Received August 14, 2020, accepted August 30, 2020, date of publication September 3, 2020, date of current version September 17, 2020.
Digital Object Identifier 10.1109/ACCESS.2020.3021469

Self-Supervised Learning Based on Spatial Awareness for Medical Image Analysis

XUAN-BAC NGUYEN, GUEE SANG LEE, SOO HYUNG KIM, AND HYUNG JEONG YANG
Department of Electronic and Computer Engineering, Chonnam National University, Gwangju 61186, South Korea
Corresponding author: Guee Sang Lee (gslee@jnu.ac.kr)

This work was supported in part by the Bio and Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean Government (MSIT) under Grant NRF-2019M3E5D1A02067961 and Grant HCRI 19136, and in part by the Chonnam National University Hwasun Hospital Institute for Biomedical Science and National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) under Grant NRF-2020R1A4A1019191.

ABSTRACT
Medical image analysis is one of the research fields that had huge benefits from deep learning in recent years. To earn a good performance, the learning model requires large scale data with full annotation. However, it is a big burden to collect a sufficient number of labeled data for the training. Since there are more unlabeled data than labeled ones in most of medical applications, self-supervised learning has been utilized to improve the performance. However, most of current methods for self-supervised learning try to understand only semantic features of the data, but have not fully utilized properties inherent in medical images. Specifically, in CT or MR images, the spatial or structural information contained in the dataset has not been fully considered. In this paper, we propose a novel method for self-supervised learning in medical image analysis that can exploit both semantic and spatial features at the same time. The proposed method is experimented in the problems of organ segmentation, intracranial hemorrhage detection and the results show the effectiveness of the method.

INDEX TERMS
Self-supervised learning, spatial awareness, medical image analysis, deep learning, convolutional neural network.

I. INTRODUCTION
For the past decades, deep learning and artificial intelligence have been rising as hot topics in computer vision and image processing. The main reason is obviously their superb effectiveness outperforming the conventional algorithms. Traditional methods require the manual feature extraction followed by classification algorithms, while deep learning methods provide the compelling ability to automatically learn multi-level visual features from raw or minor amended images and successfully accomplished a variety of tasks such as image classification [1]–[3], object detection [4]–[6] and semantic segmentation [7]–[9]. Not only are these deep neural networks successful in general tasks but they also proved the dominance over traditional methods in medical imaging. In general, there are two main things leading to the high performance of deep convolutional neural networks: the large amount of trained data and the inherent capability of the network. With regard to network capacity, many models have been formed and replaced each other to become the state-of-the-art performers in classification, namely AlexNet [2], VGG [1], GoogLeNet [3], ResNet [10], DenseNet [11], etc. For large datasets, ImageNet [12] and OpenImage [13] are the representative ones that consist of millions of images and hundreds of labels. Mostly, with the larger dataset for training, the model efficiency increases. Nonetheless, large labelled datasets are seldom available, while unlabelled dataset is relatively easier to obtain, especially in the medical imaging applications. Self-supervised learning (SSL) was introduced as a new machine learning strategy to leverage unlabelled dataset. SSL generates useful visual features via an objective task which is related to the main task. The result is a form of transfer learning through pre-training of the network using unlabelled dataset and finally enhancing the performance of the main model. The objective task accompanying the self-supervised learning is called the pretext task. There are varied pretext tasks and each pretext task is proposed to learn different kinds of visual features. Some examples of typical pretext tasks are context prediction, colorizing gray-scale images [14], image
in-painting [15], image jigsaw puzzle [16], and Rotatopm [] etc., and the choice of pretext tasks depends a lot on what the main task is. For example, for CT images the colorizing option is not applicable. The labels for the pretext tasks are generated automatically from the input data, e.g. the input image is the ground truth of the in-painting task.

Medical imaging such as organ segmentation and disease classification commonly could not contribute a large and comprehensive dataset because the medical data is often limited in quantity and the annotation work demands extensive experiences in clinical practices. Therefore, they certainly benefit a lot from the pre-trained models when a large-scale unlabelled dataset is available. Besides the huge advantage, there are still limitations to medical utilization. The pre-trained models on natural images are not able to pose their potential since the intensity distribution of natural images is not correlated with that of medical images. Hence, we need a solution to the shortage of labeled data for performance improvement in medical datasets, which provides a motivation for the utilization of SSL in medical imaging.

SSL in medical imaging is in its early stage and mostly the existing methods use semantic information in the pretext tasks. However, the structural information is essential in many medical imaging applications, which is not fully considered in the previous methods of SSL. The contribution of this paper lies on the introduction of the spatial information together with semantic information in the pretext tasks, thus resulting in the pre-training of spatial structures. To summarize, in this paper, our contributions are:

- It is the first time that self-supervised learning based on spatial awareness is proposed for medical image analysis.
- We proved that our proposed method works well in various problems of medical field, specifically when the dataset contains spatial information as in CT or MR images.

II. RELATED WORKS
In self-supervised learning, there are two phases of training, which are self-supervised pretext task training and supervised downstream task training. The downstream task is actually the common task like image classification, object detection or semantic segmentation of which we want to improve the performance. The pretext task could be self-designed for ConvNets to learn the visual features of the small datasets from the pseudo labels which we are able to generate based on the attributes of data. After training the pretext task, its parameters function as the pre-trained weight for the downstream task, transferring the knowledge from objectively auxiliary task to major task. Pretext task varies in several types and many ways of implementation. Depending on the useful features corresponding to the downstream task, the option of pretext task must be carefully considered so that the transferred knowledge is beneficial to the further training.

According to [17], based on the data attributes, pretext tasks can be divided into four groups, which are generation-based methods, context-based methods, semantic label-based methods and cross-modal-based methods. Generation-based methods extract the features by creating analogous data and are able to be applied to both images and videos. These include colorization [14], image super resolution [18], image inpainting [15], data generation with Generative Adversarial Networks (GANs) [19]–[22], video future prediction [23]. In image colorization, CNNs are assigned to convert greyscale images into colored images and at the same time, they learn the semantic information which gives clues to the coloration of objects. Image inpainting teaches the models what are missing in different areas of a picture so that models implicitly have the awareness of object location. The next methods for pretext tasks are context-based methods, which focus on context similarity, spatial structure and temporal structure. Methods for this kind of pretext tasks are image clustering [24], [25], graph constraint [26] (context similarity), image jigsaw puzzle [16], [27]–[29], context prediction [30], geometric transformation recognition [31] (spatial structure), frame order correction [32], frame sequence prediction [33] (temporal structure). Image clustering is a technique that automatically gathers images with similar characteristics into groups and then transfers the knowledge to the Convnets. Image jigsaw puzzle or image context prediction aims at the understanding of spatial relationship through the position prediction of patches which are cropped and shuffled from one image. In videos, frames in a small sequence regularly have little alteration in the content and follow a rational temporal order. The temporal-based pretext task may change the order of frames in a video and the model could learn the strict chronological connection of frame sequences. Semantic label-based methods are based on the trivial attributes of objects such as contour or depth and include the tasks of contour detection [26], [34], relative depth prediction [35]. Cross modal-based methods emphasize the concurrence of multi-input such as different angles of one object by a few cameras or the harmony of audio and visual in a specific context. These methods include Visual-Audio Correspondence Verification [36], [37], and RGB-Flow Correspondence Verification [38]. To sum up, depending on the downstream task, we can use one of the mentioned pretext tasks for our strategy or even build one which best suits what features we want to extract.

In medical imaging, the common used types of images are Magnetic Resonance Imaging (MRI) scan and Computer Tomography (CT) scan. Both techniques generate a sequence of images from the top to bottom or from left to right of the patient’s body. Due to the high complexity of annotating these scans, the datasets related to them are not sufficient. In [39], Liang et al. proposed a self-supervised method based on context restoration. The medical image was divided into patches and two patches were randomly exchanged. The pretext task will try to bring back the original image by learning the semantic features. Despite the effectiveness of context restoration for abdominal multi-organ localization and brain tumor segmentation, this pretext task focuses only
on two-dimensional semantic features, or the context of a single slice, while MR scans and CT scans mostly consist of many slices and they require the learning of 3D structures contained in the input data.

The spatial context holds an important role contributing to the performance, our pretext task is designed to learn not only the visual features of each slice but also the relationship between neighborhood slices. We do not utilize the auto-encoder network like the common pretext tasks for context restoration because the training process takes a great amount of time and includes latent complexity. Instead of that, a normal classification network is employed. Our network does not attempt to restore the edited image but detects the locations of corrupted, abnormal and unharmed area. Besides, it will predict the order of slices in a sequence. Therefore, both spatial and semantic features are involved in the perception of main model.

### III. MATERIAL AND METHODS

#### A. SELF-SUPERVISED LEARNING BASED ON SPATIAL AWARENESS

As the typical self-supervised learning method, in the pretext task, we generate input images and their ground truths. Our SSL method works with all the data which contain spatial, 3D information such as: CT image and MR image. Given dataset \( D = \{i_1, i_2, \ldots, i_n\} \) contain \( n \) samples without any annotation. We denote: \( S^m_k \) is \( m^{th} \) slice of \( k^{th} \) image where \( 1 \leq m \leq s_k \) and \( s_k \) is the number of slice that image \( i_k \) contains. The image \( \tilde{S}^m_k = f(S^m_k, S^{m+\delta}_k) \) is a generated by using two slices \( S^m_k \) and \( S^{m+\delta}_k \). We call \( \delta \) is spatial index representing for spatial information. The process to generate \( \tilde{S}^m_k \) is described as follows. At the slice \( S^m_k \), we random select \( \delta \). The range values of \( \delta \) are pre-defined as a configuration. The function \( f(.) \) takes two parameters which are \( m^{th} \) and \((m+\delta)^{th}\) slice of image \( i_k \). A rectangle patch \( P^m_{x+y+w+h}(x, y, w, h) \) of \((m+\delta)^{th}\) slice is randomly selected at the position \((x, y)\) and has width and height of \((w, h)\) correspondingly. At the next step, we replace \( P^m_{x+y+w+h}(x, y, w, h) \) by \( P^{m+\delta}_{x+y+w+h}(x, y, w, h) \) at the same position. There is a chance that \( P^m_{x+y+w+h}(x, y, w, h) = P^{m+\delta}_{x+y+w+h}(x, y, w, h) \) in the case \( \delta \neq 0 \). Thus, we do the replacing process in \( T \) times to make sure the generated image contains the corrupted patches. Algorithm 1 summarizes the process in the detail.

The example of generated image is illustrated in Figure 2.

The pretext model is trained by two objectives: 1.) Classify if the input image is normal or there are some corrupted patches exist inside. 2.) If there are corrupted patches inside, the model is able to know which slice do these patches come from. The first objective guides models to learn the semantic feature while the second helps model awareness of the spatial information. More specially, model is trained to be able to predict values of \( \delta \) from the image \( \tilde{S}^m_k \). We treat this training process as classification problem. The following reasons explain why \( \delta \) is representing for both semantic and spatial information. Firstly, \( \delta \) can be any integer, negative or positive values. When \( \delta = 0 \), we have \( \tilde{S}^m_k = S^m_k = S^{m+\delta}_k \). The original image is kept without any modifications and corrupted patches. Similarly, when \( \delta \neq 0 \), \( \tilde{S}^m_k \neq S^m_k \), there are some broken of border or heterogeneous color and contrast inside the image. If a model is successful in predicting values of \( \delta \) equals to zero or non-zero, the model understands how normal of objects should be. In other words, models already utilize the semantic features. Secondly, in the case \( \delta \neq 0 \), when a model predicts \( \delta \) correctly, the model is able to answer the question: which slice do the abnormal patches come from?! It means the model remember and understand the overall 3D structure of objects in the image. In conclusion, by predicting \( \delta \) a model learns both semantic and spatial features.

#### B. PRETEXT MODEL

The model we used for the pretext task is using CNNs. There are various architectures such as: Residual network, Densenet, VGG, etc. that is often used for the classification problem. The model then is used as an encoder in the segmentation task or transferred learning backbone in the classification task. In our work, we choose ResNet34 for training pretext task as shown in the figure 1.

#### C. LOSS FUNCTION

By learning to predict value of spatial index, the model is able to learn both semantic and spatial features at the same time. It means that the pretext model has only one target but satisfies two objectives. Therefore, we use only one cross-entropy function for the training pretext model, which is sufficient for the classification.

---

**Algorithm 1 Generate Corrupted Images Using Spatial Information for Pretext Task**

| Input: original slice image \( S^m_k \) |
| Output: image with spatial mixing \( \tilde{S}^m_k \) and value of \( \delta \) |

- R is range of \( \delta \) |
- \( s_k \) is the number of slice of \( k^{th} \) image |
- \( T \) is the number of repeat times |
- \((w, h)\) are the width and height of slice correspondingly |

**Repeat**

- random \( \delta \) in range of R |

**Until** \( 0 \leq m + \delta \leq s_k \):

- select slice: \( S^{m+\delta}_k \) |

**For** \( t = 1, t \leq T \) **Do**

**Repeat**

- random select \((x, y)\) |

**Until** \( x + w \leq W, y + h \leq H \):

- replace \( P^m(x, y, w, h) \) by \( P^{m+\delta}(x, y, w, h) \)
FIGURE 1. Overview of our proposed self-supervised learning method for Organ-at-risk segmentation. The system includes two stages: training pretext task (classification) and training the main task (segmentation). At the first stage, the backbone Resnet34 is trained to predict value of spatial index $\delta$. The weights of this stage will be used as initialized parameters for the encoders of segmentation task in second stage. In the case of Intracranial hemorrhage detection, the main task is different.

FIGURE 2. Generating corrupted image using spatial information for the pretext task. (a) is the $m^{th}$ slice, (b) is $(m + \delta)^{th}$ and (c) is the corrupted image containing spatial information. The bounding box with different colors representing for the different random patches cropped and replaced from slice $(m + \delta)^{th}$.

D. IMPLEMENTATION

Our model is implemented by using Pytorch\textsuperscript{1} framework. For the further research, we release the source code in here.\textsuperscript{2} We use Adam optimizer [40] to optimize the loss function. The parameters of optimizer are: $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-8}$. The model is trained in 50 epochs following One Cycle [41] scheduler with warm-up. The learning rate increases from 0.0001 to 0.0005 in the first 5 epochs, then decreases gradually to 0.000001 in the rest of the process.

In the pretext task, we train the model from scratch with a small dataset (i.e 50% of the whole dataset), we apply augmentation for more generalization: horizontal flip, elastic transformation, grid distortion, optical distortion, shift-scale and rotation. All the models in this paper are experimented on a desktop PC with AMD Ryzen 7 2700X equipped a NVIDIA GTX 2080Ti GPU processor.

IV. EXPERIMENTS AND RESULTS

To evaluate the performance of our proposed self-supervised learning, we did two experiments which are the most common problems in the medical image analysis. The first problem is Organ-at-risk segmentation and the second one is Intracranial Hemorrhage Detection. Both of them have datasets of CT images. We compare them in different SSL methods, namely: Jigsaw and context restoration [39]. For each dataset, the SSL is based on the whole and a half of the training set, respectively.

A. ORGAN-AT-RISK SEGMENTATION

1) DATASET OVERVIEW

We used StructSeg dataset which has been used in the StructSeg2019 challenge to evaluate the performance of our proposed method. The dataset has CT scans of 60 lung cancer patients where 50 patients for training and 10 patients for testing. Each scan is annotated by one expert and verified by another one. There are six annotated OARs with the important weights: left lung (100), right lung (100), spinal cord (100), esophagus (70), heart (100), trachea (80). The number of slice per patient varies from 60 up to 90 slices. Figure 3 shows an example of dataset.

FIGURE 3. Example of OARs from StrugSeg-2019 challenge. (a) 2D image, (b) 3D image. Each OAR is shown in a different color.

2) DATASET PREPROCESSING

For each patient, we extract all the slices from scanned 3D volume to 2D images. The Houndsfield Unit (HU) values are clipped and mapped into $[0-1]$ range as follows:

$$I = \frac{I - L}{U - L}$$  \hspace{1cm} (1)

where $U, L$ are the upper and lower boundary of HU. We select: $U = 1000, L = -400$ in our experiments.

\textsuperscript{1}https://pytorch.com
\textsuperscript{2}https://github.com/ngxbac/self-supervised-learning-spatial-awareness
The weighted cross entropy for multi-classes segmentation is defined as follows. The range of spatial index $\delta$ is $[-2, 2]$. Since $\delta$ is the integer, the possible values of $\delta$ each time sampled randomly are $-2, -1, 0, 1, 2$. Thus, our pretext task is a 5-class classification problem. The number of time that we random replace patches repeatedly is $T = 50$ and each patch is $20 \times 20$ pixels. We trained Resnet34 as the pretext model and achieved 97% accuracy on predict value of spatial index $\delta$. We analyze the Gradient Class Activation Map (Grad-CAM) to visualize how model looks at the corrupted input image as the figure below. It is clear that a pretext model performs well on detecting normal/abnormal and how it looks like in the next or previous slice.

3) PRETEXT TASK

The pretext task for organ-at-risk segmentation is designed as follows. The range of spatial index $\delta$ is $[-2, 2]$. Since $\delta$ is the integer, the possible values of $\delta$ each time sampled randomly are $-2, -1, 0, 1, 2$. Thus, our pretext task is a 5-class classification problem. The number of time that we random replace patches repeatedly is $T = 50$ and each patch is $20 \times 20$ pixels. We trained Resnet34 as the pretext model and achieved 97% accuracy on predict value of spatial index $\delta$. We analyze the Gradient Class Activation Map (Grad-CAM) to visualize how model looks at the corrupted input image as the figure below. It is clear that a pretext model performs well on detecting normal/abnormal and how it looks like in the next or previous slice.

4) MAIN TASK

We treat the problem as seven-class segmentation including six OARs and a background. One pixel can be only assigned to only one class. Therefore, Softmax function is applied to only one class. Therefore, Softmax function is applied to each time sampled $\delta = 0.1, 0.2, 0.2, 0.3, 0.4, 0.4$ and $0.4$ for background, left lung, right lung, spinal cord, esophagus, heart and trachea respectively.

5) EVALUATION

To evaluate models, dice score is used to measure overlapped volume between predicted segmentation $S$ and ground-truth $G$.

\[
DSC(S, G) = \frac{2 \times |S \cap G|}{|S| + |G|}
\]

6) RESULTS

We run the main task with 3 different scenarios: 50%, 100% and 100% + Combined (CT-MR) Healthy Abdominal Organ Segmentation (CHAOS) dataset to demonstrate how the amount of training data of pretext task has effects on the performance of the main task. Firstly, we take 50% data of StructSeg to construct dataset for the pretext task. The pretrained weights of pretext models are used as the initialize weights of encoder in the main task. We compare our method with the recent works: Jigsaw [16] and Context Restoration [39]. Those two methods only focus on the semantic structure while our method exploits the 3D spatial information. It is clear there is a significant improvement in our method in comparison with others. The 3D spatial gets 91.02% dice score while Jigsaw and Context Restoration get 88.68% and 89.60% correspondingly.

In the second scenario, we use all 100% data of training set for training a pretext task. In the main task, we also compare to two addition methods: random initialization weights and using ImageNet pretrained weights. From the table, obviously, training from scratch has lowest performance of 83.74%. Using ImageNet pretrained weights brings a better dice score of 89.62%. All of self-supervised learning methods achieve higher results since the data for pretext task is increased. However, the margins between methods are still unchanged. Our method is still better than the rest. The performance of our method is 91.24% while others are 89.18% and 89.62% for Jigsaw and Context Restoration respectively. However, there are only 40 patients in total. It means there are only 20 patients originally that is relatively small.

---

**TABLE 1.** The performance of different methods on Organ-at-risk segmentation. The self-supervised learning methods are trained on different training ratios.

| Training | Method                  | DSC(%) |
|----------|-------------------------|--------|
| 50%      | Jigsaw [16]             | 86.68  |
| 50%      | Context restoration     | 89.60  |
| 50%      | Spatial awareness (Our) | 91.02  |
| 100%     | Random initialization   | 83.74  |
| 100%     | Jigsaw [16]             | 89.18  |
| 100%     | Pretrained ImageNet     | 89.62  |
| 100%     | Context restoration [39]| 90.27  |
| 100%     | Spatial awareness (Our) | 91.24  |
| 100% + CHAOS | Spatial awareness (Our) | 91.68  |
Therefore, even the 100% of 40 patients may not be enough to expect significant difference in the performance and the final improvement is not much. To show the effectiveness of using more number of unlabeled data, we experimented the third scenario by adding external dataset called CHAOS. Now, we have 120 patients for training pretext model. The performance has been increased from 91.24% to 91.68%. It proved that, the more unlabeled dataset we have, the performance would be better.

For organ-at-risk segmentation dataset, performance is almost stable and saturated. It is a big obstacle to improve the performance of even 0.5%. However, by using more the unlabeled dataset, our proposed method gains a significant improvement by roundly 1.5% and showing an impressive results. The performance is meaningful and significant. Even though the performance increase is not drastic, the improvement is meaningful and in some sense it can be stated as significant.

**B. INTRACRANIAL HEMORRHAGE DETECTION**

1) DATASET OVERVIEW

RSNA Intracranial Hemorrhage is a CT scan dataset used to detect bleeding that occurs inside the cranium. The data includes 17079 patients. During the treatment and diagnostic, the patient might have been taken more than one time of doing CT scan. Thus, there are 19530 case studies in the dataset. The number of slices is variant between 20 to 60 per case. In total, we have 674258 slices. In this dataset, we aim to detect accurate intracranial hemorrhages and its sub-types. There are five types of intracranial hemorrhages corresponding to five labels: epidural, intraparenchymal, intraventricular, subarachnoid and subdural. Figure 6 shows examples of five sub-types. In an image, there might be more than one type existing. In addition, there is a kind of labels called: any which indicates that whether a hemorrhage, regardless of sub-types, appears in the image. In conclusion, it is the multi-label problem with there are up to 6 labels probably included in the image.

2) DATASET PREPROCESSING

In the radiologist’s workflow on determining abnormalities of CT brain images, there is an important setting that a radiologist must know: window. A window is an instruction to the computer to highlight only voxels which fulfils a specific value. The different values of the window will lead to different visualization and display of a CT image. For the different of sub-types, the window values are not the same. Thus, choosing a window carefully is an important step when working with CT images. More specifically, there are five windows that a radiologist should use for each scan: Brain window (W: 80, L: 40), Blood/subdural window (W:130-300, L:50-100), Soft tissue window (W:350–400, L:20–60), Bone window (W:2800, L:600) and Grey-white differentiation window (W:8, L:32 or W:40, L:40) where L is window level or center. W is window width or range. For example, with brain window (W: 80, L: 40), the range of pixels will be: Lower limit = 40 − 80/2, and upper limit = 40 + 80/2. Any voxel values outside this range will be completely or white.
To deal with multiple-window dependencies, we construct RGB images and each channel is a processed image with a window. We select three windows for the construction: Brain window (W:80, L:40), Bone window (W:2800, L:600) and Subdual window (W:215, L:75).

3) PRETEXT TASK
Similar to section IV-A3, we do pretext task setup and training is the same as Organ-at-risk segmentation task. The pretext model is ResNet34.

4) MAIN TASK
In this experiment, we take Resnet34 as the model for intracranial hemorrhages classification. We take pretrained weights from the pretext task for transfer learning. The final fully connected layer will be replaced by the new one that is corresponding to 6 subtypes of intracranial hemorrhages. To deal with multi-label classification problem, the sigmoid function will be used instead of softmax. The ground truth should be a one-hot vector where $i$ indicates the existence of a subtype and $s$ indicates no-subtype of intracranial hemorrhage. The model is trained in 15 epochs. To maintain the learning rate, we use ReduceLROnPlateau scheduler with patience is equal to 0 and $\gamma = 0.1$. It means whenever there is not improvement of validation loss, the learning rate is reduced 10 times. The heavy augmentations are applied, including: Random horizontal flip, elastic transform, grid distortion, optical distortion, shift, scale and rotation.

5) EVALUATION
For the detection problems, average precision (AP) or mean average precision (mAP) is used. But, Intracranial Hemorrhage Detection is a classification problem. This dataset is provided by Radiological Society of North America (RSNA) for Hemorrhage Detection Challenge hosted in the Kaggle platform. In this competition, the goal is to predict the type of intracranial hemorrhages, which is a classification problem. Besides, the challenge asks for the loss values for the performance measure. Therefore, in this paper, we follow the measure to compare with other results. Thus, the weighted multi-label logarithmic loss (log-loss) is the main evaluation metric, in which the lower the better.

At first, the log loss is taken for each subtype $s$ given and image $i$

$$l_{i,s} = l_{i,s} \cdot \ln(y_{i,s}) + (1 - l_{i,s}) \cdot \ln(1 - y_{i,s})$$

(6)

The weighted loss is computed as

$$l_i = \sum_s w_s \cdot l_{i,s}$$

(7)

where $w_s$ is the weight of subtype $s$ where the any-subtype has weight equals to 2 and rest of subtypes have same weight as 1.

6) RESULT
Table 2 shows the results of performance of methods under different configurations. We can see that there is a correlation of performance when increasing the number of across training methods. When using only half of the data, the spatial awareness method performs better than random initialization, jigsaw and context restoration. Interestingly, when we use full data, the loss decreases a double from 0.128 to 0.063 while other methods are not much improved.

In this experiment, the number of data is much larger than in the OAR segmentation above, thus, the efficiency of self-supervised method is clear to see. In addition, using pretrained imagenet gets 0.094 log loss on full data, which is lower than that of our spatial awareness method. In this competition, the leaderboard showed that, there many overfitting methods and drop performance dramatically in the final test set. Thus, achieving better performance in this dataset is a challenging. Our method demonstrate the stable and generalization when testing in the final test set. This suggests that proposed SSL be more useful for the image classification in this cases.

For Intracranial Hemorrhage Detection, the evaluation metric log-loss is very sensitive with the distribution of test set. The leaderboard of the competition showed that there are many methods perform poor on the final test set because of overfitting. Thus, achieving better performance in this dataset is a very challenging task. Our method demonstrate the stable performance and generalization despite of the sensitivity of the measure when testing in the final test set. In this condition, the performance improvement achieved in the experiment is meaningful.

V. CONCLUSION
In this paper, we presented a self-supervised learning method based on spatial awareness for medical image processing. This method allows CNN models to learn not only useful context features but also spatial features from a sequence of images such as: CT or MR images without annotations. Specifically, in pretext task, spatial index has been introduced for learning spatial structures in medical images. The experiment shows that the method has been successful in improving the performance significantly and through the examples of organ segmentation and disease classification. Therefore, proposed method is effective when the dataset contains spatial

| Training | Method | LogLoss |
|----------|--------|---------|
| 50%      | Random initialization | 0.262   |
| 50%      | Jigsaw [16]         | 0.193   |
| 50%      | Context restoration [39] | 0.181   |
| 50%      | Spatial awareness (Ours) | 0.128   |
| 100%     | Random initialization | 0.172   |
| 100%     | Jigsaw [16]         | 0.134   |
| 100%     | Pretrained Imagenet  | 0.094   |
| 100%     | Context restoration [39] | 0.089   |
| 100%     | Spatial awareness (Ours) | 0.063   |
information as in CT or MR images and is flexible to be applied to various problems of main task.

DECLARATION OF COMPETING INTEREST

We confirm that all authors of this manuscript have no conflicts of interests to declare.

REFERENCES

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.

[4] R. Girshick, “Fast R-CNN.,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[7] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2881–2890.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.

[12] 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.

[13] A. Kaznetsova, H. Rom, N. Allidrin, J. Uijlings, I. Krasin, J. Pot-tuset, S. Kamali, S. Popov, M. Allocci, A. Kolesnikov, T. Duerig, and V. Ferrari, “The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale,” 2018, arXiv:1811.00982. [Online]. Available: http://arxiv.org/abs/1811.00982

[14] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 649–666.

[15] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2536–2544.

[16] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 613–621.

[17] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, “MCoGAN: Decomposing motion and content for video generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1526–1535.

[18] N. Srivastava, E. Mansimov, and R. Salakhudinov, “Unsupervised learning of video representations using LSTMs,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 843–852.

[19] M. Norouzi, A. Vinjimoore, P. Favaro, and H. Pirsiavash, “Boosting self-supervised learning via knowledge transfer,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Jun. 2018, pp. 9359–9367.

[20] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 132–149.

[21] D. Li, W.-C. Hung, J.-B. Huang, S. Wang, N. Ahuja, and M.-H. Yang, “Unsupervised visual representation learning by graph-based consistent constraints,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 678–694.

[22] U. Ahsan, R. Madhok, and I. Essa, “Video jigaw: Unsupervised learning of spatiotemporal context for video action recognition,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2019, pp. 179–189.

[23] C. Wei, L. Xie, X. Ren, Y. Xia, C. Su, J. Liu, Q. Tian, and A. L. Yuille, “Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1910–1919.

[24] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, “Learning image representations by completing damaged jigsaw puzzles,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2018, pp. 793–802.

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

[26] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” 2018, arXiv:1803.07728. [Online]. Available: http://arxiv.org/abs/1803.07728

[27] I. Misra, C. L. Zitnick, and M. Hebert, “Shuffle and learn: Unsupervised learning using temporal order verification,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 527–544.

[28] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang, “Unsupervised representation learning by sorting sequences,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 667–676.

[29] Z. Ren and Y. J. Lee, “Cross-domain self-supervised multi-task feature learning using synthetic imagery,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 762–771.

[30] H. Jiang, G. Larsson, M. M. G. Shakhnarovich, and E. Learned-Miller, “Self-supervised relative depth learning for urban scene understanding,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 19–35.

[31] B. Korbar, D. Tran, and L. Torresani, “Cooperative learning of audio and video models from self-supervised synchronization,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 7763–7774.

[32] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in Proc. Int. Conf. Mach. Learn. (ICML), 2014, arXiv:1412.6980.

[33] N. Sayed, B. Brattoli, and B. Ommer, “Cross and learn: Cross-modal self-supervision,” in Proc. German Conf. Pattern Recognit. Cham, Switzerland: Springer, 2018, pp. 228–243.

[34] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujwara, and D. Rueckert, “Self-supervised learning for medical image analysis using image context restoration,” Med. Image Anal., vol. 58, Dec. 2019, art. no. 101539.

[35] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[36] L. N. Smith and N. Topin, “Super-convergence: Very fast training of neural networks using large learning rates,” Proc. SPIE, vol. 11006, May 2019, Art. no. 1100612.
XUAN-BAC NGUYEN received the B.S. degree in electronic and telecommunication from the University of Engineering and Technology, Vietnam National University, Vietnam, in 2015. He was a Software Engineer in Japan for a period of two years. Since then, he has been with the Department of Electronics and Computer Engineering, Chonnam National University. His main research interests include emotion recognition, image processing, and deep learning. He is an Aggressive Competitor with the Kaggle which is the largest community of data science, machine learning, and deep learning. He received several competition medals and he horned to receive Master of Competition tier with the ranking as top-50 worldwide.

GUEE SANG LEE (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in computer engineering from Seoul National University, South Korea, in 1980 and 1982, respectively, and the Ph.D. degree in computer science from Pennsylvania State University, in 1991. He is currently a Professor with the Department of Electronics and Computer Engineering, Chonnam National University, South Korea. His research interests include image processing, computer vision, and video technology.

SOO HYUNG KIM (Member, IEEE) received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. Since 1997, he has been a Professor with the School of Electronics and Computer Engineering, Chonnam National University, South Korea. His research interests include pattern recognition, document image processing, medical image processing, and ubiquitous computing.

HYUNG JEONG YANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Chonbuk National University, South Korea. She is currently a Professor with the Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea. Her main research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.

* * *