Self-Supervised Pretraining with DICOM metadata in Ultrasound Imaging

Szu-Yeu Hu SDCJIMMY@GMAIL.COM
Center for Ultrasound Research & Translation
Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

Shuhang Wang SWANG38@MGH.HARVARD.EDU
Center for Ultrasound Research & Translation
Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

Wei-Hung Weng CKBJIMMY@MIT.EDU
Computer Science & Artificial Intelligence Laboratory
Massachusetts Institute of Technology, Cambridge, MA, USA

JingChao Wang JWANG116@MGH.HARVARD.EDU
Center for Ultrasound Research & Translation
Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

XiaoHong Wang XWANG91@MGH.HARVARD.EDU
Center for Ultrasound Research & Translation
Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

Arinc Ozturk AOZTURK@MGH.HARVARD.EDU
Center for Ultrasound Research & Translation
Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

Qian Li LI.QIAN@MGH.HARVARD.EDU
Center for Ultrasound Research & Translation
Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

Viksit Kumar VKUMAR14@MGH.HARVARD.EDU
Center for Ultrasound Research & Translation
Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

Anthony E. Samir ASAMIR@MGH.HARVARD.EDU
Center for Ultrasound Research & Translation
Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

Abstract

Modern deep learning algorithms geared towards clinical adaption usually rely on a large amount of high fidelity labeled data. Low-resource settings pose challenges like acquiring high fidelity data and becomes the bottleneck for developing artificial intelligence applications. Ultrasound images, stored in Digital Imaging and Communication in Medicine (DICOM) format, have additional metadata data corresponding to ultrasound image parameters and medical exams. In this work, we leverage DICOM metadata from ultrasound images to help learn representations of the ultrasound image. We demonstrate that the proposed method outperforms the approaches without using metadata across a variety of downstream tasks.

1. Introduction

In recent years, deep learning algorithms have made foray into the clinical domain and has emerged as a successful technique in various medical imaging applications. It has shown the potential to automate disease detection, severity grading, and clinical diagnosis in different domain (Hu et al., 2019; Gulshan et al., 2016; Esteva et al., 2017). However, clinically accepted deep learning algorithms require a considerable amount of annotated data. For example, Gulshan et al. (2016) utilizes more than 100,000 images to train and validate the algorithm. Unfortunately, obtaining accurate annotations from clinicians is extremely expensive, constraining supervised learning approaches in low-resource settings.

Unsupervised or semi-supervised learning provides potential solutions to alleviate the problems by learning the data distribution without or with limited labels. Studies have shown that unsupervised pretraining can serve as a regularization method and lead to better generalization (Erhan et al., 2010). Recently, weakly-supervised and self-supervised learning have also drawn significant attention with their ability to learn high-quality feature representations. In this paper, we will explore one of the self-supervised technique, the context encoder (Pathak et al., 2016), and use the metadata in medical imaging as the weak labels to reinforce its capability to learn representation features.

In most of the modern medical imaging acquisition devices, such as ultrasound imaging, the data is stored in DICOM (Digital Imaging and Communications in Medicine) format. Besides the image pixel data, the DICOM headers contain the metadata, such as the patient information, study descriptions, and the reported results. The abundant information encoded in DICOM format provides a unique opportunity for modern deep learning applications. Recent studies have shown that the metadata can be leveraged for series categorization using machine learning (Gauriau et al.). Nevertheless, DICOM has not been a popular supervision target in machine learning. One major concern about DICOM is that they are often noisy and may contain wrong tags (Gueld et al., 2002). In practice, clinical personnel often adjust the examination protocol and imaging presets to improve the image quality, but these changes may not be properly reflected in the DICOM tags. However, using DICOM metadata as weak labels can help incorporation of valuable information into the deep learning algorithm while minimizing the noise.

In this work, we investigated weakly-supervised learning using metadata and proposed a framework build on top of the self-supervised learning method. We showed that incorporating DICOM metadata as weak labels can improve the quality of representation learning and improve the performance of the downstream segmentation and classification tasks.

2. Related Work

2.1. Pretraining Techniques

It is usually beneficial to train a model from pretrained weights, rather than from random initialization, especially in medical imaging field, where the labels are expensive to obtain. (Erhan et al., 2010) There are multiple ways for pretraining. The first is transfer learning, which first trains the model on a large amount of labeled data, and then tune the pretrained weights for new target tasks. ImageNet-pretrained convolutional neural networks, which is arguably the most successful transfer learning model, have boosted the growth of the modern deep learning applications (Deng et al., 2009). Even in medical imaging, the standard approach is to take an existing architecture trained on ImageNet and then fine-tune on the domain-specific data such as X-ray (Rajpurkar et al., 2017) or retinal fundus photography (Abràmoff et al., 2016). However, given the substantial difference between the natural images and the medical imaging, recent studies raised the questions of the precise effects of the pretrained features and suggested that transfer learning does not always improve the final performance (Raghu et al., 2019; Kornblith et al., 2019; He et al., 2019).

While transfer learning relies on supervision from large-scale hand-labeled databases without employing the rich information presented in the image structure, unsupervised learning, another popular approach for pretraining, tries to build useful feature representation using the data itself (Bengio, 2012). For example, Hinton et al. (2006) presented a greedy layerwise unsupervised pretraining methods to build representations of different levels. Variants of the autoencoder (Baldi, 2012), such as stacked denoising autoencoder (Vincent et al., 2010) or contractive autoencoder (Rifai et al., 2011), build the encoder in the process of reconstructing the original image. In recent years, the generative adversarial network also emerged as a powerful framework representation learning (Donahue et al., 2016; Donahue and Simonyan, 2019). These methods trained a network without labels, and the learned weights can be used either as high-level image feature inputs or as initialization for a target downstream task.

2.2. Self-supervised learning

Self-supervised learning is a unique form of supervised learning which eliminates the demand for manual labels. The key idea is to generate labels from the data itself and trains the network in a supervised manner. Such methods, also known as pretext tasks, have proved to be an effective technique for representation learning, and have been widely used in natural language processing. For example, BERT(Devlin et al., 2018), one of the recent breakthroughs in language model pretraining, was trained to predict the masked words given the input sequences. In image-based tasks, many methods have also been proposed. Gidaris et al. (2018) randomly rotates the images while maintaining the semantic content unchanged, and the network was trained to predict the rotation angles. Noroozi and Favaro (2016) formulated the pretext task as a jigsaw puzzle and pretrain the model by solving it. Contrastive predictive coding (Oord et al., 2018) learns an encoder to encode image patches and utilized an autoregressive decoder to predict the future vectors with a contrastive loss. Chen et al. (2020) further improved the training techniques of contrastive learning, and had achieved the performance close to the supervised pretraining.

In this study, we employed the context encoder(Pathak et al., 2016) as the foundation of our proposed framework, in which the network is trained to predict the missing parts of the images. We leave the detail descriptions of the context encoder in section 3.1.

2.3. Weakly-supervised learning

Weakly-supervised learning is another subclass of supervised learning, in which the labels can be either inexact or inaccurate. Inexact supervision usually involves annotations at a higher abstraction level. For example, Wang et al. (2017) and Yan et al. (2018) localize the disease position in Chest X-ray with image-level classes; Hu et al. (2018) shows that the model trained on the position coordinates can improve the segmentation task.

Inaccurate supervision uses a large quantity of low quality or noisy labels. One remarkable illustration is the work in Mahajan et al. (2018), which took advantage of billions of Instagram hashtag for weakly-supervised pretraining to boost the ImageNet classification. Recently, Xie et al. (2019) leverage noisy labels from a teacher-student framework and achieved the state-of-the-art classification accuracy on ImageNet.

Inspired by these work, we propose to incorporate DICOM metadata, which has noisy labels embedded within the medical imaging raw data, for weakly-supervised pretraining.

2.4. Adversarial Training

When training the context encoder, an adversarial loss was added to encourage realistic output. Adversarial training originates from the generative adversarial network (GAN)(Radford et al., 2015), which utilizes a discriminator network to distinguish the generative image and the real input. Beyond the great success in image generation, it also shows a substantial impact on other areas like domain adaption(Ganin et al., 2016) or adversarial attack(Tramèr et al., 2018). In typical tasks such as semantic segmentation, adversarial training can also boost the performance under semi-supervised(Hung et al., 2018) or unsupervised settings(Chen et al., 2019a).

A standard adversarial network does not require supervision, but recent studies have shown that the class labels can stabilize the training and improve image qualities. For example, Brock et al. (2018) fed the labels as the generator inputs to produce high-quality images. AC-GAN (Odena et al., 2017) used the discriminator to classify the class labels as an auxiliary loss. Miyato and Koyama (2018) proposed a linear projection layer, which was also employed in Lučić et al. (2019) to generate high fidelity images with a limited number of labels.

3. Method

In this study, we proposed a new self-supervised representation learning framework, which incorporates the DICOM metadata as weak labels to improve the training. In particular, we employed the context encoder as the self-supervised pretext task. The overview of the framework is demonstrated in Figure 1.

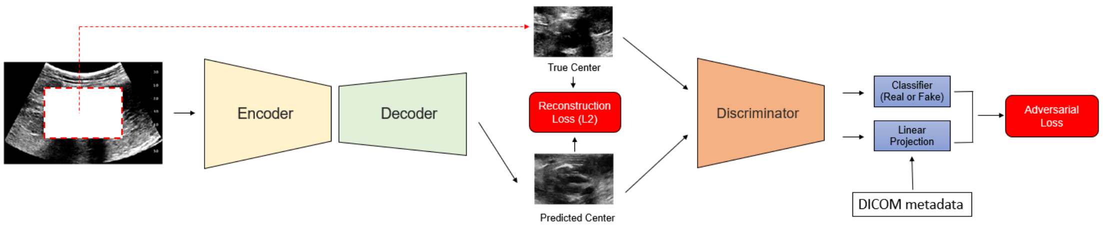


Figure 1: The proposed frame work for representation learning

3.1. Context Encoder

The idea of the context encoder is that given an input image with intentionally masked out areas, we train a deep learning model to reconstruct the missing part (semantic in-painting). The network utilizes an encoder-decoder structure. The encoder encodes the image context into a compact latent representation, and the decoder employs them to generate the missing image content. The network is trained to minimize the mean square reconstruction loss.

In the original context encoder paper, it is proposed that the in-painting area can be either fixed or random blocks. Typically, models using random blocks tend to generalize better. However, due to the nature of ultrasound images, where informative context is located in the central region, we crop a square patch in the center of the image with a fixed size equal to half of the image width and height.

3.2. Discriminator with Linear Projection Layer

We also added the discriminator for adversarial training to encourage realistic output. The standard $\mathcal{L}_{adv}$ is formulated as

$$\mathcal{L}_{adv} = \max_{D} \mathbb{E}_{x \in \mathcal{X}}[\log(D(x)) + \log(1 - D(F(\hat{x})))] \tag{1}$$

where F is the context encoder, D is the discriminator, x is original image, and $\hat{x}$ is cropped input image. (Noted that F is often denoted as G in most of the GAN literature; here we used F to distinguish the context encoder and a regular generator.)

To incorporate the DICOM metadata, we employ a linear projection layer as proposed in Miyato and Koyama (2018) and Lučić et al. (2019). The discriminator was decomposed into a learned discriminator representation, $\tilde{D}(x)$, and the representation then fed into two different parts: (1) A classifier C_{rf} to distinguish whether the image is real or fake; (2) A linear project layer P, with a learned weight matrix W applied to a feature vector $\tilde{D}(x)$ and the encoded DICOM tags y as an input. The output of the discriminator becomes:

$$D(x, y) = C_{rf}(\tilde{D}(x)) + P(\tilde{D}(x), y)$$

, where $P(\tilde{D}(x), y) = \tilde{D}(x)^{\top} W y$. Also, we adopoted a hinge version of the adversarial loss. With the above modification, the loss function for context encoder F and the discriminator D can be rewritten as:

$$\mathcal{L}_D = -\mathbb{E}_{(x,y)\sim p(x,y)}[\min(0, -1 + D(x, y)))] - \mathbb{E}_{(\hat{x},y)\sim p(x,y)}[\min(0, -1 - D(F(\hat{x}), y))] \tag{2}$$

$$\mathcal{L}_F = -\lambda_{adv} \times \mathbb{E}_{(\hat{x},y)\sim p(x,y)}[D(F(\hat{x}), y)] + \lambda_{rec} \times \mathbb{E}_{(x,\hat{x})\sim p(x)}[(F(\hat{x}) - x)^2] \tag{3}$$

The second term of Equation(3) is the reconstruction loss(mean square error). We included two hyperparameters λ_{adv} and λ_{rec} to balance the two different losses.

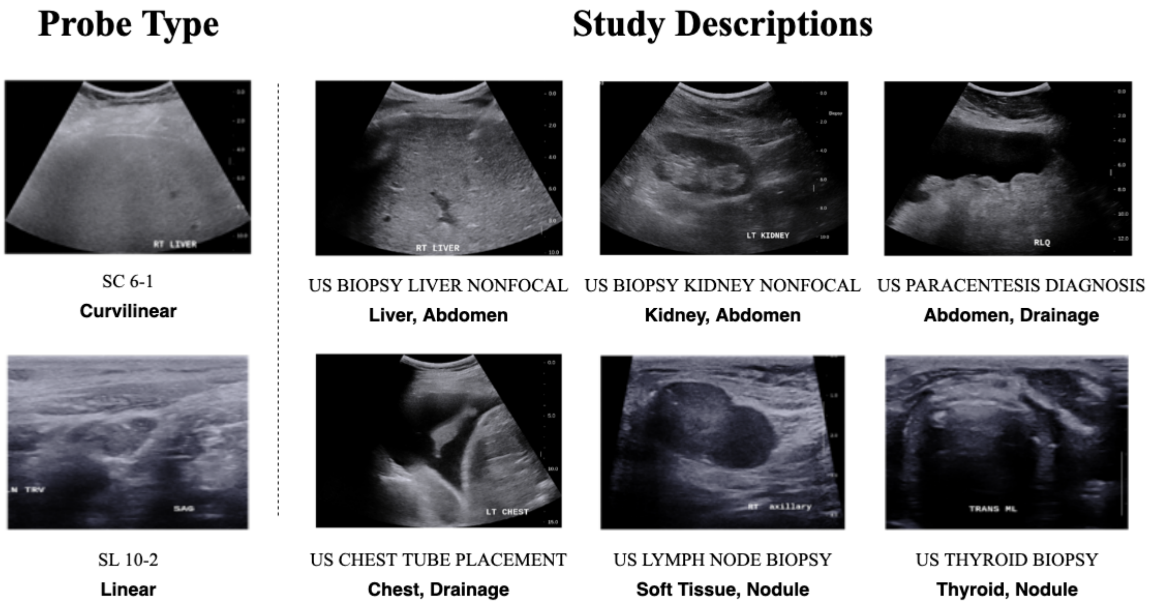
3.3. DICOM MetaData

We select two DICOM tags as the target since they directly relate with the image semantic context:

- **Transducer data** (DICOM tag: (0018, 5010)), which indicates the probe type used for examination. There are three different transducer probes in the dataset – SC6-1, SL10-2, SL15-4, where S represents single crystal, C or L represents curvilinear or linear probe geometry, and the numbers represent the ultrasound frequency bandwidth in MHz. We classify the probes into two groups - linear (SL10-2, SL15-4) and curvilinear (SC6-1).

- **Study Description** (DICOM tag: (0008, 1030)). The study description illustrates the protocol when performing the ultrasound exam. For example, images of "US BIOPSY LIVER NONFOCAL" are acquired during an ultrasound-guided liver biopsy. Therefore, we can expect these images are predominantly liver. We identified 45 different study description in our dataset (Appendix Table A1). Due to the spurious nature of the tags, we categorized the study series into eight different groups according to procedure type or site, including liver, kidney, thyroid, abdomen, chest, soft tissue, nodule,and drainage. The DICOM categorization was performed manually by a board-certified radiologist. Each study series can belong to more than one group. For example, the tag "US BIOPSY LIVER NONFOCAL" is mapped to two groups – liver and abdomen. We binarized the DICOM labels in a multi-label format.

Figure 2 demonstrates some image examples of the DICOM tags.



SC 6-1
Curvilinear

SL 10-2
Linear

US BIOPSY LIVER NONFOCAL
Liver, Abdomen

US BIOPSY KIDNEY NONFOCAL
Kidney, Abdomen

US PARACENTESIS DIAGNOSIS
Abdomen, Drainage

US CHEST TUBE PLACEMENT
Chest, Drainage

US LYMPH NODE BIOPSY
Soft Tissue, Nodule

US THYROID BIOPSY
Thyroid, Nodule

Figure 2: Example of DICOM metadata

4. Experiments

4.1. Dataset

Table 1: Description of the dataset

Dataset	Task	# of images	# Train	# Val	# Test
Private	Semantic Inpainting	12267	9814	2453	0
Private	Quality Classification	3226	2548	343	335
Private	Liver/Kindey Segmentation	591	391	100	100
Public	Thyroid Nodule Segmentation	466	298	74	94

A retrospective database was collected from September 2018 to November 2019 after proper approval from the Institutional Review Board. Informed consent was waived, and HIPAA compliance was ensured. A total of 12,267 images from 1,188 unique patients were collected. All images were acquired using Supersonic Aixplorer ultrasound machine (SuperSonic Imagine S.A., Aix-en-Provence, France).

We evaluated the results on three different downstream tasks: (1) Quality score classification on a private dataset. (2) Liver and kindey segmentation on private dataset. (3) Thyroid nodule segmentation on open dataset. The two private datasets were retrospectively collected from the same institution. All the images were acquired using a GE Logiq E9 ultrasound machine(GE Healthcare, Chicago, IL, USA). There is no overlap between our pretraining dataset and the downstream evaluation dataset. The description for the open dataset can be found in Pedraza et al. (2015). The overview of the dataset is shown in Table 1.

4.2. Context Encoder Pretraining

4.2.1. Architecture

The proposed framework consists of three parts - the encoder, the decoder and the discriminator. For the encoder, we employed two different existing network architecture - VGG16 with batch normalization (Simonyan and Zisserman, 2014) and Resnet-50 (He et al., 2016) as the backbone. The decoder has four up-sampling blocks each with a 3×3 up-convolutional, a batch normalization, and a ReLU layer. The discriminator also has four blocks, each with a 3×3 convolutional, a batch normalization, and a LeakyReLU layer as suggested by Radford et al. (2015).

4.2.2. Training

The dataset was split into training (80%, 9814 images) and validation set (20%, 2453 images) randomly while ensuring that all images from the same patient were within one set. All images were resized to an input size of 256×384 pixels and Z-score normalized before feeding into the network. Data augmentation was performed using random flipping, vertical and horizontal shifting.

We update the context encoder and discriminator parameters using Adam optimizer, minimizing $\mathcal{L}_F$ and $\mathcal{L}_D$ alternatively, with hyper-parameters set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch size = 32, context encoder learning rate = 0.0001, and discriminator learning rate = 0.00001.The models were trained over 200 epochs without early stopping, and the ones with the lowest $\mathcal{L}_F$ on the validation set were selected for downstream evaluation. We didn't split a held-out test set since we presented our quantitative results on the downstream tasks.

4.3. Experiment Settings

For both encoder, VGG16 and ResNet50, we compare the following pretraining configurations:

- **Baseline**: The encoder was randomly initialized without pretraining.
- **ImageNet**: The encoder was trained on ImageNet classfication.
- **DCM**: The encoder was trained to predict the DICOM metadata directly.
- **CE**: The encoder was trained with the context encoder without the DICOM metadata.
- **CE + DCM**: The encoder was trained using our proposed framework with DICOM metadata.
- **CE + DCM + F**: The encoder was trained using our proposed framework with DICOM metadata, and the parameters were frozen while training the downstream tasks.

4.4. Downstream Tasks Evaluation

4.4.1. Architecture

After the model was trained, only the encoder part was fine-tuned for downstream tasks. Downstream classification tasks using a classifier layer, consisting of a 1×1 convolutional, a dropout, a global average pooling, and a fully connected layer, was appended followed by a sigmoid activation function. For downstream segmentation task, we adopted an architecture simmilar to U-Net (Ronneberger et al., 2015), where the encoder arm is modified to be the pretrained VGG16 or ResNet. We followed an implementation similar to Iglovikov and Shvets (2018), adding five up-convolutional blocks and skip connection to complete the network.

4.4.2. Classification

The downstream classification task, quality score classification, is to identify an optimal view for Morrison's pouch - an anatomic site between the right lobe of the liver and the right kidney. Clinically, the view is important to identify ascites and hemoperitoneum when abnormal fluid accumulation is present. Furthermore, it is the reference view to estimate the severity of steatosis using the hepatorenal index. Therefore, quantifying the optimal view is crucial in an ultrasound examination. The images used in the classification task were reviewed by a board-certified radiologist and given five different rankings as the quality score (Figure 3). Class 0 indicates the view does not include the liver or the kidney, and should not be used; class 1 and 2 are the correct Morrison's pouch view, but the anatomic structure is not clear enough for clinical applications; class 3 and 4 are the clinically acceptable views, and class 4 represent the optimal Morrison's pouch view that will be used by an experienced operator. We used the ordinal encoding for the labels. (class 0: [0,0,0,0], class 1: [1,0,0,0], class 2: [1,1,0,0], class 3: [1,1,1,0], class 4: [1,1,1,1])

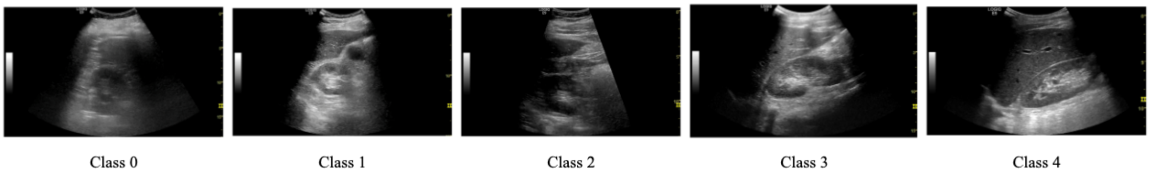


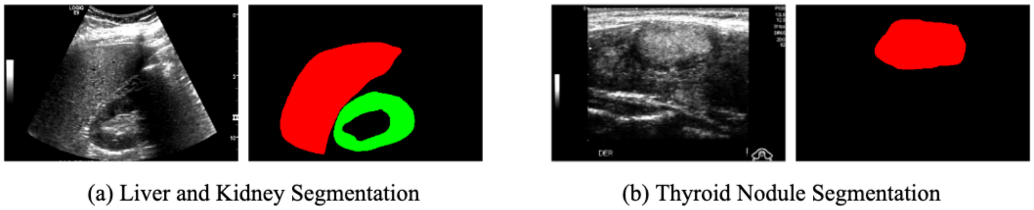
Figure 3: Quality Score classification examples.

4.4.3. Segmentation

We evaluated two different segmentation tasks. The first is to segment the kidney and liver in B-mode ultrasound imaging. This work is related to quality score classification. A board-certified radiologist selected the images representing optimal Morrison's pouch view from the institutional database and manually annotated the kidney and liver anatomy. (Figure 4(a)). The second task is thyroid nodule segmentation, including cystic nodules, adenomas and thyroid cancers, using an open access B-mode thyroid ultrasound image dataset. An example is shown in Figure 4(b).

4.4.4. Training

For all three downstream tasks, the images follow the same pre-processing procedure described in section 4.2.2. The training hyper-parameters are summarized in Appendix Table A2. Models were trained without early stopping, and the epochs with the lowest validation loss were selected. All reported values were evaluated on the held-out test set.



(a) Liver and Kidney Segmentation

(b) Thyroid Nodule Segmentation

Figure 4: Examples of the downstream segmentation tasks.

5. Results

5.1. Context Encoder with DICOM

The qualitative results of the context encoder with and without DICOM tags are shown in Figure 5. We observe that trainings without DICOM tags are more prone to mode collapse in our experiments, making it difficult to obtain optimal results. With DICOM tags, the generated images look sharper and can resemble the actual organ texture simmilar to liver and kidney. Figure 5 qualitatively shows that the joined weakly-supervised training with DICOM tags improved the prediction quality.

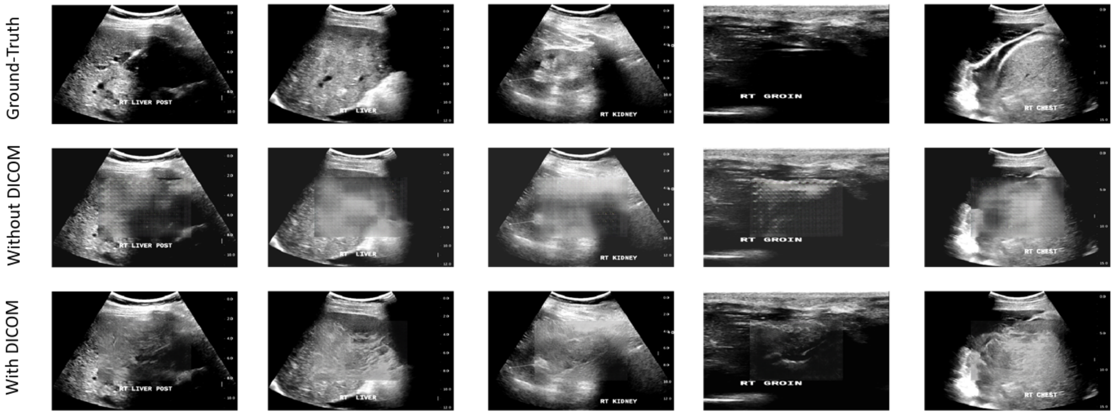


Figure 5: Example of semantic in-painting on various organs using the validation set

5.2. Downstream Tasks

Next, we examined whether downstream segmentation and classification can benefit from the pretrained encoder. The results are summarized in Table 2.

The performance with pretraining improved significantly across different tasks and configurations compared to the random-initialized baseline models. The pretraining from ImageNet and DICOM also work fairly well however, our proposed method, context encoder with DICOM, consistently obtained the best results. The effect of freezing the encoder differs between the two backbone. When freezing the encoder, we are reusing the learned features directly, and the ResNet models benefit more from this approach; when unfreezing the encoders, we treat it as the self-supervised initialization and the VGG16 gains more from this approach. The observation is consistent with the conclusion in (Kolesnikov et al., 2019), i.e. quality of the representation learned in self-supervised tasks deteriorates toward the final layers of the VGG network. In contrast, the skip connections in ResNet architecture help preventing the degradation of the representation and is the best performer when reusing the features up to the pre-logit layers.

To emphasize the impact of adding DICOM metadata, we further repeat the experiments using a smaller data regime. We compared three configurations: CE + DCM (+F), CE, and baseline model, but only using 5% of the data for all three tasks. The results are shown in Figure 6; note that we only freeze the encoder for ResNet backbone given the previous conclusion. The boxplots show that adding the metadata improved the performance in all cases. The difference between CE + DCM (+F) and CE are statistically significant (p-value < 0.05) in all combination, except for the quality score classification using VGG16 as the backbone (p-value $= 0.111$).

Table 2: Performance evaluation of downstream tasks undergoing pretraining. The reported value is mean ± standard deviation on the held-out test set. The standard deviation is derived from bootstrapping for 1000 times on the test set. Each time we sample 50% of the test data with replacement. Best performance is highlighted in boldface. The abbreviation of each pretraining configuration is specified in section 4.3.

Backbone	**Pretraining**	**Downstream Task**		
		Quality	Liver/Kidney	Thyroid
VGG16-BN	Baseline	0.558 ± 0.031	0.801 ± 0.011	0.856 ± 0.012
VGG16-BN	ImageNet	0.656 ± 0.034	0.824 ± 0.008	0.876 ± 0.009
VGG16-BN	DCM	0.652 ± 0.020	0.829 ± 0.008	0.859 ± 0.011
VGG16-BN	CE	0.629 ± 0.022	0.816 ± 0.009	0.858 ± 0.011
VGG16-BN	CE+DCM	**0.657 ± 0.031**	**0.832 ± 0.009**	**0.883 ± 0.010**
VGG16-BN	CE+DCM+F	0.645 ± 0.035	0.826 ± 0.009	0.879 ± 0.011
RESNET50	Baseline	0.706 ± 0.029	0.765 ± 0.013	0.843 ± 0.011
RESNET50	ImageNet	0.708 ± 0.035	0.811 ± 0.014	0.849 ± 0.011
RESNET50	DCM	0.706 ± 0.028	0.753 ± 0.016	0.850 ± 0.011
RESNET50	CE	0.715 ± 0.028	0.781 ± 0.011	0.849 ± 0.010
RESNET50	CE+DCM	0.715 ± 0.024	0.807 ± 0.011	0.852 ± 0.011
RESNET50	CE+DCM+F	**0.754 ± 0.024**	**0.814 ± 0.014**	**0.865 ± 0.010**

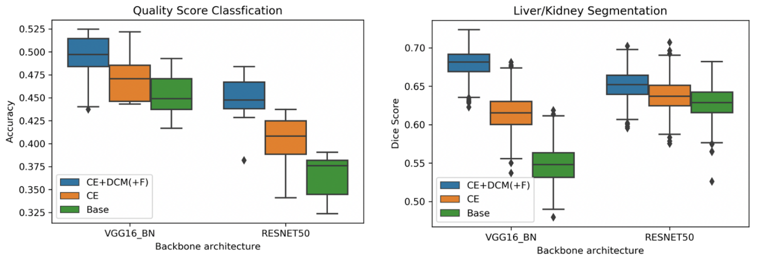


Figure 6: Boxplots between the three configurations **CE+DCM(+F)**, **CE** and **Baseline** across two backbone architectures and three downstream tasks. **CE+DCM(+F)** here denotes **CE+DCM+F** when we use ResNet as backbone and **CE+DCM** when using VGG16.

6. Discussion

In this study, we demonstrate that the performance of existing self-supervision techniques can be consistently boosted with DICOM metadata as weak labels. Comparing to other pretraining data sources like ImageNet, which often comprises millions of entries, our methods achieve comparable performance with only around 10,000 images. Mahajan et al. (2018) suggested that, while increasing the size of the pretraining dataset may be beneficial, selecting a label space for the source task to match that of the target task is even more fruitful. In our experiments, the pretraining and the downstream dataset share a similar distribution; they both cover standard ultrasound examination views such as the abdominal and thyroid scans. Our results further emphasize the benefits of pretraining from data of the same domain. Such a method can be particularly useful in the low-resource settings where obtaining and training large-scale annotated data is not feasible, and that leads to reducing the gap toward building a generalized and robust medical imaging pretraining technique.

The choice of DICOM tags is crucial to the success of the application. We only experimented with ultrasound images and two DICOM tags. Different image modality like CT and MRI have their distinct metadata and would require further investigation to identify the proper candidates. Potential targets such as voxel information(pixel spacing, Hounsfield units), study details(anatomic structure, patient orientation), or patient-level data(demographics, diagnosis) can provide meaningful semantic information for supervised learning. Though, some tags like study descriptions or study findings may be inconsistent among different acquisition devices or institutions, they can still be valuable with proper categorization by the clinical experts.

In our experiments, we focused on the advantages of DICOM metadata and only investigated one self-supervised method. However, the methodology used in this paper, adding the DICOM weak labels to the discriminator, can be generalized to other pretext tasks given that adversarial training is often used as an auxiliary loss to many existing models. For example, Chen et al. (2019b) extended the self-supervised rotation loss with a GAN-based structure. Exploration of different pretext tasks will be necessary for future studies.

7. Conclusion

In this paper, we demonstrate the potential of using DICOM metadata from ultrasound images as weak labels to improve representation learning in a self-supervised schema. The method can have great impacts in resource-limited regions by leveraging its ability to effectively utilize the pre-existing information, curtailing the need of additional annotations which require high skill and are expensive. The method can be extended to other medical image modalities with DICOM tags like CT or MRI.

References

Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science*, 57(13):5200–5206, 2016.

Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *Advances in Neural Information Processing Systems*, pages 12705–12716, 2019a.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019b.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pages 10541–10551, 2019.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Romane Gauriau, Christopher Bridge, Lina Chen, Felipe Kitamura, Neil A Tenenholtz, John E Kirsch, Katherine P Andriole, Mark H Michalski, and Bernardo C Bizzo. Using dicom metadata for radiological image series categorization: a feasibility study on large clinical brain mri datasets. *Journal of Digital Imaging*, pages 1–16.

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

Mark Oliver Gueld, Michael Kohnen, Daniel Keysers, Henning Schubert, Berthold B Wein, Joerg Bredno, and Thomas Martin Lehmann. Quality of dicom header information for image categorization. In *Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation*, volume 4685, pages 280–287. International Society for Optics and Photonics, 2002.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4918–4927, 2019.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Szu-Yeu Hu, Andrew Beers, Ken Chang, Kathi Höbel, J Peter Campbell, Deniz Erdogumus, Stratis Ioannidis, Jennifer Dy, Michael F Chiang, Jayashree Kalpathy-Cramer, et al. Deep feature transfer between localization and segmentation tasks. *arXiv preprint arXiv:1811.02539*, 2018.

Szu-Yeu Hu, Wei-Hung Weng, Shao-Lun Lu, Yueh-Hung Cheng, Furen Xiao, Feng-Ming Hsu, and Jen-Tang Lu. Multimodal volume-aware detection and segmentation for brain metastases radiosurgery. In *Workshop on Artificial Intelligence in Radiation Therapy*, pages 61–69. Springer, 2019.

Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

Mario Lučić, Marvin Ritter, Michael Tschannen, Xiaohua Zhai, Olivier Frederic Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. In *International Conference on Machine Learning*, 2019.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *International Conference on Learning Representations*, 2018.

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

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, volume 9287, page 92870W. International Society for Optics and Photonics, 2015.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, pages 3342–3352, 2019.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. 2011.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 2097–2106, 2017.

Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019.

Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang. Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 103–110, 2018.

Appendix A.

Table A1: The full list of the DICOM Study Descriptions and the corresponding encoding

Study Descriptions	**Encoding**
US BIOPSY LIVER NONFOCAL	liver,abdomen
US BIOPSY LIVER FOCAL	liver,abdomen
US LYMPH NODE BIOPSY	soft tissue,nodule
US BIOPSY KIDNEY NONFOCAL (EITHER SIDE)	kidney,abdomen
US PARACENTESIS THERAPEUTIC	abdomen,drainage
US BIOPSY TRANSPLANTED KIDNEY	kidney,abdomen
US PARACENTESIS DIAGNOSTIC AND THERAPEUTIC	abdomen,drainage
US THYROID BIOPSY	thyroid,nodule
US PARACENTESIS DIAGNOSTIC	abdomen,drainage
US THORACENTESIS DIAGNOSTIC AND THERAPEUTIC	chest,drainage
US THYROID ASPIRATION/FNA	thyroid,nodule
US DRAINAGE INTERVENTION NOT OTHERWISE SPECIFIED	soft tissue,drainage
US DRAINAGE ABDOMEN	abdomen,drainage
US DRAINAGE GALLBLADDER (CHOLECYSTOSTOMY)	abdomen,drainage
US THORACENTESIS THERAPEUTIC (RIGHT)	chest,drainage

Study Descriptions	**Encoding**
US THORACENTESIS THERAPEUTIC (LEFT)	chest,drainage
US BIOPSY MESENTERY	abdomen,drainage,soft tissue
US NECK SOFT TISSUE BIOPSY	soft tissue,nodule
US DRAINAGE CATHETER PLACEMENT	soft tissue,drainage
US DRAINAGE PELVIS	abdomen,drainage
US SOFT TISSUE BIOPSY	soft tissue,nodule
US BIOPSY KIDNEY NONFOCAL (LEFT)	kidney,abdomen
US CHEST TUBE PLACEMENT (RIGHT)	chest,drainage
US BIOPSY NOT OTHERWISE SPECIFIED	soft tissue,nodule,drainage
US ABDOMINAL PELVIC BIOPSY NOT OTHERWISE SPECIFIED	soft tissue,nodule,drainage
US CHEST TUBE PLACEMENT (LEFT)	chest,drainage
CT BIOPSY LIVER FOCAL	liver,abdomen
US BIOPSY KIDNEY FOCAL (LEFT)	liver,abdomen
US ASPIRATION ABDOMINAL COLLECTION	abdomen,drainage
CT LYMPH NODE BIOPSY	soft tissue,nodule
US DRAINAGE LIVER	liver,drainage,abdomen
US BIOPSY RETROPERITONEUM	abdomen
US LYMPH NODE ASPIRATION/FNA	soft tissue,nodule,drainage
US SOFT TISSUE ASPIRATION	soft tissue,drainage
US ASPIRATION PELVIS	abdomen,drainage
US THORACENTESIS DIAGNOSTIC (RIGHT)	chest,drainage
US THORACENTESIS DIAGNOSTIC (LEFT)	chest,drainage
US DRAINAGE KIDNEY/PARARENAL (RIGHT)	abdomen,kidney,drainage
US HEAD/NECK INTERVENTION NOT OTHERWISE SPECIFIED	soft tissue
US BIOPSY KIDNEY FOCAL (RIGHT)	kidney,abdomen
CT ABDOMINAL PELVIC BIOPSY NOT OTHERWISE SPECIFIED	abdomen
US DRAINAGE KIDNEY/PARARENAL (LEFT)	kidney,abdomen
IR PARACENTESIS (THERAPEUTIC)	abdomen,drainage
US PSEUDOANEURYSM THROMBIN INJECTION	soft tissue,nodule

Table A2: Training Details for the Downstream Tasks

	Quality Score	**Liver/Kidney Segmentation**	**Thyroid Segmentation**
Optimizer	Adam (($\beta_1 = 0.9$, $\beta_2 = 0.999$))	Adam (($\beta_1 = 0.9$, $\beta_2 = 0.999$))	Adam (($\beta_1 = 0.9$, $\beta_2 = 0.999$))
Batch Size	4	8	8
Training Epochs	300	500	500
Loss Function	Weighted Binary CrossEntropy	Soft Dice loss	Soft Dice Loss
Learning Rate	VGG16:0.00005 RESNET: 0.00005	VGG16: 0.0001 RESNET:0.00005	VGG16: 0.0001 RESNET:0.0001