides supervision for visual learning, in: Proceedings of the European Conference on Computer Vision, pp. 801–816.

Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B., 2017. Learning features by watching objects move, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2701–2710.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544.

Purushwalkam, S., Gupta, A., 2016. Pose from action: Unsupervised learning of pose features based on motion. arXiv preprint arXiv:1609.05420 .

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, pp. 91–99.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252.

Sermanet, P., Lynch, C., Hsu, J., Levine, S., 2017. Time-contrastive networks: Self-supervised learning from multi-view observation. arXiv preprint arXiv:1704.06888 .

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Suk, H.I., Lee, S.W., Shen, D., Initiative, A.D.N., et al., 2014. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage 101, 569–582.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI conference on Artificial Intelligence, pp. 4278–4284.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., Mori, K., Hajnal, J.V., Rueckert, D., 2015. Discriminative dictionary learning for abdominal multi-organ segmentation. Medical Image Analysis 23, 92–104.

de Vos, B., Wolterink, J., de Jong, P., Leiner, T., Viergever, M., Isgum, I., 2017. ConvNet-based localization of anatomical structures in 3D medical images. IEEE Transactions on Medical Imaging .

Walker, J., Gupta, A., Hebert, M., 2015. Dense optical flow prediction from a static image, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2443–2451.

Wang, X., Gupta, A., 2015. Unsupervised learning of visual representations using videos, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3462–3471.

Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization, in: Proceedings of the European Conference on Computer Vision, pp. 649–666.

Zhang, R., Isola, P., Efros, A.A., 2017. Split-brain autoencoders: Unsupervised learning by cross-channel prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1058–1067.