# Self-supervised Representation Learning via Image Out-painting for Medical Image Analysis

by

Vatsal Sodha

A Thesis Presented in Partial Fulfillment
of the Requirement for the Degree
Master of Science

Approved July 2020 by the
Graduate Supervisory Committee:

Jianming Liang, Chair
Baoxin Li
Murthy Devarakonda

ARIZONA STATE UNIVERSITY

August 2020

# ABSTRACT

In recent years, Convolutional Neural Networks (CNNs) have been widely used in not only the computer vision community but also within the medical imaging community. Specifically, the use of pre-trained CNNs on large-scale datasets (*e.g.,* ImageNet) via transfer learning for a variety of medical imaging applications, has become the *de facto* standard within both communities. However, to fit the current paradigm, 3D imaging tasks have to be reformulated and solved in 2D, losing rich 3D contextual information. Moreover, pre-trained models on natural images never see any biomedical images and do not have knowledge about anatomical structures present in medical images. To overcome the above limitations, this thesis proposes an image out-painting self-supervised proxy task to develop pre-trained models directly from medical images without utilizing systematic annotations. The idea is to randomly mask an image and train the model to predict the missing region. It is demonstrated that by predicting missing anatomical structures when seeing only parts of the image, the model will learn generic representation yielding better performance on various medical imaging applications via transfer learning.

The extensive experiments demonstrate that the proposed proxy task outperforms training from scratch in six out of seven medical imaging applications covering 2D and 3D classification and segmentation. Moreover, image out-painting proxy task offers competitive performance to state-of-the-art models pre-trained on ImageNet and other self-supervised baselines such as in-painting. Owing to its outstanding performance, out-painting is utilized as one of the self-supervised proxy tasks to provide generic 3D pre-trained models for medical image analysis.

*To my parents, Arvind and Bharti Sodha*

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Dr. Jianming Liang for his immense support throughout my thesis. I especially thank him for providing solutions whenever I was stuck in my research. Further, I thank Dr. Baoxin Li and Dr. Murthy Devrakonda for accepting to be part of my thesis committee and reviewing my work towards betterment.

This work has utilized the GPUs provided partially by the ASU Research Computing and partially by the Extreme Science and Engineering Discovery Environment (XSEDE) funded by the National Science Foundation (NSF) under grant number ACI-1548562.

Further, I would like to thank my lab-mates: Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Dr. Nima Tajbakhsh, Dr. Ruibin Feng, Mohammad Reza Hosseinzadeh Taher, Shivam Bajpai, Diksha Goyal, and Fatemeh Haghighi for providing suggestions and useful discussions.

Finally, I would like to thank my parents, my sisters, and my family for the immense support and continuous encouragement throughout my masters. This accomplishment could not have been possible without them.

TABLE OF CONTENTS

Page

LIST OF TABLES ..... vi

LIST OF FIGURES ..... vii

CHAPTER

1 INTRODUCTION ..... 1

1.1 Background ..... 1

1.2 Limitations of the Current Paradigm in Medical Image Analysis ..... 2

1.3 Research Question ..... 3

1.4 Self-supervised Representation Learning ..... 3

1.5 Hypothesis and Proposed Solution ..... 5

1.6 Terminology ..... 6

2 RELATED WORK ..... 7

2.1 Supervised Representation Learning ..... 7

2.2 Unsupervised Representation Learning ..... 7

2.3 Self-supervised Representation Learning ..... 8

3 METHODOLOGY ..... 10

3.1 Image Out-painting as a Self-supervised Proxy Task ..... 10

3.2 Region Masks ..... 12

3.3 Limitations of Out-painting ..... 13

4 EXPERIMENTS AND RESULTS ..... 14

4.1 Implementation Details ..... 14

4.1.1 2D Out-painting ..... 14

4.1.2 3D Out-painting ..... 14

4.1.3 Target Tasks ..... 15

4.1.4 Hyper-parameters ..... 17

CHAPTER Page

4.2 Qualitative Results of Image Out-painting . . . . . 17
4.3 Ablation Study . . . . . 18
4.4 2D Results . . . . . 19
4.4.1 Out-painting vs. Training from Scratch . . . . . 20
4.4.2 Out-painting vs. Autoencoder . . . . . 21
4.4.3 Out-painting vs. Self-supervised Baselines . . . . . 21
4.4.4 Out-painting vs. ImageNet . . . . . 22
4.5 3D Results . . . . . 22
4.6 Application of Out-painting . . . . . 24
5 CONCLUSION . . . . . 25
REFERENCES . . . . . 27

LIST OF TABLES

Table Page

4.1 Summary of the Target Tasks. . . . . 15

4.2 Comparison of 3D Out-Painting With Training From Scratch, Autoencoder, and In-Painting. . . . . 23

LIST OF FIGURES


| Figure | | Page |
|---|---|---|
| 1.1 | The General Pipeline of Self-supervised Learning. | 4 |
| 1.2 | Medical Images Are Highly Structured. Hence, Proxy Tasks Can Exploit Consistent and Recurrent Anatomy Present in Medical Images. | 5 |
| 3.1 | Qualitative Illustration of the Out-Painting Proxy Task. Give an Image With Missing Region (a), the Model Is Trained to Predict the Missing Region Utilizing (B) as a Ground Truth. | 10 |
| 3.2 | The Proposed Out-Painting Proxy Task. The Random Binary Mask M Is Applied on Arbitrarily-Size Patch X to Obtain Transformed Patch $\tilde{X}$. The Encoder-Decoder Architecture Is Trained to Learn a Generic Representation By Restoring the Original Patch X From the Transformed Ones $\tilde{X}$, Aiming to Yield Better Performance on Target Tasks via Transfer Learning. | 11 |
| 3.3 | An Example of (a) Central Region and (B) Random Region Binary Mask M Applied to Original Images X. | 12 |
| 3.4 | Limitation of Image Out-Painting Is That the Proxy Task Will Fail to Learn Generic Representation When Masked Region Does Not Contain Any Informative Part (*e.g.* Chest). | 13 |
| 4.1 | Qualitative Results of Image Out-painting. | 18 |
| 4.2 | Ablation Study. | 19 |
| 4.3 | Statistical Analysis Between Out-Painting in 2D and Training From Scratch, ImageNet, and the High Performing Self-Supervised Baseline Among Rotation and In-Painting. | 20 |

# Chapter 1

# INTRODUCTION

## 1.1 Background

Convolutional Neural Networks (CNNs) have been used in the computer vision for a long time (Lecun *et al.*, 2015). However, their true potential was realized when (Krizhevsky *et al.*, 2012) used CNNs to win ImageNet (Deng *et al.*, 2009) challenge in 2012 by designing considerably deep network containing 60 million parameters and training the network efficiently with graphics processing units (GPUs). Since then, CNNs have been widely used for various computer vision applications such as object detection (Ren *et al.*, 2015), semantic segmentation (Long *et al.*, 2015), etc. Nowadays, the popularity of CNNs is not limited to computer vision but across various applications such as natural language processing and medical image analysis.

However, the performance of CNNs greatly depends upon the capacity of network models and the amount of training data. Hence, the computer vision community have developed deep architectures to increase the network capability and larger datasets are collected. Various deep architectures such as AlexNet (Krizhevsky *et al.*, 2012), ResNet (He *et al.*, 2015), DenseNet (Huang *et al.*, 2017), etc have been proposed that achieves state-of-the-art performance on large scale dataset such as Places (Zhou *et al.*, 2014) and ImageNet (Deng *et al.*, 2009) containing 1.2 million annotated images of 1000 categories.

Interestingly, on continue training (fine-tuning) the deep models trained on the large labeled dataset (*e.g.,* ImageNet) achieve better performance than training from scratch on various target tasks. In other words, the deep models pre-trained on Im-

ageNet can be transferred via transfer learning to various medical imaging applications that outperform the models without any pre-training *i.e.,* training from scratch. This outstanding performance of pre-trained models on ImageNet is attributable to learned generic representation while classifying 1.2 million images among 1000 categories. Hence, this practice of using pre-trained CNN on ImageNet for various target tasks like image segmentation (Dai *et al.*, 2016), image captioning (Karpathy and Fei-Fei, 2015), and others (Lecun *et al.*, 2015), has become the *de facto* standard within the computer vision community.

## 1.2 Limitations of the Current Paradigm in Medical Image Analysis

The current paradigm of fine-tuning pre-trained models on ImageNet has become the *de facto* standard not only in the computer vision community but also in the medical imaging community. Surprisingly, even though the pre-trained models on ImageNet never see any biomedical images, they offer superior performance over training from scratch for a wide range of medical imaging applications. Moreover, Tajbakhsh *et al.* (2016) demonstrates that pre-training models achieve either equivalent or better performance over training from scratch and pre-trained models are always desirable for medical imaging applications. However, there are three limitations of current paradigm in medical image analysis:

1. **Domain gap:** Zhou *et al.* (2019) demonstrated that pre-trained models based on medical images are more powerful than pre-trained models on natural images for various medical imaging applications. In other words, *same-domain* transfer learning where models are pre-trained and fine-tuned within the same domain, is superior to *cross-domain* transfer learning where models are pre-trained and fine-tuned for a different domain.

2. **3D imaging tasks have to be solved in 2D:** The pre-trained models on ImageNet utilizes 2D convolutions and are not designed to process 3D cubes. However, to fit the current state-of-the-art ImageNet fine-tuning paradigm, 3D imaging tasks in the most prominent imaging modalities (e.g., CT and MRI) have to be reformulated and solved in 2D, losing rich 3D anatomical information and inevitably compromising the performance (Zhou *et al.*, 2019).

3. **Annotation cost:** Medical imaging community does not have a dataset as large as ImageNet because annotating biomedical images is not only tedious and time-consuming but also demanding of costly, specialty-oriented knowledge and skills, which are not easily accessible (Zhou *et al.*, 2017).

## 1.3 Research Question

To overcome the limitations of the current state-of-the-art paradigm of ImageNet fine-tuning (see section 1.2), we seek to answer the following question: *Can we develop a learning method directly from medical images, without utilizing systematic annotations, to pre-train a deep model that can achieve better performance on target tasks via transfer learning?* A promising direction to answer the question is to pre-train the models on *unlabeled images*. However, the unlabeled images do not provide any supervision to the models to learn generic representation. Hence, self-supervised representation learning can be utilized where *data intrinsically provides the supervision*. The following section discusses self-supervised representation learning in detail.

## 1.4 Self-supervised Representation Learning

Self-supervised learning methods aim at learning the representation from unlabeled images without utilizing human-annotated labels. To realize this aim, a popular

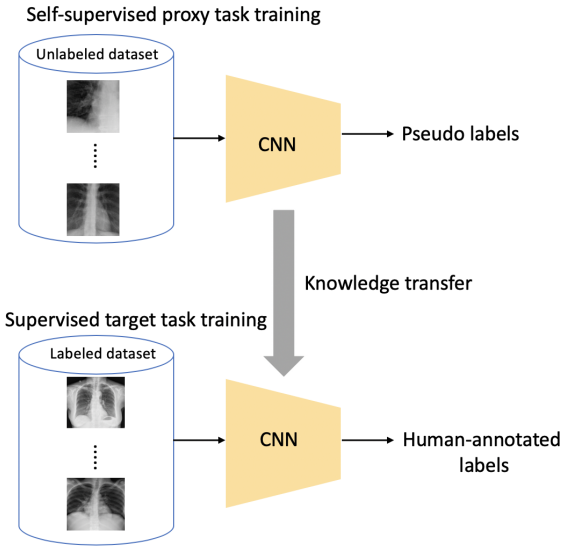



Figure 1.1: The general pipeline of self-supervised learning. First, CNN learns generic representation by solving pre-defined proxy tasks with pseudo labels. Subsequently, the learned parameters serve as a pre-train model and are transferred to target tasks utilizing human-annotated labels.

solution is to propose the proxy tasks for the models to solve, while the models are trained with the learning objectives of the proxy task. The proxy tasks are designed in such a way that the model needs to learn the representation to solve the proxy tasks. Note that the proxy tasks utilize pseudo labels that can be automatically generated based on the attributes of images.

Figure 1.1 shows the training pipeline of self-supervised learning methods. First, a proxy task is designed for models to solve, and pseudo labels are generated based on some attributes of the data. Then the models are trained with learning objectives of the proxy tasks and utilizing pseudo labels as ground truths. Once the self-supervised learning is finished, the learned visual representation can be transferred to target tasks via transfer learning. While there are various ways to transfer learned representation, we have used the entire pre-trained model and fine-tuned all the layers in the model on target tasks.

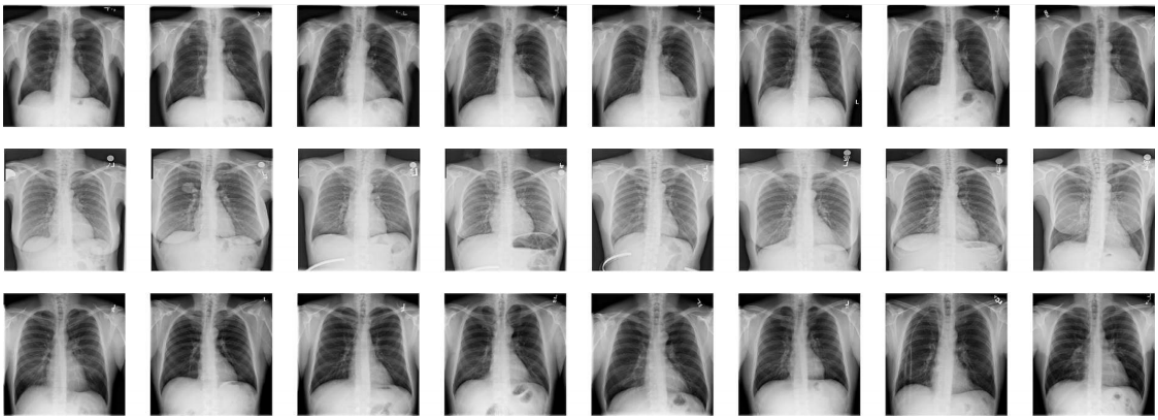


Figure 1.2: Medical images are highly structured. Hence, proxy tasks can exploit consistent and recurrent anatomy present in medical images.

## 1.5 Hypothesis and Proposed Solution

Unlike natural images, medical images are highly structured (Figure 1.2). For example, the anatomy of rib cage *i.e.,* arrangements of ribs is similar to a certain extent for most of the patients. Hence, we designed out-painting as a self-supervised proxy task to exploit consistent and recurrent anatomy present in medical images. We believe that if the model can predict the missing content when seeing only parts of the image, the model can learn a generic representation yielding better performance on target tasks via transfer learning. Moreover, by predicting the missing content, the model will learn global geometry and spatial layout of the organs in medical images. We describe out-painting as a self-supervised proxy task in detail in Chapter 3.

Once the model is pre-trained on out-painting proxy task, we evaluate the learned representation via transfer learning in seven medical imaging applications across modalities including 2D and 3D classification and segmentation target tasks. In six out of seven target tasks (see Chapter 4), fine-tuning models pre-trained on out-painting proxy task outperformed the models without any pre-training *i.e.,* training from scratch. Moreover, our proposed method is relatively more consistent and robust than other self-supervised proxy tasks and out-painting is competitive to state-

of-the-art pre-trained models on ImageNet. Owing to its outstanding performance, out-painting is utilized as one of the self-supervised proxy tasks to provide generic 3D pre-trained models for medical image analysis.

## 1.6 Terminology

To make this thesis easy to read, we define important terms used in the remaining sections. We follow the terminology conventions used in the survey paper (Jing and Tian, 2020) on self-supervised visual feature learning.

- **Human-annotated label:** It refers to the data labels that are manually annotated by a human.
- **Pseudo label:** Pseudo labels are automatically generated labels without any human intervention *i.e.* zero annotation cost.
- **Proxy task:** The neural networks are trained on proxy tasks using pseudo labels with an aim to learn generic representation by solving this task.
- **Target task:** Target tasks are various computer vision applications that are used to evaluate the learned representation by self-supervised learning.
- **Supervised learning:** The learning methods which use human-annotated labels to train the model are called as supervised learning approaches.
- **Unsupervised learning:** Unsupervised learning methods refers to learning methods without using either human-annotated labels or pseudo labels.
- **Self-supervised learning:** Self-supervised learning is a subset of unsupervised learning methods. In self-supervised learning methods CNNs are trained with pseudo labels with an aim to learn generic representation.

# Chapter 2

# RELATED WORK

Image out-painting is a self-supervised proxy task to learn generic representation. Hence, we contrast out-painting with the current state-of-the-art paradigm of supervised, unsupervised and self-supervised representation learning.

## 2.1 Supervised Representation Learning

Supervised representation learning utilizes human-annotated labels to learn generic representation. For example, ImageNet (Deng *et al.*, 2009) is one of the widely used datasets for pre-training the model contains 1.2 million human-annotated labels covering 1,000 classes. The practice of pre-training a model on ImageNet and then fine-tuning on various target tasks has become *de facto* standard within not only the computer vision community but also in the medical imaging community (Tajbakhsh *et al.*, 2016). As mentioned in section 1.2, the current state-of-the-art paradigm of fine-tuning pre-trained models on ImageNet has three limitations and our proposed method overcome these limitations by providing 2D and 3D pre-trained models directly from medical images at *zero* annotation costs.

## 2.2 Unsupervised Representation Learning

Unsupervised representation learning neither utilizes human-annotated labels nor pseudo labels to learn generic representation. Autoencoders (Hinton and Salakhutdinov, 2006) is closely related to our proposed method, as it learns representation by restoring the original input image. Unfortunately, the learned representation is likely to compress the image content without learning a semantically meaningful represen-

tation. Moreover, fine-tuning pre-trained models on out-painting proxy task outperforms fine-tuning from autoencoder proxy task (see Chapter 4) *i.e.,* out-painting learns more generic representation than autoencoders. Denoising autoencoders (Vincent *et al.*, 2008) overcome the issue of autoencoders by corrupting the input image and train the network to restore the original image. However, this corruption process is localized and the models do not require much semantic information to restore the original image. In contrast, to solve our proposed out-painting proxy task, the model requires a much deeper understanding of the anatomical structures present in medical images.

## 2.3 Self-supervised Representation Learning

Self-supervised methods aim at learning representation from unlabeled images *i.e.,* without using any human-annotated labels. To realize this aim, the computer vision community proposed a variety of proxy tasks (Jing and Tian, 2020) for the models to solve where they are trained with learning objectives of the proxy task, and representation is learned through this process. Specifically, Doersch *et al.* (2015) designed a proxy task to classify the relative positions of neighboring patches within an image, Gidaris *et al.* (2018) proposed rotation as a self-supervised proxy task to learn representation by discriminating the 0, 90, 180, and 270 degrees rotated images. Both above proxy tasks are image classification tasks, while out-painting is an image restoration task. Pathak *et al.* (2016) proposed image in-painting as a proxy task to only recover inner missing regions of a masked image, while our method restores the entire input patches. Moreover, Pathak *et al.* (2016) only provides a pre-trained encoder while our proposed method provides both pre-trained encoder and decoder. Further, on comparing our proposed out-painting proxy task with in-painting and rotation proxy task (see section 4.4.3), we found out-painting to learn more generic

representation effective across modalities. Finally, unlike the proposed method, all of the above methods provide 2D pre-trained models and cannot be utilized for 3D imaging target tasks. The 3D target tasks have to be reformulated in 2D, losing rich 3D anatomical information and inevitably compromising the performance.

Chapter 3

# METHODOLOGY

We humans have uncanny ability to imagine the missing content when seeing only parts of the image. For example, consider an X-ray patch with the missing region as shown in Figure 3.1. Although the border pixels are missing, most of us can predict its content from the surrounding pixels. This ability comes from the fact that medical images are highly structured and have consistent and recurrent anatomy. Hence, we explore whether state-of-the-art medical imaging algorithms can learn these anatomical structures. We demonstrate that by *predicting* these structures CNNs can learn generic representation that can achieve better performance on target tasks via transfer learning.

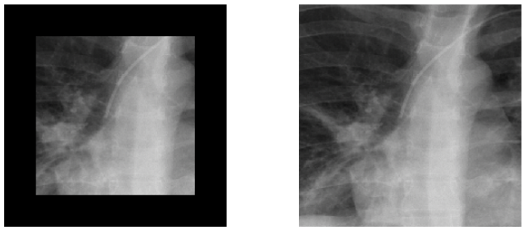



Figure 3.1: Qualitative illustration of the out-painting proxy task. Give an image with missing region (a), the model is trained to predict the missing region utilizing (b) as a ground truth.

## 3.1 Image Out-painting as a Self-supervised Proxy Task

Aiming at learning generic representation, we proposed image out-painting as a self-supervised proxy task. As shown in Figure 3.2, we first extract arbitrarily-size

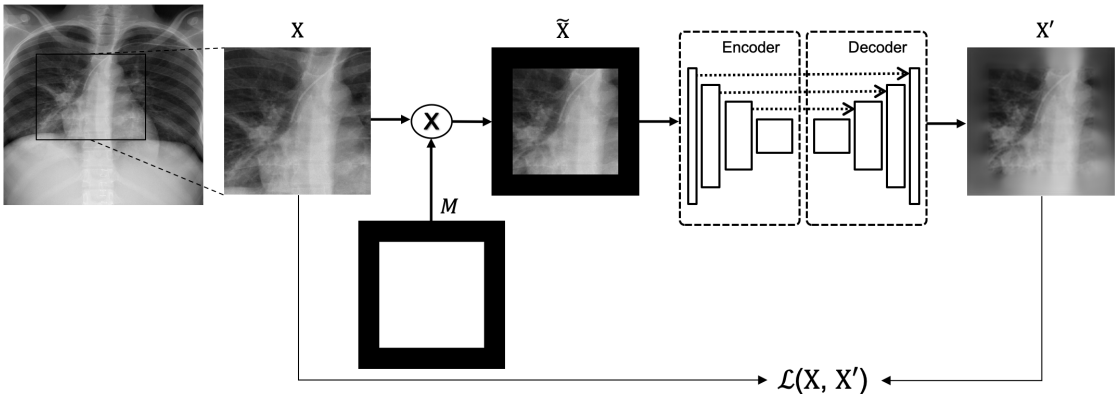



Figure 3.2: The proposed out-painting proxy task. The random binary mask M is applied on arbitrarily-size patch X to obtain transformed patch $\tilde{X}$. The encoder-decoder architecture is trained to learn a generic representation by restoring the original patch X from the transformed ones $\tilde{X}$, aiming to yield better performance on target tasks via transfer learning.

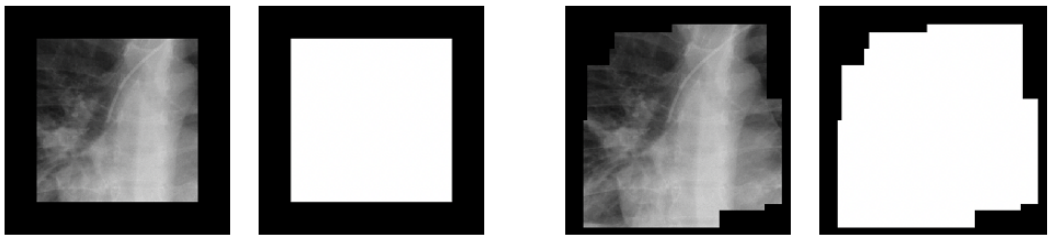
patch X cropped at a random location from an unlabeled image. Subsequently, we define a random binary mask M of the same dimensions as a patch X such that boundary pixels are set to zero. We obtain transformed patch $\tilde{X}$ with boundary pixels masked by combining binary mask M with the patch X.

Afterwards, the model $F(\cdot)$ containing an encoder-decoder architecture with skip connections in between and produces an output X' = $F(\tilde{X})$. The encoder learns to produce latent representation of transformed patches $\tilde{X}$. The decoder utilizes latent representation and restore original patch X from the transformed ones $\tilde{X}$.

**Reconstruction Loss:** The model $F(\cdot)$ is trained to minimize $L2$ distance between prediction X' and the ground truth X.

$$L(X) = ||F(X') - X)||_2^2 \tag{3.1}$$

By restoring the original patch X from the transformed ones X', the model $F(\cdot)$ will learn representation about global geometry and spatial layout of the organs via extrapolating within each patch. Once the model is pre-trained on out-painting proxy task, we can use the encoder and encoder-decoder both for target classification and



(a) Central region (b) Random region

Figure 3.3: An example of (a) central region and (b) random region binary mask M applied to original images X.

segmentation tasks respectively.

## 3.2 Region Masks

The difficulty of the out-painting proxy task depends upon the size and shape of binary mask M. To prevent the task from being too difficult or even unsolvable, we limit the masked region to be less than 1/4 of the whole image. However, the binary mask M could be of any shape, we present two different strategies here:

1. Central region: As shown in Figure 3.3, the simplest shape of binary mask M could be square. We randomly select the length and location of the square such that at most 1/4 area of the whole image is masked. There is a possibility that the model will learn low-level features such as boundary pattern or texture continuity to extrapolate the image. Such low-level features are not desirable to learn generic representation.

2. Random region: To remove the boundaries, we obtained arbitrarily shaped binary masks. Specifically, we generate an arbitrary number ($\leq 10$) of masks with various sizes and aspect ratios, then superimpose them on top of each other, resulting in a single mask of an arbitrary shape such that at most 1/4

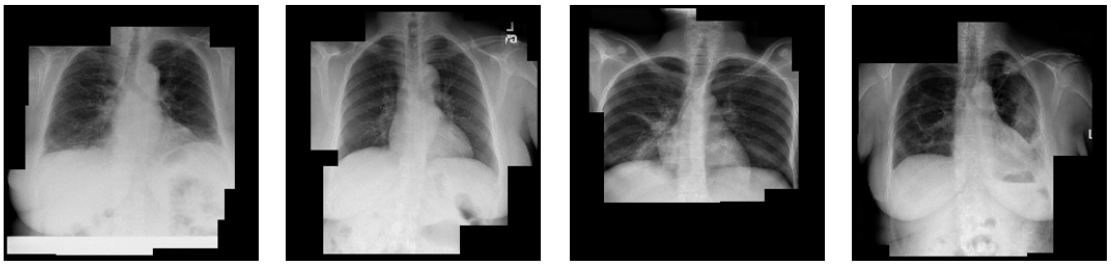


Figure 3.4: Limitation of image out-painting is that the proxy task will fail to learn generic representation when masked region does not contain any informative part (*e.g.* chest).

area of the whole image is masked. Although we found central and random region masks produce a similar generic representation, we use a random region for our experiments.

## 3.3 Limitations of Out-painting

The out-painting proxy task aims to learn anatomical structures present in medical images. However, if we apply random mask M to whole X-rays instead of the patches, we notice the drop of performance on the target tasks. This behavior was expected because the most informative part (*e.g.* chest) of X-rays is masked. In other words, we should mask the anatomical rich region of the images such that the model will learn generic representation by predicting the missing structures. For whole X-rays, we observed that in-painting performs better than out-painting on target tasks. However, this limitation is naturally eliminated in 3D because current GPUs can't fit entire CT scans and the deep models should be trained on sub-volumes. Moreover, the following chapter demonstrates that this limitation is mitigated because out-painting is a better proxy task than in-painting in four out of five 3D target tasks.

# Chapter 4

# EXPERIMENTS AND RESULTS

## 4.1 Implementation Details

### *4.1.1 2D Out-painting*

We trained our 2D out-painting proxy task on 77,074 X-rays in ChestXray8 (Wang *et al.*, 2017) dataset. Note that we did not utilize all 108,948 X-rays in the dataset for training to avoid test-image leaks between proxy and target tasks. This protocol will ensure that our pre-trained model never sees any testing images of target tasks. Moreover, out-painting proxy task utilize only unlabeled images with no annotations shipped with the datasets. U-Net (Ronneberger *et al.*, 2015) architecture with ResNet-18 (He *et al.*, 2015) encoder is used to train the proxy task.

### *4.1.2 3D Out-painting*

3D U-Net (Çiçek *et al.*, 2016) architecture is trained on unlabeled 623 Chest CT-scans in LUNA 2016 (Setio *et al.*, 2017). We did not utilize all 888 CT scans of the dataset to avoid test-image leaks between proxy and target tasks. Current GPUs cannot fit entire 3D CT scans, hence we train the model on sub-volumes of size 64×64×32 pixels. Following Zhou *et al.* (2019), we first randomly crop sub-volumes of size 64×64×32 pixels, then we exclude sub-volumes which are empty (air) or contain full tissues ensuring the model is trained only on informative sub-volumes.

Table 4.1: Summary of the target tasks.

| Code† | Modality | Dimension | Source | Description |
|---|---|---|---|---|
| DXC | X-ray | 2D | Wang *et al.* (2017) | Eight pulmonary diseases classification |
| IUC | Ultrasound | 2D | Hurst *et al.* (2010) | RoI, bulb, and background classification |
| NCC | CT | 2D/3D | Setio *et al.* (2017) | Lung nodule false positive reduction |
| LCS | CT | 3D | Bilic *et al.* (2019) | Liver segmentation |
| ECC | CT | 3D | Tajbakhsh *et al.* (2015) | Pulmonary embolism false positive reduction |
| BMS | MRI | 3D | Bakas *et al.* (2018) | Brain tumor segmentation |

† The first letter denotes the object of interest ("N" for lung nodule, "E" for pulmonary embolism, "L" for liver, etc); the second letter denotes the modality ("C" for CT, "X" for X-ray, "U" for Ultrasound, etc); the last letter denotes the task ("C" for classification, "S" for segmentation).

### *4.1.3 Target Tasks*

We have evaluated out-painting in seven medical imaging applications including 2D and 3D classification and segmentation tasks as shown in Table 4.1. We selected seven target tasks such that they are diverse in terms of organs, diseases, and modalities. In the following, we give a brief description of each target tasks:

1. **Eight pulmonary disease classification (DXC):** Wang *et al.* (2017) collected a dataset of 108,948 X-rays containing eight diseases: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax. The target task here is to classify X-rays among eight diseases where each image can have multiple diseases. The model is evaluated on 15,424 test images with Area Under the Curve (AUC) score.

2. **RoI/bulb/background classification (IUC):** The dataset contains 92 videos

in carotid intima-media thickness (CIMT) imaging with 8,021 frames (Hurst *et al.*, 2010). The target task is Region of Interest (RoI) localization *i.e.,* classify each frame among three classes: RoI, bulb, or background.

3. **Lung nodule false positive reduction (`NCC`):** Setio *et al.* (2017) provided 5,510,166 potential candidate locations of lung nodule in CT scans. The task is the binary classification of potential candidates in two classes nodules or non-nodules. Note that we can evaluate both 2D and 3D out-painting on the NCC target task as the former one uses slice-based solutions and the latter use volume-based solutions.

4. **Lung nodule segmentation (`NCS`):** The target task is to segment the lung nodules in the dataset provided by (LIDC-IDRI) (Armato III *et al.*, 2011) containing 1,018 CT scans. Each 3D CT scan and the nodules have been marked as volumetric binary masks. Following Zhou *et al.* (2020), we have re-sampled the volumes to 1-1-1 spacing and then extracted a 64×64×32 crop around each nodule. Similar to NCC, we can reformulate this target task in 2D by utilizing slice-based solution.

5. **Liver segmentation (`LCS`):** Bilic *et al.* (2019) provided 130 labeled CT scans containing binary masks of liver and lesion. In our experiments, the model is trained to segment only liver as a foreground and the rest as a background.

6. **Pulmonary embolism false positive reduction (`ECC`):** Following Tajbakhsh *et al.* (2015), we divided the dataset at the patient-level into a training set with 434 true positive Pulmonary Embolism (PE) candidates and 3,406 false positive PE candidates. The target task is binary classification whether PE candidates are true positives or false positives.

7. **Brain tumor segmentation (BMS):** Bakas *et al.* (2018) provided a dataset containing 285 brain MRI scans. Annotations include background (label 0) and three tumor subregions: GD-enhancing tumor (label 4), the peritumoral edema (label 2), and the necrotic and non-enhancing tumor core (label 1). We consider those with label 0 as negatives and others as positives and evaluate segmentation performance using Intersection over Union (IoU) scores.

#### *4.1.4 Hyper-parameters*

For all proxy tasks and target tasks, the raw image intensities were normalized to the [0, 1] range before training. The mean square error (MSE) loss or L2-norm is used as an objective function to train both proxy tasks. For 2D and 3D proxy tasks, we used stochastic gradient descent (SGD) method (Zhang, 2004) with an initial learning rate of $1e0$ for optimization. We used a learning rate scheduler such that if validation loss is not decreasing, the scheduler reduces the learning rate. For target tasks, we used Adam optimizer (Kingma and Ba, 2014) with learning rate of $1e-3$, where $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Moreover, for both proxy and target tasks, we used the *early-stop* mechanism to avoid over-fitting. We evaluate the proposed method using Area Under the Curve (AUC) and Intersection over Union (IoU) scores for classification and segmentation target tasks respectively, and further present statistical analysis based on an independent two-sample $t$-test.

### 4.2 Qualitative Results of Image Out-painting

We first qualitatively evaluate the image out-painting proxy task on test patches as shown in Figure 4.1. It is clear from the figure that the predictions from the pre-trained models are close to the original ground truth. Although the predictions are blurry for larger masked areas, the model can preserve anatomical information.

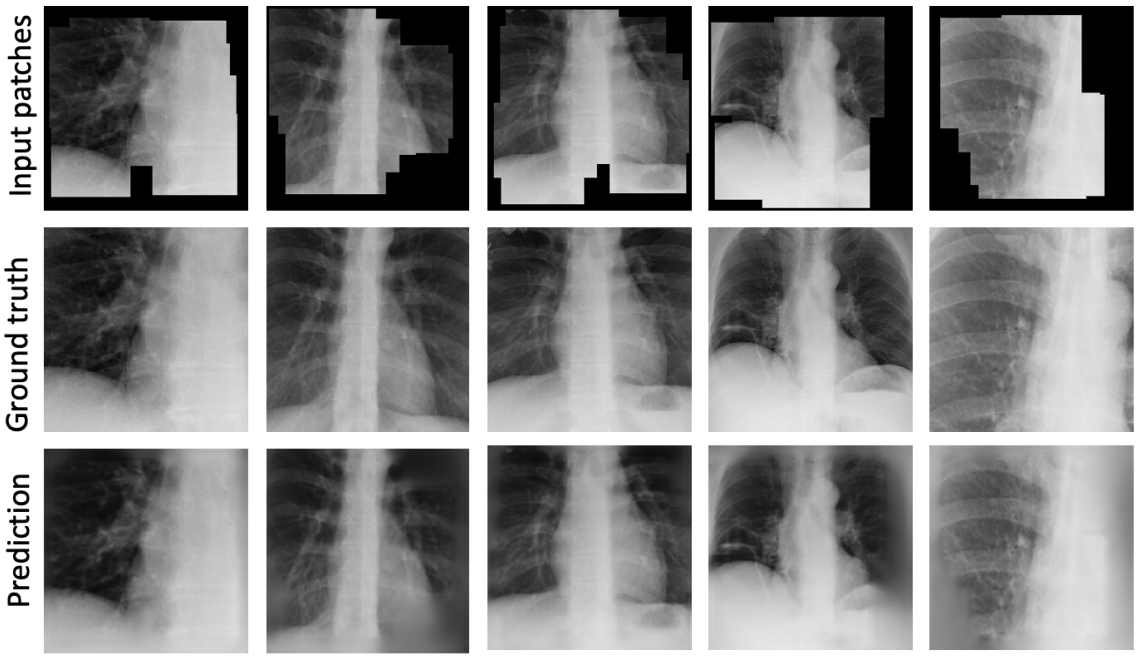



Figure 4.1: Qualitative results of image out-painting.

Figure 4.1 confirms our hypothesis that by doing the out-painting proxy task the model will learn to predict the anatomical structures present in the medical images.

## 4.3 Ablation Study

The size of the masked area is an important hyper-parameter in the proposed image out-painting proxy task. The larger masked areas make the proxy task too difficult or even unsolvable, while with smaller masked areas the model may not learn generic representation. Hence, we study the effect of the masked area in out-painting on two 3D target tasks `NCC` and `ECC`. Specifically, we trained the proposed image out-painting proxy task with different percentages of the masked area in the range [10, 70]. Once trained, we reported the mean and standard deviation of 10 rounds on two target tasks.

As shown in Figure 4.2, the model trained with just 10% of the masked area

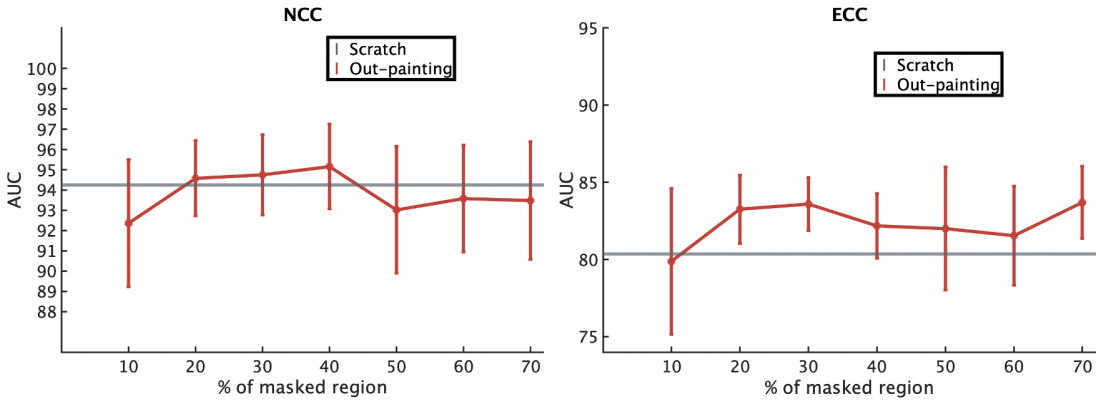



Figure 4.2: The effect of masked area in out-painting.

performs worst than training from scratch. These results confirm that smaller masked areas make the proxy task easier and the model may not learn generic representation. On the other hand, the larger masked areas (>50%) make the task too difficult and the model may not learn generic representation. However, we found out the performance of the out-painting proxy task is consistent when the masked area is 20-30%. Hence, we limit the masked area to be 25% of the whole image in our proposed image out-painting proxy task.

## 4.4 2D Results

Once the model is pre-trained on the out-painting proxy task, we evaluate the learned representation on four target tasks (see Table 4.1) via transfer learning. Specifically, we fine-tuned all the layers of the pre-trained model to perform target classification and segmentation tasks. Note that for classification target tasks (*e.g.,* `NCC`, `IUC`, and `DXC`), we only fine-tune the encoder of the pre-trained model and replacing the last layer with fully-connected layer, while for segmentation target tasks (*e.g.,* `NCS`) encoder and decoder both can be fine-tuned and replacing the last layer with $1 \times 1$ convolutional layer. In the following, we extensively compare the proposed

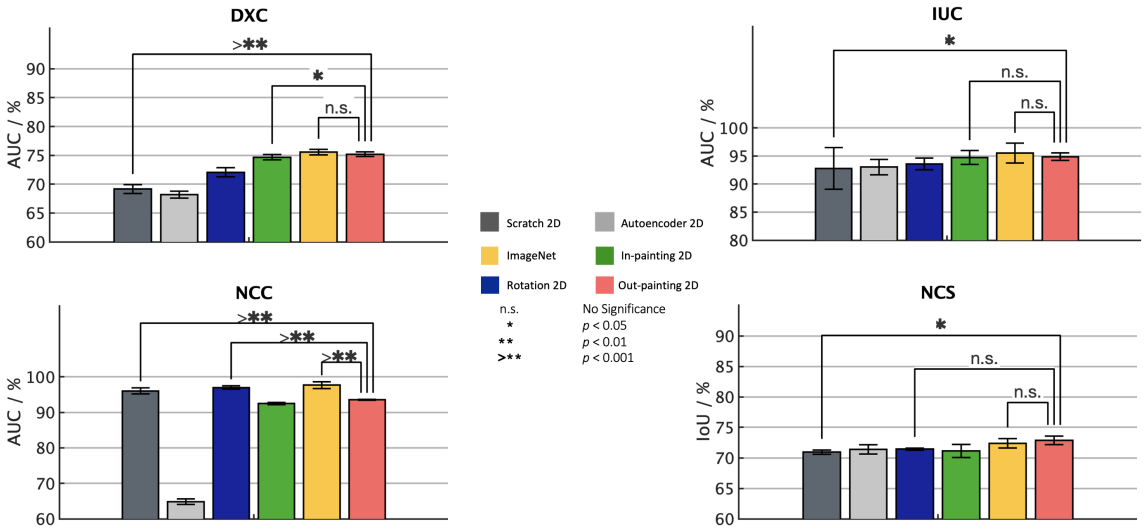



Figure 4.3: We conduct statistical analysis between out-painting in 2D and training from scratch, ImageNet, and the high performing self-supervised baseline among rotation and in-painting. The error bars represent the 95% confidence interval and the number of $*$ on the bridge indicates how significant two schemes are different from each other measured by $p$-value ("n.s." stands for "not statistically significant").

method with training from scratch, autoencoder, other self-supervised baselines such as rotation and in-painting, and state-of-the-art ImageNet fine-tuning.

### *4.4.1 Out-painting vs. Training from Scratch*

Figure 4.3 demonstrates that the fine-tuning models pre-trained on out-painting proxy task consistently outperforms the models without any pre-training *i.e.,* training from scratch. Our statistical analysis shows that in three target tasks (except `NCC`), the performance gain is significant over training from scratch. Specifically, for `DXC` where both proxy and target task is in the same domain, the pre-trained model achieves 5 points increase in AUC score over training from scratch. This is the remarkable achievement given that our pre-training comes at *zero* annotation cost. Moreover, to evaluate the generalizability of our pre-trained model on X-ray, we fine-tune the models on the target tasks in different modalities such as Ultrasound (`IUC`)

and CT (NCC and NCS). As shown in Figure 4.3, the pre-trained models achieves statistically significant performance gain over training from scratch for IUC and NCS target tasks. However, for NCC the out-painting underperform training from scratch possibly because of the large domain gap as our pre-trained model was not trained on CT slices. These mixed results in *cross-domain* transfer learning suggest that *same-domain* transfer learning should be preferred whenever possible.

### *4.4.2 Out-painting vs. Autoencoder*

We compare out-painting with simple autoencoder proxy task. Specifically, in the autoencoder proxy task, the model learns the representation by restoring the original input. As evident from Figure 4.3, out-painting consistently out-performs autoencoder in all four target tasks. This suboptimal representation learned by autoencoder *i.e.,* restoring the original input, further confirm the importance of masking the anatomical rich region of the images. In other words, it confirms our hypothesis that by predicting the anatomical structures the model can learn generic representation.

### *4.4.3 Out-painting vs. Self-supervised Baselines*

We further compare our proposed method with other self-supervised baselines such as RotNet (Gidaris *et al.*, 2018) and In-painting (Pathak *et al.*, 2016). To have a fair comparison, we used an identical experimental setup for both baselines and the proposed image out-painting proxy task. Figure 4.3 suggests that out-painting is better than both self-supervised baselines with $p > 0.05$ in three out of four target tasks (except NCC). Further analysis of the results reveals that there is no clear winner among the baselines that can always guarantee the highest performance in all four target tasks. However, out-painting is relatively more consistent and robust

than other baselines (except NCC) demonstrating the generalizability of the learned representation.

#### *4.4.4 Out-painting vs. ImageNet*

Our proposed out-painting proxy task is competitive to state-of-the-art ImageNet fine-tuning in three out of the four target tasks (except NCC). Note that ImageNet utilizes 1.2 million annotated images, whereas our pre-training has *zero* annotations costs. While our empirical results are strong, we recommend using pre-trained models on ImageNet, if available, for 2D medical image analysis.

### 4.5 3D Results

As mentioned in section 1.2, one of the limitations of the current state-of-the-art ImageNet fine-tuning paradigm is that we have to reformulate 3D imaging tasks and solved them in 2D. To overcome this limitation, we pre-train 3D U-Net (Çiçek *et al.*, 2016) using proposed out-painting proxy task and evaluate the learned representation on five 3D target tasks (see Table 4.1) via transfer learning. Specifically, we fine-tuned all the layers of the pre-trained model to perform target classification and segmentation tasks. Note that for classification target tasks (*e.g.,* NCC and ECC), we only fine-tune the encoder of the pre-trained model and replacing the last layer with fully-connected layer, while for segmentation target tasks (*e.g.,* NCS LCS, and BMS) encoder and decoder both can be fine-tuned and replacing the last layer with $1 \times 1 \times 1$ convolutional layer.

Table 4.2 demonstrates that the fine-tuning 3D models pre-trained on out-painting proxy task consistently outperforms their counterparts trained from scratch. Interestingly, even though 3D out-painting is trained on CT scans from LUNA 2016 dataset (section 4.1.2), fine-tuning the pre-trained models outperforms training from scratch

Table 4.2: Fine-tuning 3D models pre-trained on out-painting proxy task outperforms training from scratch in five 3D target tasks across organs, diseases, and modalities. Out-painting is better than in-painting in four out of five target tasks (except ECC), but out-painting is statistically equivalent to in-painting in ECC at $p = 0.05$ level. Note that the best result for each target task is highlighted in bold.

| Proxy task | NCC | NCS | ECC | LCS | BMS |
|---|---|---|---|---|---|
| Scratch | 94.25 ± 5.07 | 74.05 ± 1.97 | 79.99 ± 8.06 | 77.82 ± 3.87 | 58.52 ± 2.61 |
| Autoencoder | 91.50 ± 2.28 | 75.53 ± 0.63 | 81.32 ± 5.52 | 80.57 ± 4.89 | 57.33 ± 2.57 |
| In-painting | 94.46 ± 3.87 | 75.52 ± 0.66 | **83.90 ± 2.63** | 81.05 ± 3.25 | 61.13 ± 2.66 |
| Out-painting | **95.16 ± 1.75** | **76.20 ± 0.65** | 82.88 ± 4.38 | **82.75 ± 2.64** | **65.79 ± 1.05** |

in *cross-domain* transfer learning *i.e.,* LCS, ECC, and BMS target tasks. Moreover, out-painting not only outperforms training from scratch but also autoencoder proxy task in all five target tasks. Recall that in the autoencoder proxy task, the model learns the representation by restoring the original input. This performance gain over training from scratch and autoencoder is significant because our proposed pre-training requires no human-annotated labels by successfully utilizing unlabeled medical images.

We further compare our 3D out-painting with other self-supervised baselines such as in-painting Pathak *et al.* (2016). Note that (Pathak *et al.*, 2016) proposed in-painting for 2D natural images and it did not provide 3D pre-trained models to solve 3D imaging tasks. However, to have a fair comparison, we implemented in-painting in 3D and used identical experimental setup for both in-painting and out-painting proxy tasks. As shown in Table 4.2, out-painting is better than in-painting in four out of five target tasks (except ECC), but out-painting is statistically equivalent to in-painting in ECC at $p = 0.05$ level. Further, out-painting is more generalizable and robust than in-painting in *cross-domain* transfer learning demonstrating the superiority of learned representation by the proposed proxy tasks.

## 4.6 Application of Out-painting

Zhou *et al.* (2019) proposed Models Genesis, one of the early efforts to provide generic 3D pre-trained models for medical image analysis. Models Genesis unified four self-supervised proxy tasks including *out-painting* into single image restoration task with an aim to learn from multiple perspectives, yielding *generic* pre-trained models. Ablation study (Zhou *et al.*, 2020) suggests that out-painting is the second-best proxy task out of the four, leading Models Genesis to learn generic representation. Hence, out-painting is utilized as a self-supervised proxy task to provide pre-trained 3D models for medical image analysis.

# Chapter 5

# CONCLUSION

In this thesis, we propose a novel image out-painting as a self-supervised proxy task that can learn the visual representation directly from medical images without using any human-annotated labels. We demonstrated that the proposed proxy task, requiring no annotations, outperforms training from scratch in six out of seven 2D and 3D medical imagining applications covering both classification and segmentation. Moreover, in 2D, image out-painting is more consistent and robust as compared to current self-supervised baselines like rotation and in-painting and offers competitive performance to state-of-the-art models pre-trained on ImageNet. Further, we observed that in-painting performs better than out-painting on target tasks when proxy tasks are trained on whole X-rays rather than patches of X-rays. However, this limitation is mitigated in 3D because current GPU memory can't fit the entire 3D CT scan and deep models have to process sub-volumes. Also, fine-tuning 3D models pre-trained on out-painting proxy tasks outperforms training from scratch and in-painting in four out of five 3D target tasks. Owing to its outstanding performance, out-painting is utilized as one of the self-supervised proxy tasks to provide generic 3D pre-trained models for medical image analysis.

We have proposed a *generative* approach to learn generic representation by *predicting* the missing content when seeing only parts of the image. However, the alternative is to design the *discriminative* proxy task to learn representation. Specifically, for natural images, the discriminative approaches based on contrastive learning in the latent space have recently achieved state-of-the-art results (Chen *et al.*, 2020). Nonetheless, contrastive learning requires huge mini-batch sizes or memory banks, making it

impractical for 3D medical imaging applications. Hence, the success of contrastive representation learning methods is yet to be explored in 3D medical image analysis.

## REFERENCES


Armato III, S. G., G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans", Medical physics 38, 2, 915–931 (2011).

Bakas, S., M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge", arXiv preprint arXiv:1811.02629 (2018).

Bilic, P., P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser *et al.*, "The liver tumor segmentation benchmark (lits)", arXiv preprint arXiv:1901.04056 (2019).

Chen, T., S. Kornblith, M. Norouzi and G. Hinton, "A simple framework for contrastive learning of visual representations", arXiv preprint arXiv:2002.05709 (2020).

Çiçek, Ö., A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation", in "International conference on medical image computing and computer-assisted intervention", pp. 424–432 (Springer, 2016).

Dai, J., K. He and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades", in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", pp. 3150–3158 (2016).

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database", in "CVPR09", (2009).

Doersch, C., A. Gupta and A. A. Efros, "Unsupervised visual representation learning by context prediction", in "Proceedings of the IEEE international conference on computer vision", pp. 1422–1430 (2015).

Gidaris, S., P. Singh and N. Komodakis, "Unsupervised representation learning by predicting image rotations", arXiv preprint arXiv:1803.07728 (2018).

He, K., X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", CoRR abs/1512.03385, URL `http://arxiv.org/abs/1512.03385` (2015).

Hinton, G. E. and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", science 313, 5786, 504–507 (2006).

Huang, G., Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely connected convolutional networks", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 4700–4708 (2017).

Hurst, R. T., R. F. Burke, E. Wissner, A. Roberts, C. B. Kendall, S. J. Lester, V. Somers, M. E. Goldman, Q. Wu and B. Khandheria, "Incidence of subclinical atherosclerosis as a marker of cardiovascular risk in retired professional football players", The American journal of cardiology 105, 8, 1107–1111 (2010).

Jing, L. and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey", IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

Karpathy, A. and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 3128–3137 (2015).

Kingma, D. P. and J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980 (2014).

Krizhevsky, A., I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in "Advances in neural information processing systems", pp. 1097–1105 (2012).

Lecun, Y., Y. Bengio and G. Hinton, "Deep learning", Nature Cell Biology 521, 7553, 436–444 (2015).

Long, J., E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 3431–3440 (2015).

Pathak, D., P. Krahenbuhl, J. Donahue, T. Darrell and A. A. Efros, "Context encoders: Feature learning by inpainting", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 2536–2544 (2016).

Ren, S., K. He, R. Girshick and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", in "Advances in neural information processing systems", pp. 91–99 (2015).

Ronneberger, O., P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in "International Conference on Medical image computing and computer-assisted intervention", pp. 234–241 (Springer, 2015).

Setio, A. A. A., A. Traverso, T. De Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge", Medical image analysis 42, 1–13 (2017).

Tajbakhsh, N., M. B. Gotway and J. Liang, "Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks", in "International Conference on Medical Image Computing and Computer-Assisted Intervention", pp. 62–69 (Springer, 2015).

Tajbakhsh, N., J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?", IEEE transactions on medical imaging 35, 5, 1299–1312 (2016).

Vincent, P., H. Larochelle, Y. Bengio and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders", in "Proceedings of the 25th international conference on Machine learning", pp. 1096–1103 (2008).

Wang, X., Y. Peng, L. Lu, Z. Lu, M. Bagheri and R. 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", pp. 2097–2106 (2017).

Zhang, T., "Solving large scale linear prediction problems using stochastic gradient descent algorithms", in "Proceedings of the twenty-first international conference on Machine learning", p. 116 (2004).

Zhou, B., A. Lapedriza, J. Xiao, A. Torralba and A. Oliva, "Learning deep features for scene recognition using places database", in "Advances in neural information processing systems", pp. 487–495 (2014).

Zhou, Z., J. Shin, L. Zhang, S. Gurudu, M. Gotway and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 7340–7351 (2017).

Zhou, Z., V. Sodha, J. Pang, M. B. Gotway and J. Liang, "Models genesis", arXiv preprint arXiv:2004.07882 (2020).

Zhou, Z., V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis", in "International Conference on Medical Image Computing and Computer-Assisted Intervention", pp. 384–393 (Springer, 2019).