Recent advances and clinical applications of deep learning in medical image analysis

Xuxin Chen\textsuperscript{a}, Ximin Wang\textsuperscript{b}, Ke Zhang\textsuperscript{a}, Roy Zhang\textsuperscript{c}, Kar-Ming Fung\textsuperscript{c}, Theresa C. Thai\textsuperscript{d}, Kathleen Moore\textsuperscript{e}, Robert S. Mannel\textsuperscript{e}, Hong Liu\textsuperscript{a}, Bin Zheng\textsuperscript{a}, Yuchen Qiu\textsuperscript{a}\textsuperscript{*}

\textsuperscript{a} School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK, USA 73019
\textsuperscript{b} School of Information Science and Technology, ShanghaiTech University, Shanghai, China 201210
\textsuperscript{c} Department of Pathology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA 73104
\textsuperscript{d} Department of Radiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA 73104
\textsuperscript{e} Department of Obstetrics and Gynecology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA 73104

ABSTRACT

Deep learning has become the mainstream technology in computer vision, and it has received extensive research interest in developing new medical image processing algorithms to support disease detection and diagnosis. As compared to conventional machine learning technologies, the major advantage of deep learning is that models can automatically identify and recognize representative features through the hierarchical model architecture, while avoiding the laborious development of hand-crafted features. In this paper, we reviewed and summarized more than 200 recently published papers to provide a comprehensive overview of applying deep learning methods in various medical image analysis tasks. Especially, we emphasize the latest progress and contributions of state-of-the-art unsupervised and semi-supervised deep learning in medical images, which are summarized based on different application scenarios, including lesion classification, segmentation, detection, and image registration. Additionally, we also discussed the major technical challenges and suggested the possible solutions in future research efforts.

Keywords: Deep learning, unsupervised learning, self-supervised learning, semi-supervised learning, medical images, classification, segmentation, detection, registration

1. INTRODUCTION

In current clinical practice, accuracy of detection and diagnosis of cancers and/or many other diseases depends on the expertise of individual clinicians (i.e., radiologists or pathologists) (Kruger et al., 1972), which results in large inter-reader variability in reading and interpreting medical images. In order to address and overcome this clinical challenge, many computer-aided detection and diagnosis (CAD) schemes have been developed and tested aiming to help clinicians more efficiently read medical images and make the diagnostic decision in a more accurate and objective manner. The scientific rationale of this approach is that using computer-aided quantitative image feature analysis can help overcome many negative factors in the clinical practice, including the wide variations in expertise of the clinicians, potential fatigue of human experts, and lack of sufficient medical resources.

Although early CAD schemes have been developed in 1970s (Meyers et al., 1964; Kruger et al., 1972; Sezaki and Ukena, 1973), progress of the CAD schemes accelerates since the middle of 1990s (Doi et al., 1999) due to the development and integration of more advanced machine learning methods or models to CAD schemes. For conventional CAD schemes, a common developing procedure consists of three steps: target segmentation, feature computation, and disease classification. For example, Shi et al. (2008) developed a CAD scheme to achieve mass classification on digital

\textsuperscript{*} Corresponding author: qiuyuchen@ou.edu
mammograms. The regions of interest (ROIs) containing the target masses were first segmented from the background using a modified active contour algorithm (Sahiner et al., 2001). Then a large number of image features were applied to quantify the lesion characteristics in size, morphology, margin geometry, texture, and etc. Thus the raw pixel data were converted into a vector of representative features. Finally, a LDA (linear discrimination analysis) based classification model was applied on the feature vector to identify the mass malignancy.

As a comparison, for deep learning based models, hidden patterns inside ROIs are progressively identified and learned by the hierarchical architecture of deep neural networks (LeCun et al., 2015). During this process, important properties of the input image will be gradually identified and amplified for certain tasks (e.g. classification, detection), while irrelevant features will be attenuated and filtered out. For instance, an ultrasound image depicting suspicious lesions comes with a pixel array (Tanaka et al., 2019), and each entry is used as one input feature of the deep learning model. The first several layers of the model may initially obtain some basic lesion information, such as tumor shape, location, and orientation. The next batch of layers may identity and keep the features consistently related to lesion malignancy (e.g. shape, edge irregularity), while ignoring irrelevant variations (e.g. location). The relevant features will be further processed and assembled by subsequent higher layers in a more abstract manner. When increasing the number of layers, a higher level of feature representations can be achieved. Through the entire procedure, important features hidden inside the raw image are recognized by a general neural network based model in a self-taught manner, and thus the manual feature development is not needed.

Due to its huge advantage, deep learning related methods have become the mainstream technology in the CAD field and have been widely applied in a variety of tasks, such as disease classification (Li et al., 2020a; Shorffuzzaman and Hossain, 2021; Zhang et al., 2020a; Frid-Adar et al., 2018a; Kumar et al., 2016), ROI segmentation (Alom et al., 2018; Yu et al., 2019; Fan et al., 2020), medical object detection (Rijthoven et al., 2018; Mei et al., 2021; Nair et al., 2020; Zheng et al., 2015), and cross-modality image registration (Simonovsky et al., 2016; Sokooti et al., 2017; Balakrishnan et al., 2018). Among different kinds of deep learning techniques, supervised learning was first adopted in medical image analysis. Although it has been successfully utilized in many applications (Esteve et al., 2017; Long et al., 2017), further deployment of supervised models in many scenarios is majorly hindered by the limited size of most medical datasets. As compared to regular computer vision tasks, the medical image dataset usually contains relatively small amounts of images (less than 10,000), and in many cases, only a small percentage of images are annotated by experts. To overcome this limitation, unsupervised and semi-supervised learning methods have received extensive attention in the past three years, which are able to 1) generate more labeled images for model optimization, 2) learn meaningful hidden patterns from unlabeled image data, and 3) generate pseudo labels for the unlabeled data.

There are several review articles that have summarized deep learning applications in medical image analysis. Litjens et al. (2017) and Shen et al. (2017) reviewed relatively early deep learning techniques, which are mainly based on supervised methods. Yi et al. (2019) and Kazeminia et al. (2020a) specifically reviewed the applications of generative adversarial networks across different medical imaging tasks, such as classification, segmentation, lesion detection, and registration. Cheplygina et al. (2019) reviewed how to use semi-supervised learning and multiple instance learning in diagnosis or segmentation tasks. Tajbakhsh et al. (2020) broadly investigated ways to deal with dataset limitations (e.g., scarce or weak annotations) in medical image segmentation. In contrast, our survey is comprehensive and technically-oriented, which especially features applications of the latest deep learning technologies in various clinical scenarios. This survey aims to (1) provide an in-depth overview of recent advances in deep learning, with a focus on unsupervised and semi-supervised approaches; (2) summarize the major contributions of applying deep learning techniques in four main clinical areas: classification, segmentation, detection, and registration; (3) discuss challenges for further improving the performance of deep learning based CAD schemes and suggest possible perspectives on future research directions.

2. OVERVIEW OF DEEP LEARNING METHODS

Depending on whether labels of the training dataset are present, deep learning can be roughly divided into supervised, unsupervised, and semi-supervised learning. In supervised learning, all training images are labeled, and the model is optimized using the image-label pairs. For each testing image, the optimized model will generate a likelihood score to predict its class label (LeCun et al., 2015). For unsupervised learning, the model will analyze and learn the underlying patterns or hidden data structures without labels. If only a small portion of training data is labeled, the model learns input-output relationship from the labeled data, and the model will be strengthened by learning semantic and fine-grained features from the unlabeled data. This type of learning approach is defined as semi-supervised learning (van Engelen and Hoos, 2020).
2.1. Supervised learning

2.1.1. Convolutional neural networks (CNNs)

Convolutional neural networks (CNNs) are a widely used deep learning architecture in medical image analysis (Anwar et al., 2018). CNNs are mainly composed of convolutional layers and pooling layers. Figure 1 shows a simple CNN in the context of medical image classification task. The CNN directly takes an image as input, and transforms it via convolutional layers, pooling layers, and fully connected layers, and finally outputs a class-based likelihood of that image.

At each convolutional layer \( l \), a bunch of kernels \( W = \{W_1, ..., W_k\} \) are used to extract features from the input image, and biases \( b = \{b_1, ..., b_k\} \) are added, generating new feature maps \( W_i^l x_i^l + b_i^l \). Then a non-linear transform, an activation function \( \sigma(\cdot) \), is applied resulting in \( x_i^{l+1} = \sigma(W_i^l x_i^l + b_i^l) \) as the input of the next layer. After the convolutional layer, a pooling layer is incorporated to reduce the dimension of feature maps, thus reducing the number of parameters. Average pooling and maximum pooling are two common pooling operations. The above process is repeated for the rest layers. At the end of the network, fully connected layers are usually employed to produce the probability distribution over classes via a sigmoid or softmax function. The predicted probability distribution gives a label \( \hat{y} \) for each input instance so that a loss function \( L(\hat{y}, y) \) can be calculated, where \( y \) is the real label. Parameters of the network are iteratively optimized by minimizing the loss function.

![Figure 1. A simple CNN for disease classification from MRI images (Anwar et al., 2018).](image)

2.1.2. Recurrent neural networks (RNNs)

Due to the advantage of being able to process input of any length and historical information, recurrent neural networks (RNNs) models can be used in analyzing sequencing data, such as MR image sequences (Bai et al., 2018). Different from CNNs, RNNs allow previous outputs to be used as inputs while having hidden states. In a plain RNN model, a hidden state \( h_t \) is a nonlinear transform from the combination of its input \( x_t \) and previous hidden state \( h_{t-1} \), expressed as

\[
    h_t = \sigma(W x_t + R h_{t-1} + b)
\]

where \( W \) and \( R \) are weights shared over time. Plain RNNs suffer from vanishing gradient problems, while Long Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRUs) (Cho et al., 2014) avoid this issue and perform better by keeping relevant information and removing irrelevant information.

2.2. Unsupervised models

2.2.1. Autoencoders

Autoencoders are widely applied in dimensionality reduction and feature learning (Hinton and Salakhutdinov, 2006). The simplest autoencoder, initially known as auto-associator (Bourlard and Kamp, 1988), is a neural network with only one hidden layer that learns a latent feature representation of the input data by minimizing a reconstruction loss.
between the input and its reconstruction from the latent representation. The shallow structure of simple autoencoders limits their representation power, but deeper autoencoders with more hidden layers can improve the representation. By stacking multiple auto-encoders and optimizing them in a greedy layer-wise manner, deep autoencoders or Stacked Autoencoders (SAEs) can learn more complicated non-linear patterns than shallow ones and thus generalize better outside training data (Bengio et al., 2007). SAEs consist of an encoder network and a decoder network, which are typically symmetrical to each other. To further force models to learn useful latent representations with desirable characteristics, regularization terms such as sparsity constraints in Sparse Autoencoders (Ranzato et al., 2007) can be added to the original reconstruction loss. Other regularized autoencoders include the Denoising Autoencoder (Vincent et al., 2010) and Contractive Autoencoder (Rifai et al., 2011), both designed to be insensitive to input perturbations.

Unlike traditional autoencoders mapping input to a fixed vector, variational autoencoder (VAE) (Kingma and Welling, 2013) learns a probabilistic distribution for feature representations. The encoder network maps input data to latent space by learning two vectors: the mean and standard deviation of a Gaussian distribution; the decoder network samples vectors: the mean and standard deviation of a Gaussian distribution; the decoder network samples

2.2.2. Generative adversarial networks (GANs)

Generative adversarial networks (GANs) are a class of deep learning framework for generative modeling first proposed in by Goodfellow et al. (2014). Since then numerous research papers have been published discussing GANs’ applications in medical image analysis. In contrast to the existing review papers (Kazemini et al., 2020b; Yi et al., 2019), we reviewed the research progress of GANs from two aspects: 1) different loss (objective) functions, and 2) conditional and unconditional settings. Different loss functions are usually accompanied by architecture changes of the generator and discriminator networks. Interested readers can refer to the paper (Yi et al., 2019) for a detailed review of architecture changes.

I. Background of GANs

Vanilla GAN is the first GAN model (Goodfellow et al., 2014), in which a framework for estimating generative models is designed to directly draw samples from the desired underlying data distribution without the need to explicitly define a probability distribution. It consists of two models: a generator G and a discriminator D. The generative model G takes as input a random noise vector z sampled from a prior distribution P_r(z), often either a Gaussian or a uniform distribution, and then maps z to data space as G(z, \theta_g), where G is a neural network with parameters \theta_g. The fake samples denoted as G(z) or x_g are expected to resemble real samples from the training data P_r(x), and these two types of samples are sent into D. The discriminator, a second neural network parameterized by \theta_d, outputs the probability D(x, \theta_d) that a sample comes from the training data rather than G. The overall objective training function L(D, G) can be expressed as follows.

\[
\min_G \max_D L(D, G) = \mathbb{E}_{x \sim P_r(x)}[\log D(x)] + \mathbb{E}_{z \sim P_z(z)}[\log (1 - D(G(z)))]
\]

(1)

G and D are optimized alternately using the following objective functions.

\[
L_D = \max_D \mathbb{E}_{x \sim P_r(x)}[\log D(x)] + \mathbb{E}_{z \sim P_z(z)}[\log (1 - D(G(z)))]
\]

(2)

\[
L_G = \min_G \mathbb{E}_{x \sim P_g(x)}[\log (1 - D(x_g))]
\]

(3)

The training procedure is like playing a minimax two-player game. The discriminative network D is optimized to maximize the log likelihood of assigning correct labels to fake samples and real samples, while the generative model G is trained to maximize the log likelihood of D making a mistake. Through the adversarial process, G is desired to gradually estimate the underlying data distribution and generate realistic samples.

II. Modified loss functions

II. A. New probability distance metrics

Non-saturating loss (Vanilla GAN): The original loss function L_G saturates quickly at the early stage of training. Goodfellow et al. (2014) proposed to train G by minimizing −\log D(G(z)). The corresponding loss functions for D and G are

\[
L_D = -\mathbb{E}_{x \sim P_r(x)}[\log D(x)] - \mathbb{E}_{x \sim P_g(x)}[\log (1 - D(x))]
\]

(4)
\[ L_G = -\mathbb{E}_{x \sim p_d(x)} \log D(x_g) \] (5)

Minimizing \( L_G \) was proved equivalent to minimizing Jensen-Shannon (JS) divergence between the real data distribution and the synthetic data distribution.

**Wasserstein loss (WGAN):** In Wasserstein GAN (WGAN), Earth-Mover (EM) distance or Wasserstein-1, commonly known as the Wasserstein distance, was proposed to replace the JS divergence to measure the distance between the real and synthetic data distribution (Arjovsky et al., 2017a). The critic of WGAN has the advantage to provide useful gradients information where JS divergence saturates and results in vanishing gradients. WGAN could also improve the stability of learning and alleviate problems like mode collapse. The corresponding loss functions are

\[ L_D = -\mathbb{E}_{x \sim p_r(x)} [D(x)] - \mathbb{E}_{x \sim p_g(x)} [1 - D(x_g)] \] (6)

\[ L_G = -\mathbb{E}_{x \sim p_g(x)} [D(x_g)] \] (7)

There exist a variety of probability distance metrics (Saxena and Cao, 2020), such as least square (Mao et al., 2017), f-divergence (Nowozin et al., 2016), and Maximum Mean Discrepancy (Li et al., 2017a). For example, in Least Squares GAN (LSGAN), the authors claimed that the sigmoid cross entropy loss function adopted in regular GANs might lead to the vanishing gradients problem, and proposed to use the least-squares (LS) loss, which saturates slower than the cross-entropy loss. Optimizing the LS loss yields minimizing the Pearson \( \chi^2 \) divergence (Mao et al., 2017).

**II. B. Regularizations and normalizations (WGAN-GP, SN-GAN)**

As an improvement to WGAN, WGAN-GP utilizes an alternative way to enforce the Lipschitz constraint by introducing a penalty term to the gradient norm of D (Gulrajani et al., 2017). Empirically WGAN-GP has demonstrated better performance and stability than WGAN across a variety of architectures. Compared to WGAN and WGAN-GP, the authors of SN-GAN proved that the so-called spectral normalization could not only improve the quality of generated images, but also make discriminator training more stable. Spectral normalization normalizes the weight matrix of each layer with their spectral norm so that the Lipschitz constraint \( \text{Lipschitz} = 1 \) can be satisfied.

To conclude, new loss functions to stabilize training and reduce mode collapse is a hot research topic in GANs. For a more comprehensive and in-depth review, we refer interested readers to (Wiatrak et al., 2019; Saxena and Cao, 2020; Kurach et al., 2019).

**III. GANs in conditional settings**

cGAN: An unconditional generative model like the vanilla GAN cannot explicitly control the modes of data being synthesized. To guide the data generation process, the conditional GAN (cGAN) is constructed by conditioning its generator and discriminator with additional information (i.e., the class labels) (Mirza and Osindero, 2014). Specifically, the noise vector \( z \) and class label \( c \) are jointly provided to \( G \); the real/fake data and class label \( c \) are together presented as the inputs of \( D \). The conditional information can also be images or other attributes, not limited to class labels. The loss function of cGAN is as below.

\[ \min_G \max_D \mathbb{L}(D, G) = \mathbb{E}_{x \sim p_r(x)} [\log D(x|c)] + \mathbb{E}_{z \sim p_z(x)} [\log (1 - D(G(z)|c))] \] (8)

ACGAN: The auxiliary classifier GAN (ACGAN) presents a new strategy to employ label conditioning to improve image synthesis (Odena et al., 2017). Unlike the discriminator of cGAN, \( D \) in ACGAN is no longer provided with the class conditional information. Instead, \( D \) apart from separating real and fake images, is tasked with reconstructing class labels. When forcing \( D \) to perform the additional classification task, ACGAN can generate high-quality images easily.

**2.2.3. Self-supervised learning**

Unsupervised learning is more than just generating data (in generative models); a broader goal is to learn rich and meaningful feature representations from the raw unlabeled data, which can be useful for a wide variety of downstream tasks, such as classification, object detection, instance segmentation, etc. Another motivation behind unsupervised learning is to avoid supervised tasks that are often expensive and time-consuming due to the need to establish new labeled datasets or acquire high-quality annotations in certain fields like medicine. Meanwhile, there usually exist large amounts of cheap unlabeled data remaining unexploited in many fields; naturally, unsupervised learning is used to leverage the power of unlabeled data to further improve the performance and efficiency in supervised tasks. Since unsupervised learning touches upon vaster data than supervised learning, the learnt features can generalize better potentially in the real world. In recent literature, the term *self-supervised learning* is used interchangeably with *unsupervised learning*; however, self-supervised
learning actually refers to a form of deep unsupervised learning, where inputs and labels are created from the unlabeled data itself. The supervision can be created in two ways: pretext tasks based methods and contrastive learning based methods.

**Pretext task** is designed to learn representative features for the downstream tasks, but the pretext itself is not of our true interest (He et al., 2020). The pretext tasks learn representations by hiding certain information (e.g., channel, patches, etc.) for each input image, and then predict the missing information from the image’s remaining parts. Examples include image inpainting (Pathak et al., 2016), colorization (Zhang et al., 2016a), relative patch prediction (Doersch et al., 2015), jigsaw puzzles (Noroozi and Favaro, 2016), rotation (Gidaris et al., 2018), etc. However, the learnt representations’ generalizability is heavily dependent on the quality of hand-crafted pretexts (Chen et al., 2020a).

**Contrastive learning** approaches can date back to the pioneering work conducted by Becker and Hinton (1992). For this method, a contrastive loss is adopted to optimize the model, so that similarity can be maximized for positive (similar) pairs, and minimized for negative (dissimilar) pairs (Chen et al., 2020a; Chaitanya et al., 2020). For each image, positive pairs are mostly built through different image transformations, while the rest images are taken as negative examples. Momentum Contrast (MoCo) (Figure 2(a)) consists of an encoder to encode an image as a query \( q \), and a momentum encoder to generate keys \( k_0, k_1, k_2, \ldots \) from a set of data samples; a query and a key are considered as a positive pair if they stem from the same image (He et al., 2020). In another study, Chen et al. (2020a) proposed a framework called SimCLR (Figure 2(b)) consisting of four parts: 1) stochastic image augmentation; 2) encoder networks \( f(\cdot) \) extracting feature representations from augmented images; 3) multilayer perceptron (MLP) projection heads \( g(\cdot) \) that map the feature representations to a lower-dimensional space; and 4) contrastive loss computation. The MLP heads are important in achieving satisfactory results, as demonstrated in MoCo v2 (Chen et al., 2020b).

![Contrastive learning](image)

**Figure 2.** (a) MoCo (He et al., 2020); (b) SimCLR (Chen et al., 2020a).

### 2.3. Semi-supervised learning

Different from unsupervised learning that can work just on unlabeled data, semi-supervised learning (SSL) needs some labeled data and large-scale unlabeled data. Recent SSL methods for analyzing medical images can be divided into three groups: 1) consistency regularization based approach; 2) pseudo labeling based approach; 3) semi-supervised adversarial training.

Methods in the first category share one same idea that the prediction for an unlabeled example should not change significantly if some perturbations (e.g., adding noise, data augmentation) are applied. The loss functions of an SSL model generally consist of two parts. More concretely, given an unlabeled data example \( x \) and its perturbed version \( \tilde{x} \), the SSL model outputs logits \( f_\theta(x) \) and \( f_\theta(\tilde{x}) \). On the unlabeled data, the objective is to give consistent predictions by minimizing the mean squared error \( d(f_\theta(x), f_\theta(\tilde{x})) \), and this leads to the consistency (unsupervised) loss \( L_u \) on unlabeled data. On the labeled data, a cross entropy supervised loss \( L_s \) is computed. Example SSL models that are regularized by consistency constraints include Ladder Networks (Rasmus et al., 2015), II-Model (Laine and Aila, 2017), and Temporal Ensembling (Laine and Aila, 2017). A more recent example is the Mean Teacher paradigm (Tarvainen and Valpola, 2017a), composed of a teacher model and a student model (Figure 3). The student model is optimized by minimizing \( L_u \) on unlabeled data and \( L_s \) on labeled data; as an Exponential Moving Average (EMA) of the student model, the teacher model is used to guide
the student model for consistency training. Most recently, several works such as unsupervised data augmentation (UDA) (Xie et al., 2020) and MixMatch (Berthelot et al., 2019a) have brought SSL to a new level.

For pseudo labeling (Lee, 2013), an SSL model itself generates pseudo annotations for unlabeled examples; the pseudo-labeled examples are used jointly with labeled examples to train the SSL model. This process is iterated for several times, during which the quality of pseudo labels and the model’s performance get enhanced. The naïve pseudo-labeling process can be combined with MixUp augmentation (Zhang et al., 2018a) to further improve SSL model’s performance (Arazo et al., 2020). Pseudo labeling also works well with multi-view co-training (Qiao et al., 2018). For each view of the labeled examples, co-training learns a separate classifier, and then the classifier is used to generate pseudo labels for the unlabeled data; co-training maximizes the agreement of assigning pseudo annotations among each view of unlabeled examples.

The purpose of semi-supervised adversarial learning aims at solving the target task (e.g., classification) rather than generating high-fidelity samples. One simple way to adapt GAN to semi-supervised settings is by modifying the discriminator to perform additional tasks. For example, in the task of image classification, Salimans et al. (2016) and Odena (2016) changed the discriminator in DCGAN forcing it to serve as a classifier. For an unlabeled image, the discriminator functions as in the vanilla GAN, providing a probability of the input image being real; for a labeled image, the discriminator predicts its class besides generating a realness probability.

Figure 3. Mean Teacher model application in medical image analysis (Li et al., 2021a).

### 3. DEEP LEARNING APPLICATIONS

#### 3.1. Classification

Medical image classification is the goal of computer-aided diagnosis (CADx), which aims at either distinguishing malignant lesions from benign ones or identifying certain diseases from input images (Shen et al., 2017; van Ginneken et al., 2011). CADx-generated classification score serves as an objective “second opinion” for radiologists to make the final decisions regarding a disease (Doi, 2007; Doi, 2005). However, developing conventional CADx schemes can be labor-intensive and time-consuming due to the difficulty in automated lesion segmentation and determination of hand-crafted features.

Deep learning methods avoid lesion segmentation and handcrafted feature identification and computation. Thus, using deep learning has potential to more reliably and efficiently develop CADx schemes. Autoencoder and its variants are once prevalent in analyzing medical images because of the advantage of automatically exploring representative patterns (Shin et al., 2013; Zhang et al., 2016b; Suk et al., 2016; Chen et al., 2017a). For example, Cheng et al. (2016) exploited the stacked denoising auto-encoder (SDAE) (Vincent et al., 2010) to discover diverse representative features for the identification of ultrasound breast lesions and CT pulmonary nodules. The SDAE-based model was first pre-trained using resized ROIs with random noise corruptions and then fine-tuned with the original ROIs’ information, including aspect ratios and resizing factors. The SDAE-based CADx scheme exhibited a significant performance boost over conventional schemes.

However, the performance of deep learning models is highly dependent on the size of training dataset and quality of image annotations. In many medical image analysis tasks especially 3D scenarios, it can be challenging to establish a sufficiently large and high-quality training dataset because of difficulties in data acquisition and annotation (Tajbakhsh et al., 2016; Chen et al., 2019a). Classic data augmentation (e.g., rotation, scale, flip, translation, etc.) is simple but effective for performance boost (Simard et al., 2003; Krizhevsky et al., 2017; Ciregan et al., 2012; Li et al., 2013). In the following sections, we will introduce several common strategies to address the data scarcity issue and/or improve model performance.
in the medical domain, including transfer learning, self-supervised learning, and semi-supervised learning. In the end, we will introduce attention mechanisms that can be embedded into different models.

3.1.1. Supervised transfer learning and image synthesis

The supervised transfer learning technique (Tajbakhsh et al., 2016; Donahue et al., 2014) has been used to tackle the insufficient training data problem and improve model’s performance, where standard architectures (e.g., VGG (Simonyan and Zisserman, 2015), GoogleLeNet (Szegedy et al., 2015), ResNet (He et al., 2016), DenseNet (Huang et al., 2017), etc.) are first optimized in the source domain with a large amount of natural images (e.g., ImageNet (Deng et al., 2009)) or medical images, and then the pre-trained models are transferred to the target domain and fine-tuned using fewer training examples. Transfer learning has become a cornerstone for image classification tasks (de Bruijne, 2016), which has proven effective for a variety of modalities, including CT (Kumar et al., 2017; Shin et al., 2016), mammography (Huynh et al., 2016), MRI (Kumar et al., 2017), ultrasonography (Liu et al., 2017; Droste et al., 2019), microscopy (Kumar et al., 2017; Khan et al., 2019), optical endoscopy (Kumar et al., 2017; Tajbakhsh et al., 2016). Some researchers from the computer vision field have recently challenged the conventional practice of ImageNet pre-training for subsequent tasks (He et al., 2019a; Kornblith et al., 2019). For instance, He et al. (2019a) pointed out ImageNet pre-training, despite with faster convergence, does not necessarily guarantee higher accuracy than training from scratch. In the medical image field, Tajbakhsh et al. (2016) showed that pre-trained CNNs with adequate fine-tuning performed at least as well as CNNs trained from scratch, while Raghu et al. (2019) demonstrated that ImageNet pre-training sometimes hurt the performance of thoracic disease classification on a large chest X-ray dataset. Nonetheless, we believe transfer learning will remain as one of the most useful techniques in dealing with small medical datasets in the short term.

Given the limited amount of labeled images, GANs have been widely adopted to generate synthetic examples and improve model performance in the medical domain. Frid-Adar et al. (2018b) exploited DCGAN for synthesizing high-quality examples to improve liver lesion classification on a limited dataset. The dataset only has 182 liver lesions including cysts, metastases, and hemangiomas. Since training GAN typically needs a large number of examples, the authors applied classic data augmentation (e.g., rotation, flip, translation, scale) to create nearly 90,000 examples. The GAN-based synthetic data augmentation significantly improved the classification performance, with the sensitivity and specificity increased from 78.6% and 88.4% to 85.7% and 92.4% respectively. In their later work (Frid-Adar et al., 2018a), the authors further extended lesion synthesis from the unconditional setting (DCGAN) to a conditional setting (ACGAN). The generator of ACGAN was conditioned on the side information (lesion classes), and the discriminator predicted lesion classes besides synthesizing new examples. However, it was found that ACGAN-based synthetic augmentation delivered a weaker classification performance than its unconditional counterpart.

To alleviate data scarcity and especially the lack of positive cancer cases, Wu et al. (2018a) adopted a conditional structure (cGAN) to generate realistic lesions for mammogram classification. Traditional data augmentation was also used to create enough examples for training GAN. The generator, conditioned with malignant/non-malignant labels, can control the process of generating a specific type of lesions. For each non-malignant patch image, a malignant lesion was synthesized onto it using a segmentation mask of another malignant lesion; for each malignant image, its lesion was removed, and a non-malignant patch was synthesized. Although the GAN-based augmentation achieved better classification performance than traditional data augmentation, the improvement was relatively small, less than 1%.

Instead of synthetically generating examples, GAN can also be pre-trained on a large dataset related to the target task to learn generic features, and these features are fine-tuned on a small dataset to facilitate performing tasks such as classification (Shams et al., 2018). This also helps lessen the data scarcity issue in the medical imaging field. For example, Rubin et al. (2019) first pre-trained a DCGAN using thousands of unlabeled optical path delay (OPD) maps of sperm cells, and then they modified the discriminator by appending a few un-trained fully connected layers. The modified discriminator was finally trained on limited OPD maps of cancer/healthy cells, and it achieved better classification performance than models trained using transfer learning or classic augmentation. One key factor contributing to the success is the utilization of morphological similarity of OPD images from different biological cells. The adversarial training process of GAN can capture generic features from sperm cells, which are beneficial to cancer cells classification.

3.1.2. Self-supervised learning

Recent self-supervised learning approaches have shown great potential in improving performance of medical tasks lacking sufficient annotations (Bai et al., 2019; Tao et al., 2020; Li et al., 2020a; Shorfuzzaman and Hossain, 2021; Zhang et al., 2020a). This method is suitable to the scenario where large amounts of medical images are available, but only a small percentage are labeled. Accordingly, the model optimization is divided into two steps, namely, self-supervised pre-training and supervised fine-tuning. The model is initially optimized using unlabeled images to effectively learn good
features that are representative of the image semantics (Azizi et al., 2021). The pre-trained models from self-supervision are followed by supervised fine-tuning to achieve faster and better performance in subsequent classification tasks (Chen et al., 2020c). In practice, self-supervision can be created either through pretext tasks (Misra and Maaten, 2020) or contrastive learning (Jing and Tian, 2020) as follows.

Self-supervised pretext task based classification utilizes common pretext tasks such as rotation prediction (Tajbakhsh et al., 2019) and Rubik’s cube recovery (Zhuang et al., 2019; Zhu et al., 2020a). Chen et al. (2019b) argued that existing pretext tasks such as relative position prediction (Doersch et al., 2015) and local context prediction (Pathak et al., 2016) resulted in only marginal improvements on medical image datasets; the authors designed a new pretext task based on context restoration. This new pretext task has two steps: disordering patches in corrupted images and restoring the original images. The context restoration pre-training strategy improved the performance of medical image classification. Tajbakhsh et al. (2019) exploited three pretext tasks, namely, rotation (Gidaris et al., 2018), colorization (Larsson et al., 2017), and reconstruction (Arjovsky et al., 2017b), to pre-train models for classification tasks. After pre-training, models were trained using labeled examples. It was shown that pretext tasks based pre-training in the medical domain was more effective than random initializations and transfer learning (ImageNet pre-training) for diabetic retinopathy classification.

For self-supervised contrastive classification, Azizi et al. (2021) adopted the self-supervised learning framework SimCLR (Chen et al., 2020d) to train models (wider versions of ResNet-50 and ResNet-152) for dermatology condition classification and chest X-ray classification. They pre-trained the models by first using unlabeled natural images then with unlabeled dermatology images and chest X-rays. Feature representations were learned by maximizing agreement between positive image pairs that are either two augmented examples of the same image or multiple images from the same patient. The pre-trained models were fine-tuned using much fewer labeled dermatology images and chest X-rays. These models outperformed their counterparts pre-trained using ImageNet by 1.1% in mean AUC for chest X-ray classification and 6.7% in top-1 accuracy for dermatology condition classification. SimCLR pre-training also yielded a significant performance boost for histopathological image classification (Ciga et al., 2020). MoCo (He et al., 2020; Chen et al., 2020b) is another popular self-supervised learning framework to pre-train models for medical classification tasks, such as COVID-19 diagnosis from CT images (Chen et al., 2021) and pleural effusion identification in chest X-rays (Sowrirajan et al., 2020). Furthermore, it has been shown that self-supervised contrastive pre-training can greatly benefit from the incorporation of domain knowledge (Azizi et al., 2021; Vu et al., 2021). For example, Vu et al. (2021) harnessed patient metadata (patient number, image laterality, and study number) to construct and select positive pairs from multiple chest X-ray images for MoCo pre-training. With only 1% of the labeled data for pleural effusion classification, the proposed approach improved mean AUC by 3.4% and 14.4% compared to previous contrastive learning method (Sowrirajan et al., 2020) and ImageNet pre-training respectively.

3.1.3. Semi-supervised learning

Unlike self-supervised approaches that can learn useful feature representations just from unlabeled data, semi-supervised learning needs to integrate unlabeled data with labeled data through different ways to train models for a better performance. Madani et al. (2018a) employed GAN that was trained in a semi-supervised manner (Kingma et al., 2014) for cardiac disease classification in chest X-rays where labeled data was limited. Unlike the vanilla GAN (Goodfellow et al., 2014), this semi-supervised GAN was trained using both unlabeled and labeled data. Its discriminator was modified to predict not only the realism of input images but also image classes (normal/abnormal) for real data. When increasing the number of labeled examples, the semi-supervised GAN based classifier consistently performed better than supervised CNN. Semi-supervised GAN was also shown useful in other data-limited classification tasks, such as left ventricular hypertrophy classification from echocardiograms (Madani et al., 2018b), diabetic retinopathy classification from funduscopy images (Lecout et al., 2018), and CT lung nodule classification (Xie et al., 2019a). Besides the semi-supervised adversarial approach, consistency-based semi-supervised methods such as Π-Model (Laine and Aila, 2017) and Mean Teacher (Tarvainen and Valpola, 2017b) have also been used to leverage unlabeled medical image data for better classification (Shang et al., 2019; Liu et al., 2020a).

3.1.4. Attention mechanisms

Attention mechanisms took off in fields related to natural language processing, such as machine translation (Bahdanau et al., 2015; Vaswani et al., 2017; Luong et al., 2015) and image captioning (Xu et al., 2015; You et al., 2016; Anderson et al., 2018) and becomes popular in computer vision tasks, such as natural image classification (Wang et al.,
When processing images, attention modules can adaptively learn “what” and “where” to attend so that model predictions are conditioned on the most relevant image regions and features. In the context of medical image analysis, different types of attention modules have been used for performance boost and better model interpretability (Zhou et al., 2019a). Guan et al. (2018) introduced an attention-guided CNN, which is based on ResNet-50 (He et al., 2016). The attention heatmaps from the global X-ray image were used to suppress large irrelevant areas and highlight local regions that contain discriminative cues for the thorax disease. The proposed model effectively fused the global and local information and achieved a good classification performance. In another study, Schlemper et al. (2019) incorporated attention modules to a variant network of VGG (Baumgartner et al., 2017) and U-Net (Ronneberger et al., 2015) for 2D fetal ultrasound image plane classification and 3D CT pancreas segmentation, respectively. Each attention module was trained to focus on a subset of local structures in input images, and these local structures contain salient features useful to the target task.

### 3.2 Segmentation

Medical image segmentation, identifying the set of pixels or voxels of lesions, organs, and other substructures from background regions, is another challenging task in medical image analysis (Litjens et al., 2017). Among all common image analysis tasks such as classification and detection, segmentation needs the strongest supervision (large amounts of high-quality annotations) (Tajbakhsh et al., 2020). Since its introduction in 2015, U-Net (Ronneberger et al., 2015) has become probably the most well-known architecture for segmenting medical images; afterwards, different variants of U-Net such as V-Net (Milletari et al., 2016) and UNet++ (Zhou et al., 2018) have been proposed to further improve the segmentation performance. At the same time, a number of semi- and self-supervised learning based approaches have also been proposed to alleviate the need for large annotated datasets. Accordingly, in this section, we will first 1) review the original U-Net and its important variants, and summarize useful performance enhancing strategies; 2) introduce semi- and self-supervised learning based segmenting models.

#### 3.2.1 U-Net and its variants

In a convolutional network, the high-level coarse-grained features learned by higher layers capture semantics beneficial to the whole image classification; in contrast, the low-level fine-grained features learned by lower layers contain useful details for precise localizations (i.e., assigning a class label to each pixel) (Hariharan et al., 2015), which is important to perform image segmentation. U-Net is built on the fully convolutional network (Long et al., 2015), and a basic convolutional unit is the forward convolutional layer as shown in Figure 5(a). Like the encoder-decoder architecture, U-Net consists of a down-sampling subnetwork with convolutional layers and a symmetric up-sampling subnetwork with deconvolutional layers; these two subnetworks are connected to form the U-shaped network (Figure 4). The key innovation of U-Net is the so-called skip connections between opposing convolutional layers and deconvolutional layers, which successfully concatenate features learned at different levels to improve the segmentation performance. Meanwhile, skip connections is also helpful in recovering the network’s output to be of the same spatial resolution as the input. U-Net takes 2D images as input, and it generates several segmentation maps, each of which corresponds to one respective pixel class.

*Figure 4. The structure of U-Net (Ronneberger et al., 2015)*
Based on the basic architectures, Drozdzal et al. (2016) further studied the influence of long and short skip connections in biomedical image segmentation. They concluded that adding short skip connections is important to train very deep segmentation networks. In one study, Zhou et al. (2018) claimed that the plain skip connections between U-Net’s encoder and decoder subnetworks leads to fusion of semantically dissimilar feature maps; they proposed to reduce the semantic gap prior to fusing feature maps. In the proposed model UNet++, the plain skip connections were replaced by nested and dense skip connections. The suggested architecture outperformed U-Net and wide U-Net across four different medical image segmentation tasks.

Figure 5. Units of different segmentation networks (a) forward convolutional unit (U-Net), (b) recurrent convolutional block (RCNN), (c) residual convolutional unit (residual U-Net), and (d) recurrent residual convolutional unit (R2U-Net) (Alom et al., 2018).

Aside from redesigning the skip connections, Çiçek et al. (2016) replaced all 2D operations with their 3D counterparts to extend the 2D U-Net to 3D U-Net for volumetric segmentation with sparsely annotated images. Further, Milletari et al. (2016) proposed V-Net for 3D MRI prostate volumes segmentation. A major architecture difference between U-Net and V-Net lies in the change of the forward convolutional units (Figure 5(a)) to residual convolutional units (Figure 5(c)), so V-Net is also referred as residual U-Net. A new loss function based on Dice coefficient was proposed to deal with the imbalanced number of foreground and background voxels. To tackle the scarcity of annotated volumes, the authors augmented their training dataset with random non-linear transformations and histogram matching. Gibson et al. (2018) proposed the Dense V-network that modified V-Net’s loss function of binary segmentation to support multiorgan segmentation of abdominal CT images. Although the authors followed the V-Net architecture, they replaced its relatively shallow down-sampling network with a sequence of three dense feature stacks. The combination of densely linked layers and the shallow V-Net architecture demonstrates its importance in improving segmentation accuracy, and the proposed model yielded significantly higher Dice scores for all organs compared to multi-atlas label fusion (MALF) methods.

Alom et al. (2018) proposed to integrate the architectural advantages of recurrent convolutional neural network (RCNN) (Ming and Xiaolin, 2015) and residual network (ResNet) (He et al., 2016) when designing U-Net based segmentation networks. In their first network (RU-Net), the authors replaced U-Net’s forward convolutional units using RCNN’s recurrent convolutional layers (RCL) (Figure 5(b)), which can help accumulate useful features to improve segmentation results. In their second network (R2U-Net), the authors further modified RCL using ResNet’s residual units (Figure 5(d)), which learns a residual function by using identity mapping for shortcut connections, thus allowing for training very deep networks. Both models achieved better segmentation performance than U-Net and residual U-Net. Dense convolutional blocks (Huang et al., 2017) also demonstrated its superiority in enhancing segmentation performance on liver and tumor CT volumes (Li et al., 2018).

Besides the redesigned skip connections and modified architectures, U-Net based segmentation approaches also benefit from adversarial training (Xue et al., 2018; Zhang et al., 2020b), attention mechanism (Jetley et al., 2018; Anderson et al., 2018; Oktay et al., 2018; Nie et al., 2018; Sinha and Dolz, 2021), and uncertainty estimation (Wang et al., 2019; Yu et al., 2019; Baumgartner et al., 2019; Mehrtash et al., 2020). For example, Xue et al. (2018) developed an adversarial network for brain tumor segmentation, and the network has two parts: a segmentor and a critic. The segmentor is a U-Net-like network that generates segmentation maps given input images; the predicted maps and ground-truth segmentation maps are sent into the critic network. Alternatively training these two components eventually led to good segmentation results. Oktay et al. (2018) proposed incorporating attention gates (AGs) into the U-Net architecture to suppress irrelevant
features from background regions and highlight important salient features that are propagated through the skip connections. Attention U-Net consistently delivered a better performance than U-Net in CT pancreas segmentation. Baumgartner et al. (2019) developed a hierarchical probabilistic model to estimate uncertainties in the segmentation of prostate MR and thoracic CT images. The authors employed variational autoencoders to infer the uncertainties or ambiguities in expert annotations, and separate latent variables were used to model segmentation variations at different resolutions.

3.2.2 Semi- and self-supervised learning based segmenting models

For medical image segmentation, to alleviate the need for a large amount of annotated training data, researchers have adopted generative models for image synthesis to increase the number of training examples (Zhang et al., 2018b; Zhao et al., 2019a). In contrast with image synthesis, exploiting the power of unlabeled medical images seems like a much more popular choice. Although it is difficult and expensive to acquire high-quality annotations, unlabeled medical images are often available, usually coming with a large number. Given a small medical image dataset with limited ground truth annotations and a related but unlabeled large dataset, researchers have explored self-supervised and semi-supervised learning approaches to learn useful and transferrable feature representations from the unlabeled dataset, which will facilitate solving downstream tasks like segmentation with limited annotations. There exist roughly 4 different ways to leverage unlabeled dataset to improve the segmentation performance: semi-supervised learning based 1) consistency regularization and 2) pseudo labeling, and self-supervised 3) pretext tasks and 4) contrastive learning.

Semi-supervised consistency regularization: The mean teacher model is commonly used. Based on the mean teacher framework (Tarvainen and Valpola, 2017b), Yu et al. (2019) introduced uncertainty estimation (Kendall and Gal, 2017) for better segmentation of 3D left atrium from MR images. They argued that on an unlabeled dataset, the output of the teacher model can be noisy and unreliable; therefore, besides generating target outputs, the teacher model was modified to estimate these outputs’ uncertainty. The uncertainty-aware teacher model can produce more reliable guidance for the student model, and the student model could in turn improve the teacher model. The mean teacher model can also be improved by the transformation-consistent strategy (Li et al., 2020b). In one study, Wang et al. (2020a) proposed a semi-supervised framework to segment COVID-19 pneumonia lesions from CT scans with noisy labels. Their framework is also based on the Mean Teacher model; instead of updating the teacher model with a predefined value, they adaptively updated the the teacher model using a dynamic threshold for the student model’s segmentation loss. Similarly, the student model was also adaptively by the teacher model. To simultaneously deal with noisy labels and the foreground-background imbalance, the authors developed a generalized version of the Dice loss in (Milletari et al., 2016). The authors designed the segmentation network in the same spirit of U-Net but made several changes in terms of new skip connections (Pang et al., 2019), multi-scale feature representation (Chen et al., 2018a), etc. In the end, the segmentation network with the Dice loss were combined with the mean teacher framework. The proposed method demonstrated high robustness to label noise and achieved better performance for pneumonia lesion segmentation than other state-of-the-art methods.

Semi-supervised pseudo labeling: Fan et al. (2020) presented a semi-supervised framework (Semi-InfNet) to tackle the lack of high-quality labeled data in COVID-19 lung infection segmentation from CT images. To generate pseudo labels for the unlabeled images, they first used 50 labeled CT images to train their model, which produced pseudo labels for a small amount of unlabeled images. Then the newly pseudo-labeled examples were included in the original labeled training dataset to re-train the model to generate pseudo labels for another batch of unlabeled images. This process was repeated until 1600 unlabeled CT images all got pseudo-labeled. Both the labeled and pseudo-labeled examples were used to train Semi-InfNet, and its performance surpassed other cutting-edge segmentation models such as UNet++ by a large margin. Aside from the semi-supervised learning strategy, there are three critical components in the model responsible for the good performance: parallel partial decoder (PPD) (Wu et al., 2019a), reverse attention (RA) (Chen et al., 2018b), and edge attention (Zhang et al., 2019). PPD can aggregate high-level features of the input image and generate a global map indicating the rough location of lung infection regions; EA module uses low-level features to model boundary details, and RA module further refines the rough estimation into an accurate segmentation map.

In addition to the aforementioned improvement, researchers have also explored incorporating domain-specific prior knowledge to tailor the semi-supervised frameworks for a better segmentation performance. The prior knowledge varies a lot, such as the anatomical prior (He et al., 2019b), atlas prior (Zheng et al., 2019), topological prior (Clough et al., 2020), semantic constraint (Ganaye et al., 2018), and shape constraint (Li et al., 2020c) to name a few.

Self-supervised pretext tasks: Since self-supervision via pretext tasks and contrastive learning can learn rich semantic representations from unlabeled datasets, self-supervised learning is often used to pre-train the model and enable solving downstream tasks (e.g., medical image segmentation) more accurately and efficiently when limited annotated
examples are available (Taleb et al., 2020). The pretext tasks could be either designed based on application scenarios or chosen from traditional ones used in the computer vision field. For the former type, Bai et al. (2019) designed a novel pretext task by predicting anatomical positions for cardiac MR image segmentation. The self-learnt features via the pretext task were transferred to tackle a more challenging task, accurate ventricles segmentation. The proposed method achieved much higher segmentation accuracy than the standard U-Net trained from scratch, especially when only limited annotations were available.

For the latter type, Taleb et al. (2020) extended performing pretext tasks from 2D to 3D scenarios, and they investigated the effectiveness of several pretext tasks (e.g., rotation prediction, jigsaw puzzles, relative patch location) in 3D medical image segmentation. For brain tumor segmentation, they adopted the U-Net architecture, and the pretext tasks were performed on a large unlabeled dataset (about 22,000 MRI scans) to pre-train the models; then the learned feature representations were fine-tuned on a much smaller labeled dataset (285 MRI scans). The 3D pretext tasks performed better than their 2D counterparts; more importantly, the proposed methods sometimes outperformed supervised pre-training, suggesting a good generalization ability of the self-learnt features.

The performance of self-supervised pre-training could also be improved by adding other types of information. Hu et al. (2020) implemented a context encoder (Pathak et al., 2016) performing semantic inpainting as the pretext task, and they incorporated DICOM metadata from ultrasound images as weak labels to boost the quality of pre-trained features toward facilitating two different segmentation tasks.

For self-supervised contrastive learning based segmentation approaches, early works such as (Jamaludin et al., 2017) adopted the original contrastive loss (Chopra et al., 2005) to learn useful feature representations. In recent three years, with a surge of interest in self-supervised contrastive learning, contrastive loss has evolved from the original version to more powerful ones (Oord et al., 2019) for learning expressive feature representations from unlabeled datasets. Chaitanya et al. (2020) claimed although the contrastive loss in (Chen et al., 2020a) was suitable for learning image-level (global) feature representations, it did not guarantee learning distinctive local representations that are important for per-pixel segmentation. They proposed a local contrastive loss to capture local features that can provide complementary information to boost the segmentation performance. Meanwhile, to the best of our knowledge, when computing the global contrastive loss, these authors are the first to utilize the domain knowledge that there is structural similarity in volumetric medical images (e.g., CT and MRI). In MRI image segmentation with low annotations, the proposed method substantially outperformed other semi-supervised and self-supervised methods. In addition, it was shown that the proposed method could further benefit from data augmentation techniques like MixUp (Zhang et al., 2018c).

| Author                  | Year | Application                                                                 | Model             | Dataset                            | Contributions highlights                                                                 |
|-------------------------|------|-----------------------------------------------------------------------------|-------------------|------------------------------------|------------------------------------------------------------------------------------------|
| Ronneberger et al., 2015 | 2015 | Segmentation of (1) neuronal structures in electron microscopic stacks and (2) cells in light microscopy images | U-Net             | EM ISBI 2012 dataset and ISBI cell tracking challenge dataset 2015 | (1) Skip connections between opposing convolutional layers and deconvolutional layers concatenate features learned at different levels; (2) Overlap-tile strategy; (3) Applying elastic deformations to training images to learn invariance; (4) A weighted loss for separating touching objects of the same class. |
| Drozdzel et al., 2016   | 2016 | Segmentation of neuronal structures in electron microscopic stacks            | Residual Network with skip connections | EM ISBI 2012 dataset                | (1) Using very deep network for segmentation without any post-processing; (2) Showing long and short skip connections are important for the convergence of deep networks. |
| Çeçek et al., 2016       | 2016 | Segmentation of the Xenopus kidney structures from confocal microscopic data | 3D U-Net          | Private dataset                    | (1) Extending the original 2D U-Net to its 3D counterpart; (2) A weighted softmax loss function and special data augmentation by applying a smooth dense deformation field enable the model to learn from sparse annotations. |
| Milletari et al., 2016  | 2016 | MRI prostate volumes segmentation                                            | V-Net (Residual U-Net) | PROMISE 2012 dataset               | (1) Incorporation of residual learning; (2) A new loss function based on Dice coefficient to deal with class imbalance; (3) Data augmentation by applying random non-linear transformations and histogram matching. |
| Authors                  | Year | Task                                                                 | Model/Architectures                                                                 | Notes                                                                                                                                 |
|-------------------------|------|----------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|
| Zhou et al., 2018       | 2018 | Segmentation of (1) CT lung nodules, (2) microscopic cell nuclei, (3) CT liver, and (4) colon polyps | UNet++                                                                             | (1) Proposing nested and dense skip connections to reduce the semantic gap before fusing feature maps; (2) Using deep supervision to enable accurate and fast segmentation. |
| Gibson et al., 2018     | 2018 | Segmentation of CT pancreas, gastrointestinal organs, and surrounding organs | Dense V-Net                                                                        | (1) A new loss function extends binary segmentation to multiorgan segmentation; (2) Integrating densely linked layers into the shallow V-Net architecture. |
| Alom et al., 2018       | 2018 | Segmentation of (1) retina blood vessels, (2) skin cancer lesions, and (3) lung | RU-Net and R2U-Net                                                                 | (1) Replacing U-Net’s forward convolutional units using RCNN’s recurrent convolutional layers to accumulate useful features; (2) Incorporating residual learning to train very deep networks. |
| Oktay et al., 2018      | 2018 | Multi-class CT segmentation of pancreas, spleen, and kidney          | Attention U-Net                                                                   | (1) Incorporating attention gates into the U-Net architecture to learn important salient features and suppress irrelevant features; (2) Image grid-based gating improves attention to local regions. |
| Xue et al., 2018        | 2018 | Brain tumor segmentation from MRI volumes                            | SegAN: adversarial network with a segmentor and a critic                           | (1) Using adversarial learning for segmentation; (2) Proposing a multi-scale $L_1$ loss function to facilitate learning local and global features. |
| Zhang et al., 2018b      | 2018 | Segmentation of multimodal cardiovascular images (CT and MRI)        | Modified GAN and a U-Net based segmentor                                           | (1) Training GAN by adding a cycle-consistency loss and a shape consistency loss, making the segmentor and the generator benefit from each other; (2) Updating the generator in an online manner. |
| Zhao et al., 2019a       | 2019 | Segmentation of brain MRI scans                                      | U-Net based networks and a SD-Net (Roy et al., 2017) based architecture            | Novel data augmentation (i.e., learning complex spatial and appearance transformations to synthesize additional labeled images based on limited labeled examples). |
| Baumgartner et al., 2019 | 2019 | Segmentation of prostate MR and thoracic CT images                    | PHiSeg: a probabilistic U-Net architecture                                          | (1) Applying conditional VAE for inference in the U-Net architecture; (2) Using a separate latent variable to control segmentation at each resolution level to hierarchically generate final segmentations. |
| **Semi-supervised segmentation** |       |                                                                      |                                                                                    |                                                                                                                                 |
| Yu et al., 2019         | 2019 | Segmentation of left atrium from 3D MR scans                          | UA-MT: Mean Teacher framework with V-Net as backbone                              | Enforcing the teacher model to provide more reliable guidance to the student model via uncertainty estimation, where the estimated uncertainty was used to filter out highly uncertain predictions. |
| Li et al., 2020b        | 2020 | Segmentation of (1) skin lesions, (2) fundus optic disks, and (3) CT liver | TCSM_v2: Mean Teacher framework with U-Net-like network as backbone                | Imposing transformation-consistent regularizations to unlabeled images to enhance the network’s generalization capacity. |
| Wang et al., 2020a       | 2020 | Segmentation of COVID-19 pneumonia lesions from CT scans             | COFLE-Net: Mean Teacher framework with U-Net                                         | (1) Adaptively updating the teacher model and the student model; (2) Developing a generalized Dice loss to deal with noisy labels and foreground-background imbalance; (3) |
like network as backbone | Using new skip connections and multi-scale feature representation.
--- | ---
Fan et al., 2020 | 2020 | Segmentation of COVID-19 lung infection from CT images | Semi-InfNet: Inf-Net trained in a semi-supervised manner | 2 publicly available CT datasets of COVID-19 | (1) Iterative generation of pseudo labels for unlabeled images; (2) Using the parallel partial decoder to generate a rough infection map, and reverse and edge attention modules to refine the segmentation map. (3) Multi-scale training strategy (Wu et al., 2019b).

**Self-supervised segmentation**

Bai et al., 2019 | 2019 | Cardiac MR image segmentation | Self-supervised U-Net | UK Biobank (UKB) | (1) Pre-training the network using a new pretext tasks (i.e., predicting anatomical positions) where meaningful features were learned via self-supervision; (2) Comparing three different ways for supervised fine-tuning.

Taleb et al., 2020 | 2020 | Segmentation of (1) brain tumor from MRI and (2) pancreas tumor from CT | Self-supervised 3D U-Net | (1) UKB and BraTS 2018, (2) part of medical decathlon benchmarks | (1) Extending traditional 2D pretext tasks to 3D, utilizing the 3D spatial context for better self-supervision; (2) A comprehensive comparison of the performance of five different pretext tasks.

Hu et al., 2020 | 2020 | Segmentation of (1) thyroid nodule and (2) liver/kidney from ultrasound images | Self-supervised U-Net with VGG16 or ResNet50 as backbone | (1) DDTI ultrasound dataset and (2) a private dataset | Incorporating DICOM metadata from ultrasound images as weak labels to improve the quality of pre-trained features from the pretext task.

Chaitanya et al., 2020 | 2020 | Segmentation of MRI cardiac structures and prostate regions | U-Net based encoder and decoder architecture | (1) MICCAI 2017 ACDC, (2) Medical Segmentation Decathlon, (3) STACOM 2017 | (1) Proposing a local contrastive loss; (2) Incorporation of domain knowledge (structural similarity in volumetric) in contrastive loss calculation; (3) A comprehensive comparison of a variety of pre-training techniques, such as self-supervised contrastive and pretext task pre-training, etc.

### 3.3. Detection

A natural image may contain objects belonging to different categories, and each object category may contain several instances. In the computer vision field, object detection algorithms are applied to detect and identify if any instance(s) from certain object categories are present in the image (Seraman et al., 2014; Girshick et al., 2014; Russakovsky et al., 2015). Previous works (Shen et al., 2017; Litjens et al., 2017) have reviewed the successful applications of the frameworks before 2015, such as OverFeat (Seraman et al., 2014; Ciompi et al., 2015), RCNN (Girshick et al., 2014), and fully convolutional networks (FCN) based models (Long et al., 2015; Dou et al., 2016; Wolterink et al., 2016). As a comparison, we aim at summarizing applications of more recent object detection frameworks (since 2015), such as Faster RCNN (Ren et al., 2015), YOLO (Redmon et al., 2016), and RetinaNet (Lin et al., 2017). In this section, we will first briefly review several recent milestone detection frameworks (Liu et al., 2020b), and then describe their applications in medical detection tasks.

### 3.3.1 Overview of the detection frameworks

RCNN framework (Girshick et al., 2014) is a multi-stage pipeline. Despite its impressive results in object detection, RCNN has some drawbacks namely, the multistage pipeline makes training slow and difficult to optimize; separately extracting features for each region proposal makes training expensive in disk space and time, and it also slows down testing (Girshick, 2015). These drawbacks have inspired several recent milestone detectors, and they can be categorized into two groups (Liu et al., 2020b): (1) two-stage detection frameworks (Girshick, 2015; Ren et al., 2015; Ren et al., 2017; Dai et al., 2016), which include a separate module to generate region proposals before bounding box recognition (predicting class probabilities and bounding box coordinates); (2) one-stage detection frameworks (Redmon et al., 2016; Redmon and Farhadi, 2017; Liu et al., 2016; Lin et al., 2017; Law and Deng, 2020; Duan et al., 2019) which predict bounding boxes in a unified manner without separating the process of generating region proposals. In an image,
region proposals are a collection of potential regions or candidate bounding boxes that are likely to contain an object (Liu et al., 2020b).

**Two-stage detectors:** Unlike RCNN, the Fast RCNN framework (Girshick, 2015) is an end-to-end detection pipeline employing a multi-task loss to jointly classify region proposals and regress bounding boxes. Region proposals in Fast RCNN are generated on a shared convolutional feature map rather than the original image to speed up computation. Then a Region of Interest pooling layer was applied to warp all the region proposals into the same size. The adjustments resulted in a better and faster detection performance but the speed of Fast RCNN is still bottlenecked by the inefficient process of computing region proposals. In the Faster RCNN framework (Ren et al., 2015; Ren et al., 2017), a Region Proposal Network (RPN) replaced the selective search method to produce high-quality region proposals from anchor boxes efficiently. Anchor boxes are a set of pre-determined candidate boxes of different sizes and aspect ratios to capture objects of specific classes (Ren et al., 2015). Since that time, anchor boxes have played a dominant role in top-ranked detection frameworks. Mask RCNN (He et al., 2017) is closely related to Faster RCNN but it was originally designed for pixelwise object instance segmentation. Mask RCNN also has a RPN to propose candidate object bounding boxes; this new framework extends Faster RCNN by adding an extra branch that outputs a binary object mask to the existing branch of predicting classes and bounding box offsets. Mask RCNN uses a Feature Pyramid Network (FPN) (Setio et al., 2017) as its backbone to extract features at various resolution scales. Besides instance segmentation, Mask RCNN can be used for object detection, achieving excellent accuracy and speed.

**One-stage detectors** Redmon et al. (2016) proposed a single-stage framework YOLO; instead of using a separate network to generate region proposals, they treated object detection as a simple regression problem. A single network was used to directly predict object classes and bounding box coordinates. YOLO also differs from region proposal based frameworks (e.g., Faster CNN) in that it learns features globally from the entire image rather than from local regions. Despite being faster and simpler, YOLO has more localization errors and lower detection accuracy than Faster RCNN. Later the authors proposed YOLOv2 and YOLO9000 (Redmon and Farhadi, 2017) to improve the performance by integrating different techniques, including batch normalization, using good anchor boxes, fine-grained features, multi-scale training, etc. Lin et al. (2017) identified that the central cause for the lagging performance of one-stage detectors is the imbalance between foreground and background classes (i.e., the training process was dominated by vast numbers of easy examples from the background). To deal with the class imbalance problem, they proposed a new focal loss that can weaken the influence of easy examples and enhance the contribution of hard examples. The proposed framework (RetinaNet) demonstrated higher detection accuracy than state-of-the-art two-stage detectors at that time. Law and Deng (2020) proposed CornerNet and pointed out that the prevalent use of anchor boxes in object detection frameworks, especially one-stage detectors, causes issues such as the extreme imbalance between positive and negative examples, slow training, introducing extra hyperparameters, etc. Instead of designing a set of anchor boxes to detect bounding boxes, the authors formulated bounding boxes detection as detecting a pair of key-points (top-left and bottom-right corners) (Newell et al., 2017; Tychsen-Smith and Petersson, 2017). Nonetheless, CornerNet generates a large number of incorrect bounding boxes since it cannot fully utilize the recognizable information inside the cropped regions (Duan et al., 2019). Based on CornerNet, Duan et al. (2019) proposed CenterNet that detects each object using a triplet of key-points, including a pair corners and one center key-point. Unlike CornerNet, CenterNet can extract more recognizable visual patterns within each proposed region, thus effectively suppress inaccurate bounding boxes (Duan et al., 2019).

3.3.2. Specific-type medical object (e.g., lesion) detection

Common computer-aided detection (CADE) tasks include detecting lung nodules (Gu et al., 2018; Xie et al., 2019b), breast masses (Akselrod-Ballin et al., 2017; Ribli et al., 2018), lymph nodes (Lin et al., 2019; Zhu et al., 2020b), sclerosis lesions (Nair et al., 2020), etc. The general detection frameworks, originally designed for general object detection in natural images, cannot guarantee satisfactory performance for lesion detection in medical images for two main reasons: (1) lesions can be extremely small in size compared to natural objects; (2) lesions and non-lesions often have similar appearances (e.g. texture and intensity) (Tao et al., 2019; Tang et al., 2019). To deliver good detection performance in the medical domain, these frameworks need to be adjusted through different methods, such as incorporating domain-specific characteristics, uncertainty estimation, or semi-supervised learning strategy, which are presented as follows.

Incorporating domain-specific characteristics has been a popular choice in both the radiology and histopathology domains. In the radiology domain, the intrinsic 3D spatial context information among volumetric images (e.g. CT scans) has been utilized in many studies (Roth et al., 2016; Dou et al., 2017; Yan et al., 2018a; Liao et al., 2019). For example, in the task of pulmonary nodule detection, Ding et al. (2017) argued that the original Faster RCNN (Ren et al., 2015) with the VGG-16 network (Liu and Deng, 2015) as its backbone cannot capture representative features of small pulmonary
nODULES; they introduced a deconvolutional layer at the end of Faster RCNN to recover fine-grained features that are important in detecting small objects. On the deconvolutional feature map, an FPN was applied to propose candidate regions of nodules from 2D axial slices. To reduce the false positive rate, the authors proposed to make the classification network see the full range of contexts of the nodule candidates. Instead of using 2D CNN, they chose a 3D CNN to exploit the 3D context of candidate regions so that more distinctive features can be captured for nodule recognition. The proposed method ranked the 1st place in nodule detection on the LUNA16 benchmark dataset (Setio et al., 2017). Zhu et al. (2018a) also considered the 3D nature of lung CT images and designed a 3D Faster RCNN for nodule detection. To efficiently learn nodule features, the 3D faster RCNN had the U-Net-like structure (Ronneberger et al., 2015) and was built with compact dual path blocks (Chen et al., 2017b). It should be noted that despite the effectiveness in boosting detection performance, compared to 2D CNN, 3D CNN has downsides including consuming more computational resources and requiring more efforts to acquire 3D bounding box annotations (Yan et al., 2018a; Tao et al., 2019). In a recent study, Mei et al. (2021) established a large dataset (PN9) with more than 40,000 annotated lung nodules to train 3D CNN-based models. The authors improved the model’s ability to detect both large and small lung nodules by utilizing correlations that exist among multiple consecutive CT slices. Given a slice group, a non-local operation based module (Wang et al., 2018a) was employed to seize long-range dependencies of different positions and different channels in the feature map. Furthermore, since each shallow ResNet block can generate feature maps on the same scale that carry useful spatial information, the authors reduced false positive nodule candidates by merging multi-scale features produced by 3 different blocks.

In the histopathology domain, Rijthoven et al. (2018) presented a modified version of YOLOv2 (Redmon and Farhadi, 2017) for lymphocytes detection in whole-slide images (WSI). Based on the prior knowledge of lymphocytes (e.g., average size, no overlaps), the authors simplified the original YOLO network with 23 layers by keeping only a few layers. With the prior knowledge that brown areas without lymphocytes in the WSI contain many hard negative samples, the authors also designed a sampling strategy to enforce the detection model to focus on these hard negative examples during training. The proposed method improved F1-score by 3% with a speed-up of 4.3X. In their later work, Swiderska-Chadaj et al. (2019) modified the YOLO architecture to further detect lymphocytes in a more diversified WSI dataset of breast, prostate, and colon cancer; however, it did not perform as well as the U-Net based detection architecture, which first classified each pixel and then produced detection results using post-processing techniques. The modified YOLO architecture was also shown the least robust to different staining. Readers interested in a comprehensive review of how to integrate medical domain knowledge into network designing can refer to (Xie et al., 2021).

Recently, semi-supervised methods have been used to improve the performance of medical object detection (Gao et al., 2020; Qi et al., 2020). For example, Wang et al. (2020b) developed a generalized version of the original focal loss (Lin et al., 2017) to deal with soft labels in computing semi-supervised loss function. They modified the semi-supervised approach MixMatch (Berthelot et al., 2019a) from two aspects to make it suitable for 3D medical image detection. An FPN (Setio et al., 2017; Liu et al., 2019) was first applied on unlabeled CT images (without lesion annotations) to generate pseudo-labeled object instances. Then the pseudo-labeled examples were mixed with examples having ground truth annotations through MixUp augmentation. However, the original MixUp augmentation (Zhang et al., 2018a) was designed for classification tasks where labels are image classes; the authors adapted this augmentation technique to the lesion detection task with annotations in the form of bounding boxes. The semi-supervised approach demonstrated a significant performance gain over supervised learning baselines in pulmonary nodule detection.

In addition, uncertainty estimation is another useful technique to facilitate the detection of small objects (Ozdemir et al., 2017; Nair et al., 2020). For example, in the task of multiple sclerosis lesion detection where uncertainties mostly result from small lesions and lesion boundaries, Nair et al. (2020) explored using uncertainty estimates to improve detection performance. Specifically, four uncertainty measures were computed: a predicted variance from training data (Kendall and Gal, 2017), variance of Monte Carlo (MC) samples, a predictive entropy, and mutual information. A threshold formed by these measures was used to filter out the most uncertain lesion candidates and thus improve detection performance.

### 3.3.3 Universal lesion detection

Traditional lesion detectors have focused on a specific type of lesions but there is a rising research interest in identifying and localizing different kinds of lesions from the whole human body at once (Yan et al., 2018a; Yan et al., 2019; Tao et al., 2019; Yan et al., 2020; Cai et al., 2020; Li et al., 2020d). DeepLesion is a large and comprehensive dataset (32K lesions) that contains a variety of lesion types such as lung nodule, liver tumor, abdominal mass, pelvic mass, etc. (Yan et al., 2018b; Yan et al., 2018c). Tang et al. (2019) proposed ULDor based on Mask R-CNN for universal lesion detection. Training Mask-RCNN requires ground truth masks for lesions; however, the DeepLesion dataset does not contain such annotated masks. With the RECIST (Response Evaluation Criteria In Solid Tumors) annotations (Eisenhauer
et al., 2009), the authors estimated real masks via ellipse fitting for each lesion region. In addition, hard negative examples were used to re-train the model to reduce false positives. Yan et al. (2019) further improved the performance of universal lesion detection by enforcing a multitask detector (MULAN) to jointly perform lesion detection, tagging, and segmentation. It was previously shown that combining different tasks may provide complementary information to each other and thus enhance the performance of a single task (Wu et al., 2018b; Tang et al., 2019). MULAN is modified from Mask RCNN (He et al., 2017) with three head branches. The detection branch predicts whether each proposed region is lesion and regresses bounding boxes; the tagging branch predicts 185 tags (e.g., body part, lesion type, intensity, shape, etc.) for each lesion proposal; the segmentation branch outputs a binary mask (lesion/non-lesion) for each proposed region. MULAN significantly surpassed previous lesion detection models such as ULDor (Tang et al., 2019) and 3DCE (Yan et al., 2018a). Furthermore, Yan et al. (2020) have recently shown that learning from heterogenous lesion datasets and partial labels can also boost detection performance.

In addition to the above strategies, attention mechanism is another useful way to improve lesion detection. Tao et al. (2019) trained a universal lesion detector on the DeepLesion dataset, and the attention mechanism (Wang et al., 2017; Woo et al., 2018b) was introduced to incorporate 3D context and spatial information into a R-FCN based detection architecture (Dai et al., 2016). A contextual attention module outputs a vector indicating the importance of features learned from different axial CT slices, so the detection framework can adaptively aggregate features from different slices (i.e., enhancing relevant contextual features); a spatial attention module outputs a weight matrix so that discriminative regions on feature maps can be amplified, through which richer and more representative features can be well learned for small lesions. The proposed method demonstrated a significant performance improvement despite using much less fewer slices. Li et al. (2019) presented an FPN based architecture (Setio et al., 2017) with an attention module that can incorporate clinical knowledge. In clinical practice, it is common for radiologists to inspect multiple CT windows for an accurate lesion diagnosis. The authors first employed three FPNs to generate feature maps from three frequently inspected windows; then the attention module (Woo et al., 2018b) was used to reweight feature maps from different windows. The prior knowledge about lesion positions was also incorporated to further improve the performance.

We observe that, whether in the detection of specific-type of lesions or universal lesions, two-stage detectors are still quite prevalent for their high performance and robustness; however, separately generating region proposals might hinder developing streamline CADe schemes. Several very recent studies have demonstrated that good detection performance can also be obtained by one-stage detectors (Pisov et al., 2020; Zhu et al., 2021; Lung et al., 2021). We predict that advanced anchor-free one-stage detectors (e.g., CenterNet (Duan et al., 2019), PAFNet (Xin et al., 2021), etc.) if adjusted properly to accommodate the uniqueness of medical images, will attract much more attention and even become a better choice than two-stage detectors for developing new CADe schemes in the long run.

Table 2. A list of recent papers related to medical image detection

| Author          | Year | Application                          | Model                                      | Dataset          | Contributions highlights                                                                                                                                 |
|-----------------|------|--------------------------------------|--------------------------------------------|------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|
| Ding et al., 2017 | 2017 | Lung nodules detection from CT images | Faster RCNN with changed VGG16 as backbone | LUNA16           | (1) Using deconvolutional layer to recover fine-grained features; (2) Using 3D CNN to exploit 3D spatial context information for false positives reduction. |
| Zhu et al., 2018 | 2018 | Lung nodules detection from CT images | 3D Faster RCNN with U-Net-like structure, built with dual path blocks | LIDC-DRIs        | (1) Using 3D Faster RCNN considering the 3D nature of lung CT images; (2) Utilizing the compactness (i.e., fewer parameters) of dual path networks on small dataset. |
| Wang et al., 2020b | 2020 | Lung nodules detection from CT images | 3D variant of FPN with modified residual network as backbone | LUNA16 and NLST  | (1) A semi-supervised learning strategy to leverage unlabeled images in NLST; (2) MixUp augmentation for examples with pseudo labels and ground truth annotations; (3) FPN outputs multi-level features to enhance small object detection. |
| Mei et al., 2021 | 2021 | Lung nodules detection from CT images | U-shaped architecture, with 3D ResNet50 as encoder | PN9              | (1) Inserting non-local modules in residual blocks to seize long-range dependencies of different positions and different channels. (2) Using multi-scale features for false positives reduction. |
| Authors          | Year | Task Description                                                                 | Backbone/Network Description                                                                 | Dataset(s)                                                                                                                                                                                                 | Extraction/Annotation/Annotation Reduction/Pseudo Annotation Methods |
|------------------|------|-----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------|
| Ma et al., 2021  | 2020 | Breast mass detection from mammograms                                             | CVR-RCNN: Two-branch Faster RCNNs, with relation modules (Hu et al., 2018a)                  | DDSM and a private dataset                                                                                                                                                                                  | Extraction of complementary relation features on CC and MLO views of mammograms using relation modules. |
| Liu et al., 2020c| 2020 | Breast mass detection from mammograms                                             | BG-RCNN: Incorporating Bipartite Graph convolutional Network (BGN) into Mask RCNN           | DDSM and a private dataset                                                                                                                                                                                  | (1) Modeling relations (e.g., complementary information and visual correspondeces) between CC and MLO views of mammograms using BGN; (2) Defining simple pseudo landmarks in mammograms to facilitate learning geometric relations. |
| Rijthoven et al., 2018 | 2018 | Lymphocytes detection in whole-slide (WSI) histology images of breast, colon, and prostate cancer | Smaller YOLOv2 with much fewer layers                                                         | Private dataset                                                                                                                                                                                              | (1) Simplifying the original YOLO network using prior knowledge of lymphocytes (e.g., average size, no overlaps); (2) Designing a new training sampling strategy using the prior knowledge (i.e., brown areas without lymphocytes contain hard negative samples). |
| Lin et al., 2019 | 2019 | Lymph node metastasis detection from WSI histology images                         | Modified Fully convolutional network (FCN) based on VGG16                                    | Camelyon16 dataset and ISBI 2016                                                                                                                  | (1) Utilizing FCN for fast gigapixel-level WSI analysis; (2) Proposing anchor layers for model conversion to ensure dense scanning; (3) Hard negative mining.                        |
| Nair et al., 2020| 2020 | Multiple sclerosis lesion detection from MR brain images                           | 3D U-Net based segmentation network to obtain lesions                                        | Private dataset                                                                                                                                                                                              | (1) Uncertainty estimation using Monte Carlo (MC) dropout; (2) Using multiple uncertainty measures to filter out uncertain predictions of lesion candidates.                                          |
| Yan et al., 2018a| 2018 | Detection of lung, mediastinum, liver, soft tissue, pelvis, abdomen, kidney, and bone lesions from CT images | 3DCE: Modified R-FCN                                                                          | DeepLesion                                                                                                                                                                                                | (1) Exploiting 3D context information; (2) Leveraging pre-trained 2D backbones (VGG-16) for transfer learning.                                      |
| Tang et al., 2019| 2019 | Detection of various types of lesions in DeepLesion                              | ULDor: Mask R-CNN with ResNet-101 as backbone                                               | DeepLesion                                                                                                                                                                                                | (1) Pseudo mask construction using RECIST annotations; (2) Hard negative mining to learn more discriminative features for false positives reduction.                                      |
| Yan et al., 2019 | 2019 | Detection of various types of lesions in DeepLesion                              | MULAN: Modified Mask RCNN with DenseNet-121 as backbone                                       | DeepLesion                                                                                                                                                                                                | (1) Jointly performing three different tasks (detection, tagging, and segmentation) for better performance; (2) A new 3D feature fusion strategy.                                             |
| Tao et al., 2019 | 2019 | Detection of various types of lesions in DeepLesion                              | Improved R-FCN                                                                                | DeepLesion                                                                                                                                                                                                | Contextual attention module aggregates relevant context features, and spatial attention module highlights discriminative features for small objects.                                           |
| Li et al., 2019  | 2019 | Detection of various types of lesions in DeepLesion                              | MVP-Net: a three pathway architecture with FPN as backbone                                   | DeepLesion                                                                                                                                                                                                | Using an attention module to incorporate clinical knowledge of multi-view window inspection and position information.                                                                    |

### 3.4. Registration

Registration, the process of aligning two or more images into one coordinate system with matched contents, is also an important step in many (semi-)automatic medical image analysis tasks. Image registration can be sorted into two groups: rigid and deformable (non-rigid) (Fu et al., 2020). In rigid registration, a simple transform (e.g., rotation, translation) is uniformly applied to all the image pixels (Shang et al., 2006); in contrast, deformable registration allows a non-uniform mapping between images (Oh and Kim, 2017). Although the use of deep learning in medical registration tasks is not as popular as in medical image classification, segmentation, and detection, we have seen more applications of deep learning related to this research topic rising in recent years, especially for deformable image registration. Deep learning-based medical image registration approaches can be roughly categorized into three groups: (1) deep iterative
In deep iterative registration, deep learning models learn a metric that quantifies the similarity between a target/moving image and a reference/fixed image; then the learned similarity metric is used in conjunction with classical (i.e., non-learning-based) frameworks to perform image registration (Haskins et al., 2020). For example, Simonovsky et al. (2016) used a 5-layer CNN to learn a metric to evaluate the similarity between aligned 3D brain MRI T1–T2 image pairs, and then incorporated the learnt metric into a continuous optimization framework to complete deformable registration. The deep learning-based metric outperformed standard metrics such as mutual information for multimodal registration (Simonovsky et al., 2016). In essence, this work is most related to previous approach in (Cheng et al., 2018) that estimates the similarity of 2D CT–MR patch pairs using an FCN pre-trained with stacked denoising autoencoder; the major difference between these two works lies in network architecture (CNN vs. FCN), application scenario (3D vs. 2D), and training strategy (from scratch vs. pre-training). For T1–T2 weighted MR images and CT–MR images, Haskins et al. (2019) claimed it is relatively easy to learn a good similarity metric because these multimodal images share large similar views or simple intensity mappings. They extended the deep similarity metric to a more challenging scenario, 3D MR–TRUS prostate image registration, where a large appearance difference exists between the two imaging modalities. Aside from learning a similarity, deep learning can also learn parameters of classical registration frameworks, such as the Large Diffeomorphic Distance Metric Mapping (LDDMM) model (Yang et al., 2017).

Despite the success of deep iterative registration, the process of learning a similarity metric followed by iterative optimization in classic registration frameworks leads to slow registration (Haskins et al., 2020). In comparison, some supervised registration methods estimate deformation fields in just one step, without iteratively optimizing traditional registration frameworks. These methods typically require ground truth warp fields, which can be established either by simulating deformations (Fan et al., 2019) or performing classical registration (Balakrishnan et al., 2019). For 3D deformable image registration, Sokooti et al. (2017) developed multiscale CNNs based model to directly predict displacement vector fields (DVFs) between image pairs. To make their training dataset larger and more diversified, they first artificially generated DVFs with varying spatial frequency and amplitude, and then applied data augmentation on the generated DVFs, resulting in approximately 1 million training examples. After training, deformed images were registered in one-shot, and their method demonstrated close performance to a conventional B-spline registration.

Meanwhile, unsupervised learning based registration has received extensive attention in recent years (Zhao et al., 2019b; Kim et al., 2019) for two major reasons: (1) It is cumbersome to obtain ground truth warp fields via conventional registration tools (Balakrishnan et al., 2019); (2) Types of deformations used for model training are limited, leading to poor registration performance on unseen images (Balakrishnan et al., 2019b; Zhao et al., 2019a). As one of the early works related to unsupervised registration, Wu et al. (2016) argued that supervised based registration methods do not generalize well on new data; they employed a convolutional stacked autoencoder (Lee et al., 2011) to extract features from fixed and moving images to improve registration performance.

Balakrishnan et al. (2018) proposed an unsupervised registration model (VoxelMorph in Figure 6) that does not need supervised information (e.g., true registration fields or anatomical landmarks). The model has two components, including a convolutional U-Net and a spatial transformer network (STN). The authors formulated 3D MR brain volume registration as a parametric function, which was modeled using the U-Net architecture. The encoder’s input is the concatenation of a moving image and a fixed image, and the decoder outputs a registration field. The spatial transformer network (Jaderberg et al., 2015) was applied to warp the moving image with the learned registration field, resulting in a reconstructed version of the fixed image. By minimizing the difference between the reconstructed image and the fixed image, VoxelMorph can update parameters for generating desired deformation fields. The authors innovatively deployed amortized optimization (Marino et al., 2018), which can learn optimal parameters over the entire dataset of volume pairs rather than individual volume pairs. As a result, this unsupervised registration framework was able to operate orders of magnitude faster but achieved competitive performance to Symmetric Normalization (SyN) (Avants et al., 2008), a classic registration algorithm. In a later paper (Balakrishnan et al., 2019), the authors extended VoxelMorph to leverage auxiliary segmentation information (anatomical segmentation maps), and the extended model demonstrated an improved registration accuracy. Prior to this, several works had shown when there is no ground truth for voxel-level transformation, solely using auxiliary anatomical information can achieve accurate cross-modality registration (Hu et al., 2018b; Hu et al., 2018c).

DLIR is another famous unsupervised registration framework (de Vos et al., 2019), which is an extension of the previous work (de Vos et al., 2017). DLIR has four stages to progressively perform image registration. The first stage is designed for affine image registration (AIR), and the rest three stages are for deformable image registration (DIR). In the AIR stage, a CNN takes as input pairs of fixed and moving images and outputs predictions for the affine transformation
parameters so that affinely aligned image pairs can be obtained. In the subsequent DIR stage, these aligned image pairs are the input of a new CNN, whose output is a B-spline displacement vector as the deformation field. With this field, deformably registered image pairs can be obtained, and the registration results are further refined through the rest two DIR stages.

Figure 6. VoxelMorph (Balakrishnan et al., 2018).

4. DISCUSSIONS

Deep learning in the medical image analysis field has been hugely influenced by breakthroughs from computer vision. Although supervised deep learning approaches from computer vision have delivered human-level or even better performance in a variety of medical tasks such as dermatology and ophthalmology, sufficiently large annotated datasets are unfortunately not available in many scenarios (e.g., cancer images). Currently, there are two possible directions to overcome this limitation: 1) utilizing GAN model to enlarge the labeled dataset; 2) utilizing the self-supervised and semi-supervised learning models to exploit the information underlying vast unlabeled medical images. The challenges and possible future research directions of these three major technologies are presented below.

GAN has shown great promise in medical image synthesis and semi-supervised learning; however, several challenges exist when GAN is applied. (1) It still remains undermined how to build a strong connection between GAN’s generator and the target task (e.g., classifier, detector, segmentor). The lack of such connection may cause a subtle performance boost as compared to the conventional data augmentation (e.g., rotation, rescale, and flip) (Wu et al., 2018c). The connection between the generator and classifier can be strengthened by utilizing semi-supervised GAN, in which the discriminator was modified to serve as a classifier (Salimans et al., 2016). Several training strategies can be employed: identifying a “bad” generator that can significantly contribute to good semi-supervised classification (Dai et al., 2017); jointly optimizing the triple components of a generator, a discriminator, and a classifier (Li et al., 2017b). Besides these investigated strategies, it is meaningful to explore new ways that can effectively set up connections between the generator and a specific medical image task for a better performance. (2) GAN usually needs at least thousands of training examples to converge (Wang et al., 2018b), which limits its applicability on small medical datasets. This challenge can be partially addressed by using classic data augmentation for adversarial learning (Frid-Adar et al., 2018a; Frid-Adar et al., 2018b). Further, if there exist relatively large amounts of medical images that share structural, textural, and semantic similarities with the target dataset, pre-training generators and/or discriminators may facilitate faster convergence and better performance (Rubin et al., 2019). Meanwhile, some recent novel augmentation mechanisms, such as the differentiable augmentation (Zhao et al., 2020) and adaptive discriminator augmentation (Karras et al., 2020) have enabled GAN to effectively generate high-fidelity images under data-limited conditions, but they have not been applied to any medical image analysis tasks. We anticipate that these new methods can also demonstrate promising performance in future studies of the medical imaging field.
Self-supervised learning can be accomplished by large amounts of medical images with limited annotations. Many recent studies (Zhou et al., 2021; Zhou et al., 2019b; Zhu et al., 2020b; Zhang et al., 2020a; Azizi et al., 2021) uniformly reflected the great potential of self-supervised pre-training followed by supervised fine-tuning. Nonetheless, one major challenge lies in the extraction of fine-grained visual features to understand subtle differences in medical images. Given the high inter-class similarity of medical images, self-supervised learning methods originally designed based on natural images are likely to produce only suboptimal results when they are directly applied to medical image tasks (Zhang et al., 2020a). One possible solution is to leverage proper domain knowledge or task-specific properties into self-supervised pre-training, which has proven beneficial to facilitate learning useful feature representations in the medical imaging context. The domain knowledge varies a lot, such as anatomical information in MRI and CT images (Zhou et al., 2021; Zhou et al., 2019b), 3D spatial context information from volumetric images (Zhang et al., 2017; Zhuang et al., 2019; Zhu et al., 2020a), multi-instance data from the same patient (Azizi et al., 2021), patient metadata (Vu et al., 2021), radiomic features (Shorfuzzaman and Hossain, 2021), text reports accompanying images (Zhang et al., 2020a), etc. In addition, self-supervised contrastive pre-training is currently impeded by the high computing complexity of large models (e.g., ResNet-50 (4x), ResNet-152 (2x)), which require a large group of multi-core TPUs (Chen et al., 2020a; Azizi et al., 2021). When such types of computing resources are not available, directly utilizing the existing pre-trained models may lead to lower performance (Li et al., 2021b). Therefore, it should be an important direction to develop novel models or training strategies to enhance the computing efficiency. For example, Reed et al. (2021) proposed a hierarchical pre-training strategy to make the self-supervised pre-training process converge up to 80x faster with an improved accuracy across different tasks.

The latest semi-supervised learning (SSL) methods (Berthelot et al., 2019a; Berthelot et al., 2019b; Sohn et al., 2020) also require a large dataset to optimize the model. The model performance highly relies on the correctly selected data augmentation strategies. Some common strategies include weak augmentations (e.g. flipping and cropping (Berthelot et al., 2019a)), strong augmentations (e.g. cutout, invert, blur, etc. (Berthelot et al., 2019b; Cubuk et al., 2019)), and hybrid weak-strong augmentations (Sohn et al., 2020). For medical imaging applications, the weak augmentations have been widely used to enlarge the lesion dataset (Hadam et al., 2017), but the performance enhancement of the strong augmentations is limited, as useful information may be lost when applying cutout or blur to medical lesion images. To better utilize recent advances from SSL to the medical context, we believe it is important to design and select augmentation strategies that are appropriate for medical images. In addition, some recent SSL approaches (e.g., FixMatch (Sohn et al., 2020)) have proved the feasibility of using extremely small numbers of high-quality labeled samples and large amounts of unlabeled samples to achieve high classification accuracy on CIFAR-10. In our opinion, this will have a profound impact on medical tasks in low-annotation regimes.

In summary, deep learning is a fast-developing technology, and has produced promising potential in broad medical image analysis fields including disease classification, segmentation, detection, and image registration. Despite of significant research progress, we are still facing many technical challenges or pitfalls (Roberts et al., 2021) to develop deep learning based CAD schemes that can achieve high scientific rigor. Therefore, more research efforts are needed to overcome these pitfalls before the deep learning based CAD schemes can be commonly accepted by clinicians.

5. REFERENCES

[1]. Meyers, P.H., Nice Jr, C.M., Becker, H.C., Nettleton Jr, W.J., Sweeney, J.W., Meckstroth, G.R., 1964. Automated computer analysis of radiographic images. Radiology 83, 1029-1034.
[2]. Kruger, R.P., Townes, J.R., Hall, D.L., Dwyer, S.J., Lodwick, G.S., 1972. Automated Radiographic Diagnosis via Feature Extraction and Classification of Cardiac Size and Shape Descriptors. IEEE Transactions on Biomedical Engineering BME-19, 174-186.
[3]. Sezaki, N., Ukena, K., 1973. Automatic Computation of the Cardiothoracic Ratio with Application to Mass Screening. IEEE Transactions on Biomedical Engineering BME-20, 248-253.
[4]. Doi, K., MacMahon, H., Katsuragawa, S., Nishikawa, R.M., Jiang, Y., 1999. Computer-aided diagnosis in radiology: potential and pitfalls. European Journal of Radiology 31, 97-109.
[5]. Shi, J., Sahiner, B., Chan, H.-P., Ge, J., Hadijiski, L., Helvie, M.A., Nees, A., Wu, Y.-T., Wei, J., Zhou, C., Zhang, Y., Cui, J., 2008. Characterization of mammographic masses based on level set segmentation with new image features and patient information. Med Phys 35, 280-290.
[6]. Sahiner, B., Petrick, N., Heang-Ping, C., Hadijiski, I.M., Paramagul, C., Helvie, M.A., Gurcan, M.N., 2001. Computer-aided characterization of mammographic masses: accuracy of mass segmentation and its effects on characterization. IEEE Transactions on Medical Imaging 20, 1275-1284.
[7]. LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436-444.

[8]. Tanaka, H., Chiu, S.-W., Watanabe, T., Kaoku, S., Yamaguchi, T., 2019. Computer-aided diagnosis system for breast ultrasound images using deep learning. Physics in Medicine & Biology 64, 235013.

[9]. Li, X., Jia, M., Islam, M.T., Yu, L., Xing, L., 2020a. Self-Supervised Feature Learning via Exploiting Multi-Modal Data for Retinal Disease Diagnosis. IEEE Trans Med Imaging 39, 4023-4033.

[10]. Shorffuzzaman, M., Hassain, M.S., 2021. MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. Pattern recognition 113, 107700.

[11]. Zhang, Y., Jiang, H., Miura, Y., Manning, C., Langlotz, C., 2020a. Contrastive Learning of Medical Visual Representations from Paired Images and Text, arXiv preprint arXiv:2010.00747, pp. 1-15.

[12]. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H., 2018a. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 321, 321-331.

[13]. Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D., 2016. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE journal of biomedical and health informatics 21, 31-40.

[14]. Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K., 2018. Recurrent residual convolutional neural network based on u-net (r2unet) for medical image segmentation. arXiv preprint arXiv:1802.06955.

[15]. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A., 2019. Uncertainty-Aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation, In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 605-613.

[16]. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Inf-Net: Automatic COVID-19 Lung Infection Segmentation From CT Images. IEEE Transactions on Medical Imaging 39, 2626-2637.

[17]. Rijthoven, M., Swiderska-Chadaj, Z., Seeliger, K., Laak, J.v.d., Ciompi, F., 2018. You Only Look on Lymphocytes Once, Medical Imaging with Deep Learning, pp. 1-15.

[18]. Mei, J., Cheng, M., Xu, G., Wan, L., Zhang, H., 2021. SANet: A Slice-Aware Network for Pulmonary Nodule Detection. IEEE Transactions on Pattern Analysis & Machine Intelligence, pre-print.

[19]. Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. Medical Image Analysis 59, 101557.

[20]. Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D., 2015. 3D deep learning for efficient and robust landmark detection in volumetric data. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 565-572.

[21]. Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N., 2016. A deep metric for multimodal registration, International conference on medical image computing and computer-assisted intervention. Springer, pp. 10-18.

[22]. Sokooti, H., de Vos, B., Berendsen, F., Lelieveldt, B.P.F., Işgum, I., Staring, M., 2017. Nonrigid Image Registration Using Multi-scale 3D Convolutional Neural Networks, In: Descôteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. Springer International Publishing, Cham, pp. 232-239.

[23]. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Dalca, A.V., Guttag, J., 2018. An Unsupervised Learning Model for Deformable Medical Image Registration, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9252-9260.

[24]. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115-118.

[25]. Long, E., Lin, H., Liu, Z., Wu, X., Wang, L., Jiang, J., An, Y., Lin, Z., Li, X., Chen, J., Li, J., Cao, Q., Wang, D., Liu, X., Chen, W., Liu, Y., 2017. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nature Biomedical Engineering 1, 0024.

[26]. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60-88.

[27]. Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221-248.

[28]. Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: A review. Medical image analysis 58, 101552.

[29]. Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A., 2020a. GANs for medical image analysis. Artificial Intelligence in Medicine 109, 101938.
[56]. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum Contrast for Unsupervised Visual Representation Learning, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9726-9735.

[57]. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context Encoders: Feature Learning by Inpainting, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2536-2544.

[58]. Zhang, R., Isola, P., Efros, A.A., 2016a. Colorful image colorization, European conference on computer vision. Springer, pp. 649-666.

[59]. Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised Visual Representation Learning by Context Prediction, 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1422-1430.

[60]. Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles, European conference on computer vision. Springer, pp. 69-84.

[61]. Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised Representation Learning by Predicting Image Rotations, International Conference on Learning Representations, pp. 1-16.

[62]. Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A Simple Framework for Contrastive Learning of Visual Representations, In: Hal, D., III, Aarti, S. (Eds.), Proceedings of the 37th International Conference on Machine Learning. PMLR, Proceedings of Machine Learning Research, pp. 1597--1607.

[63]. Becker, S., Hinton, G.E., 1992. Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature 355, 161-163.

[64]. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations, Advances in Neural Information Processing Systems, pp. 1-13.

[65]. Chen, X., Fan, H., Girshick, R., He, K., 2020b. Improved Baselines with Momentum Contrastive Learning, arXiv preprint arXiv:2003.04297, pp. 1-3.

[66]. Rasmus, A., Valpola, H., Honkala, M., Berglund, M., Raiko, T., 2015. Semi-supervised learning with Ladder networks, Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2, pp. 3546-3554.

[67]. Laine, S., Aila, T., 2017. Temporal Ensembling for Semi-Supervised Learning.

[68]. Tarvainen, A., Valpola, H., 2017a. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 1195-1204.

[69]. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q., 2020. Unsupervised Data Augmentation for Consistency Training, Advances in Neural Information Processing Systems, pp. 1-13.

[70]. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C., 2019a. MixMatch: A Holistic Approach to Semi-Supervised Learning, Advances in Neural Information Processing Systems, pp. 1-11.

[71]. Lee, D.-H., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, Workshop on challenges in representation learning, ICML.

[72]. Zhang, H., Cisse, M., Dauphin, Y., Lopez-Paz, D., 2018a. Mixup: Beyond empirical risk minimization, International Conference on Learning Representations, pp. 1-13.

[73]. Arazo, E., Ortega, D., Albert, P., O’Connor, N.E., McGuinness, K., 2020. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning, 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-8.

[74]. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A., 2018. Deep Co-Training for Semi-Supervised Image Recognition, In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp. 142-159.

[75]. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training GANs, Proceedings of the 30th International Conference on Neural Information Processing Systems. Curran Associates Inc., Barcelona, Spain, pp. 2234–2242.

[76]. Odena, A., 2016. Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583.

[77]. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A., 2021a. Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation. IEEE Transactions on Neural Networks and Learning Systems 32, 523-534.

[78]. van Ginneken, B., Schaefer-Prokop, C.M., Prokop, M., 2011. Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 261, 719-732.

[79]. Doi, K., 2007. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph 31, 198-211.

[80]. Doi, K., 2005. Current status and future potential of computer-aided diagnosis in medical imaging. The British journal of radiology 78 Spec No 1, S3-s19.
[81]. Shin, H., Orton, M.R., Collins, D.J., Doran, S.J., Leach, M.O., 2013. Stacked Autoencoders for Unsupervised Feature Learning and Multiple Organ Detection in a Pilot Study Using 4D Patient Data. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 1930-1943.

[82]. Zhang, X., Dou, H., Ju, T., Xu, J., Zhang, S., 2016b. Fusing Heterogeneous Features From Stacked Sparse Autoencoder for Histopathological Image Analysis. IEEE Journal of Biomedical and Health Informatics 20, 1377-1383.

[83]. Suk, H.I., Wee, C.Y., Lee, S.W., Shen, D., 2016. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. NeuroImage 129, 292-307.

[84]. Chen, M., Shi, X., Zhang, Y., Wu, D., Guizani, M., 2017a. Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network. IEEE Transactions on Big Data, 1-1.

[85]. Cheng, J.Z., Ni, D., Chou, Y.H., Qin, J., Tiu, C.M., Chang, Y.C., Huang, C.S., Shen, D., Chen, C.M., 2016. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Scientific reports 6, 24454.

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

[87]. Chen, S., Ma, K., Zheng, Y., 2019a. Med3D: Transfer Learning for 3D Medical Image Analysis, arXiv preprint arXiv:1904.00625, pp. 1-12.

[88]. Simard, P.Y., Steinkraus, D., Platt, J.C., 2003. Best practices for convolutional neural networks applied to visual document analysis, Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., pp. 958-963.

[89]. Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, pp. 1–9.

[90]. Ciregan, D., Meier, U., Schmidhuber, J., 2012. Multi-column deep neural networks for image classification, 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3642-3649.

[91]. Li, W., Matthew, Z., Sixin, Z., Yann Le, C., Rob, F., 2013. Regularization of Neural Networks using DropConnect. PMLR, pp. 1058-1066.

[92]. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T., 2014. Decaf: A deep convolutional activation feature for generic visual recognition, International conference on machine learning. PMLR, pp. 647-655.

[93]. Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition.

[94]. Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9.

[95]. He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.

[96]. Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q., 2017. Densely Connected Convolutional Networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269.

[97]. Deng, J., Dong, W., Socher, R., Li, L., Kai, L., Li, F.-F., 2009. ImageNet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255.

[98]. de Bruijne, M., 2016. Machine learning approaches in medical image analysis: From detection to diagnosis. Medical Image Analysis 33, 94-97.

[99]. Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D., 2017. An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification. IEEE Journal of Biomedical and Health Informatics 21, 31-40.

[100]. Shin, H., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Transactions on Medical Imaging 35, 1285-1298.

[101]. Huynh, B.Q., Li, H., Giger, M.L., 2016. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. Journal of medical imaging (Bellingham, Wash.) 3, 034501.

[102]. Liu, T., Xie, S., Zhang, Y., Yu, J., Niu, L., Sun, W., 2017. Feature selection and thyroid nodule classification using transfer learning, 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, pp. 1096-1099.

[103]. Droste, R., Cai, Y., Sharma, H., Chatelain, P., Dukker, L., Papageorghiou, A.T., Noble, J.A., 2019. Ultrasound image representation learning by modeling sonographer visual attention, International conference on information processing in medical imaging. Springer, pp. 592-604.

[104]. Khan, S., Islam, N., Jan, Z., Ud Din, I., Rodrigues, J.J.P.C., 2019. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters 125, 1-6.
[105]. He, K., Girshick, R., Dollar, P., 2019a. Rethinking ImageNet Pre-Training. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4917-4926.
[106]. Kornblith, S., Shlens, J., Le, Q.V., 2019. Do Better ImageNet Models Transfer Better?, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2656-2666.
[107]. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding Transfer Learning for Medical Imaging. Advances in Neural Information Processing Systems, pp. 1-11.
[108]. Frid-Adar, M., Klang, E., Amiatii, M., Goldberger, J., Greenspan, H., 2018b. Synthetic data augmentation using GAN for improved liver lesion classification, 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 289-293.
[109]. Wu, E., Wu, K., Cox, D., Lotter, W., 2018a. Conditional Infilling GANs for Data Augmentation in Mammogram Classification, arXiv preprint arXiv:1807.08093, pp. 1-8.
[110]. Shams, S., Patania, R., Zhang, J., Kim, J., Lee, K., Park, S.-J., 2018. Deep Generative Breast Cancer Screening and Diagnosis, In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 859-867.
[111]. Rubin, M., Stein, O., Turko, N.A., Ngyate, Y., Roitshtain, D., Karako, L., Barnea, I., Giryes, R., Shaked, N.T., 2019. TOP-GAN: Stain-free cancer cell classification using deep learning with a small training set. Medical Image Analysis 57, 176-185.
[112]. Bai, W., Chen, C., Tarroni, G., Duan, J., Guittot, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D., 2019. Self-Supervised Learning for Cardiac MR Image Segmentation by Anatomical Position Prediction, In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 541-549.
[113]. Tao, X., Li, Y., Zhou, W., Ma, K., Zheng, Y., 2020. Revisiting Rubik’s Cube: Self-supervised Learning with Volume-Wise Transformation for 3D Medical Image Segmentation, In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer International Publishing, Cham, pp. 238-248.
[114]. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M., 2021. Big Self-Supervised Models Advance Medical Image Classification, arXiv preprint arXiv:2101.05224, pp. 1-19.
[115]. Chen, T., Liu, S., Chang, S., Cheng, Y., Amini, L., Wang, Z., 2020c. Adversarial Robustness: From Self-Supervised Pre-Training to Fine-Tuning, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 696-705.
[116]. Misra, I., Maaten, L.v.d., 2020. Self-supervised learning of pretext-invariant representations, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6707-6717.
[117]. Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[118]. Tajbakhsh, N., Hu, Y., Cao, J., Yan, X., Xiao, Y., Lu, Y., Liang, J., Terzopoulos, D., Ding, X., 2019. Surrogate Supervision for Medical Image Analysis: Effective Deep Learning From Limited Quantities of Labeled Data, 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1251-1255.
[119]. Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., Zheng, Y., 2019. Self-supervised Feature Learning for 3D Medical Images by Playing a Rubik’s Cube, In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 420-428.
[120]. Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S.K., Zheng, Y., 2020a. Rubik’s Cube+: A self-supervised feature learning framework for 3D medical image analysis. Medical Image Analysis 64, 101746.
[121]. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2019b. Self-supervised learning for medical image analysis using image context restoration. Medical Image Analysis 58, 101539.
[122]. Larsson, G., Mairé, M., Shakhnarovich, G., 2017. Colorization as a Proxy Task for Visual Understanding. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 840-849.
[123]. Arjovsky, M., Chintala, S., Bottou, L., 2017b. Wasserstein Generative Adversarial Networks, In: Doina, P., Yee Whye, T. (Eds.), Proceedings of the 34th International Conference on Machine Learning. PMLR, Proceedings of Machine Learning Research, pp. 214–223.
[124]. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E., 2020d. A Simple Framework for Contrastive Learning of Visual Representations, International Conference on Machine Learning, pp. 1597-1607.
[125]. Ciga, O., Martel, A., Xu, T., 2020. Self supervised contrastive learning for digital histopathology. arXiv preprint arXiv:2011.13971, pp. 1-37.
[126]. Chen, X., Yao, L., Zhou, T., Dong, J., Zhang, Y., 2021. Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images. Pattern recognition 113, 107826.
[127]. Sowrirajan, H., Yang, J., Ng, A., Rajpurkar, P., 2020. MoCo-CXR: MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models. arXiv preprint arXiv:2010.05352, pp. 1-12.
[128]. Vu, Y.N.T., Wang, R., Balachandar, N., Liu, C., Ng, A., Rajpurkar, P., 2021. MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation, arXiv preprint arXiv:2102.10663, pp. 1-13.
[129]. Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T., 2018a. Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation, 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1038-1042.
[130]. Kingma, D.P., Rezende, D.J., Mohamed, S., Welling, M., 2014. Semi-supervised learning with deep generative models, Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. MIT Press, Montreal, Canada, pp. 3581–3589.
[131]. Madani, A., Ong, J.R., Tibrewal, A., Mofrad, M.R.K., 2018b. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. npj Digital Medicine 1, 59.
[132]. Lecouat, B., Chang, K., Foo, C., Unnikrishnan, B., Brown, J., Zenati, H., Beers, A., Chandrasekhar, V., Kalpathy-Cramer, J., Krishnaswamy, P., 2018. Semi-Supervised Deep Learning for Abnormality Classification in Retinal Images, arXiv preprint arXiv:1812.07832, pp. 1-9.
[133]. Xie, Y., Zhang, J., Xia, Y., 2019a. Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT. Medical Image Analysis 57, 237-248.
[134]. Tarvainen, A., Valpola, H., 2017b. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 1195–1204.
[135]. Shang, H., Sun, Z., Yang, W., Fu, X., Zheng, H., Chang, J., Huang, J., 2019. Leveraging Other Datasets for Medical Imaging Classification: Evaluation of Transfer, Multi-task and Semi-supervised Learning. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 431-439.
[136]. Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A., 2020a. Semi-Supervised Medical Image Classification With Relation-Driven Self-Ensembling Model. IEEE Transactions on Medical Imaging 39, 3429-3440.
[137]. Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural Machine Translation by Jointly Learning to Align and Translate.
[138]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 6000–6010.
[139]. Luong, M., Pham, H., Manning, C., 2015. Effective Approaches to Attention-based Neural Machine Translation, arXiv preprint arXiv:1508.04025, pp. 1-11.
[140]. Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: neural image caption generation with visual attention, Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. JMLR.org, Lille, France, pp. 2048–2057.
[141]. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J., 2016. Image Captioning with Semantic Attention, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4651-4659.
[142]. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6077-6086.
[143]. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual Attention Network for Image Classification, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6450-6458.
[144]. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018a. CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp. 3-19.
[145]. Jetley, S., Lord, N., Lee, N., Torr, P., 2018. Learn to Pay Attention, International Conference on Learning Representations, pp. 1-14.
[146]. Chen, L., Yang, Y., Wang, J., Xu, W., Yuille, A.L., 2016. Attention to Scale: Scale-Aware Semantic Image Segmentation, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3640-3649.
[147]. Ren, M., Zemel, R.S., 2017. End-to-End Instance Segmentation with Recurrent Attention, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 293-301.

[148]. Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., Shao, L., 2019a. Collaborative Learning of Semi-Supervised Segmentation and Classification for Medical Images, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2074-2083.

[149]. Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y., 2018. Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification, arXiv preprint arXiv:1801.09927, pp. 1-10.

[150]. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis 53, 197-207.

[151]. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D., 2017. SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. IEEE Transactions on Medical Imaging 36, 2204-2215.

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

[153]. Milletari, F., Navab, N., Ahmadi, S., 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, 2016 Fourth International Conference on 3D Vision (3DV), pp. 565-571.

[154]. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation, In: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., Madabhushi, A. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer International Publishing, Cham, pp. 3-11.

[155]. Hariparan, B., Arbeláz, P., Girshick, R., Malik, J., 2015. Hypercolumns for object segmentation and fine-grained localization, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 447-456.

[156]. Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431-3440.

[157]. Drozdal, M., Vorontsov, E., Charttrand, G., Kadoury, S., Pal, C., 2016. The Importance of Skip Connections in Biomedical Image Segmentation, In: Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J.M.R.S., Belagiannis, V., Papa, J.P., Nascimento, J.C., Loog, M., Lu, Z., Cardoso, J.S., Cornebise, J. (Eds.), Deep Learning and Data Labeling for Medical Applications. Springer International Publishing, Cham, pp. 179-187.

[158]. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation, International conference on medical image computing and computer-assisted intervention. Springer, pp. 424-432.

[159]. Gibson, E., Gigi, F., Hu, Y., Bonnati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic Multi-Organ Segmentation on Abdominal CT With Dense V-Networks. IEEE transactions on medical imaging 37, 1822-1834.

[160]. Ming, L., Xiaolin, H., 2015. Recurrent convolutional neural network for object recognition, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3367-3375.

[161]. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, F.A., 2018. H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes. IEEE Transactions on Medical Imaging 37, 2663-2674.

[162]. Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X., 2018. SegAN: Adversarial Network with Multi-scale L1 Loss for Medical Image Segmentation. Neuroinformatics 16, 383-392.

[163]. Zhang, Y., Miao, S., Mansi, T., Liao, R., 2020b. Unsupervised X-ray image segmentation with task driven generative adversarial networks. Medical Image Analysis 62, 101664.

[164]. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas, Medical Imaging with Deep Learning, pp. 1-14.

[165]. Nie, D., Gao, Y., Wang, L., Shen, D., 2018. ASDNet: Attention Based Semi-supervised Deep Networks for Medical Image Segmentation, In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 370-378.

[166]. Sinha, A., Dolz, J., 2021. Multi-Scale Self-Guided Attention for Medical Image Segmentation. IEEE Journal of Biomedical and Health Informatics 25, 121-130.
[167]. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34-45.

[168]. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E., 2019. PHiSeg: Capturing Uncertainty in Medical Image Segmentation, In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 119-127.

[169]. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T., 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging 39, 3868-3878.

[170]. Zhang, Z., Yang, L., Zheng, Y., 2018b. Translating and Segmenting Multimodal Medical Volumes with Cycle-and Shape-Consistency Generative Adversarial Network, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9242-9251.

[171]. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V., 2019a. Data Augmentation Using Learned Transformations for One-Shot Medical Image Segmentation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8535-8545.

[172]. Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision?, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 5580–5590.

[173]. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A., 2020b. Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation. IEEE Transactions on Neural Networks and Learning Systems 32, 523-534.

[174]. Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., Meng, T., Li, K., Huang, N., Zhang, S., 2020a. A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions From CT Images. IEEE Transactions on Medical Imaging 39, 2653-2663.

[175]. Pang, Y., Li, Y., Shen, J., Shao, L., 2019. Towards Bridging Semantic Gap to Improve Semantic Segmentation, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4229-4238.

[176]. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018a. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp. 833-851.

[177]. Wu, Z., Su, L., Huang, Q., 2019a. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3902-3911.

[178]. Chen, S., Tan, X., Wang, B., Hu, X., 2018b. Reverse Attention for Salient Object Detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp. 236-252.

[179]. Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., Shao, L., 2019. ET-Net: A Generic Edge-atTention Guidance Network for Medical Image Segmentation, Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer-Verlag, pp. 442–450.

[180]. He, Y., Yang, G., Chen, Y., Kong, Y., Wu, J., Tang, L., Zhu, X., Dillenseger, J.-L., Shao, P., Zhang, S., Shu, H., Coatrieux, J.-L., Li, S., 2019b. DPA-DenseBiasNet: Semi-supervised 3D Fine Renal Artery Segmentation with Dense Biased Network and Deep Priori Anatomy, In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 139-147.

[181]. Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Tong, R., Wu, J., 2019. Semisupervised Segmentation of Liver Using Adversarial Learning with Deep Atlas Prior. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 148-156.

[182]. Clough, J., Byrne, N., Oksuz, I., Zimmer, V.A., Schnabel, J.A., King, A., 2020. A Topological Loss Function for Deep-Learning based Image Segmentation using Persistent Homology. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1-1.

[183]. Ganayev, P.-A., Sdika, M., Benoit-Cattin, H., 2018. Semi-supervised Learning for Segmentation Under Semantic Constraint, In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 595-602.
[184]. Li, S., Zhang, C., He, X., 2020c. Shape-Aware Semi-supervised 3D Semantic Segmentation for Medical Images, In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racocoeanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer International Publishing, Cham, pp. 552-561.

[185]. Taleb, A., Loetzscht, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C., 2020. 3D Self-Supervised Methods for Medical Imaging, Advances in Neural Information Processing Systems, pp. 1-15.

[186]. Hu, S.-Y., Wang, S., Weng, W.-H., Wang, J., Wang, X., Ozturk, A., Li, Q., Kumar, V., Samir, A., 2020. Self-Supervised Pretraining with DICOM metadata in Ultrasound Imaging, In: Finale, D.-V., Jim, F., Ken, J., David, K., Rajesh, R., Byron, W., Jenna, W. (Eds.), Proceedings of the 5th Machine Learning for Healthcare Conference. PMLR, Proceedings of Machine Learning Research, pp. 732–749.

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

[188]. Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), pp. 539-546 vol. 531.

[189]. Oord, A.v.d., Li, Y., Vinyals, O., 2019. Representation Learning with Contrastive Predictive Coding, arXiv preprint arXiv:1807.03748, pp. 1-13.

[190]. Zhang, H., Cisse, M., Y.Dauphin, Lopez-Paz, D., 2018c. mixup: Beyond Empirical Risk Minimization International Conference on Learning Representations, pp. 1-13.

[191]. Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C., 2017. Error Corrective Boosting for Learning Fully Convolutional Networks with Limited Data, In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. Springer International Publishing, Cham, pp. 231-239.

[192]. Wu, Z., Su, L., Huang, Q., 2019b. Stacked cross refinement network for edge-aware salient object detection, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7264-7273.

[193]. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2014. OverFeat: Integrated Recognition,Localization and Detection using Convolutional Networks.

[194]. Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587.

[195]. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115, 211-252.

[196]. Ciompi, F., de Hoop, B., van Riel, S.J., Chung, K., Scholten, E.T., Oudkerk, M., de Jong, P.A., Prokop, M., van Ginneken, B., 2015. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Med Image Anal 26, 195-202.

[197]. Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V.C., Shi, L., Heng, P., 2016. Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks. IEEE Transactions on Medical Imaging 35, 1182-1195.

[198]. Wolterink, J.M., Leiner, T., de Vos, B.D., van Hamersvelt, R.W., Viergever, M.A., Išgum, I., 2016. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. Med Image Anal 34, 123-136.

[199]. Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Advances in Neural Information Processing Systems, pp. 1-9.

[200]. Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788.

[201]. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection, 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999-3007.

[202]. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M., 2020b. Deep Learning for Generic Object Detection: A Survey. International Journal of Computer Vision 128, 261-318.

[203]. Girshick, R., 2015. Fast R-CNN, 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448.

[204]. Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 1137-1149.
[205]. Dai, J., Li, Y., He, K., Sun, J., 2016. R-FCN: object detection via region-based fully convolutional networks, Proceedings of the 30th International Conference on Neural Information Processing Systems. Curran Associates Inc., Barcelona, Spain, pp. 379–387.

[206]. Redmon, J., Farhadi, A., 2017. YOLO9000: Better, Faster, Stronger, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517-6525.

[207]. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. SSD: Single Shot MultiBox Detector, In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016. Springer International Publishing, Cham, pp. 21-37.

[208]. Law, H., Deng, J., 2020. CornerNet: Detecting Objects as Paired Keypoints. International Journal of Computer Vision 128, 642-656.

[209]. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q., 2019. CenterNet: Keypoint Triplets for Object Detection, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6568-6577.

[210]. He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988.

[211]. Setio, A.A.A., Traverso, A., de Bel, T., Berens, M.S.N., Bogaard, C.V.D., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., Gugten, R.V., Heng, P.A., Jansen, B., de Kaste, M.M.J., Kotov, V., Lin, J.Y., Manders, J., Sőñora-Mengana, A., García-Naranjo, J.C., Papavasileiou, E., Prokop, M., Saletta, M., Schafer-Prokop, C.M., Scholten, E.T., Scholten, L., Snoeren, M.M., Torres, E.L., Vandemeulebroucke, J., Walasek, N., Zuidhof, G.C.A., Ginneken, B.V., Jacobs, C., 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 42, 1-13.

[212]. Newell, A., Huang, Z., Deng, J., 2017. Associative embedding: end-to-end learning for joint detection and grouping, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 2274–2284.

[213]. Tychsen-Smith, L., Petersson, L., 2017. DeNet: Scalable Real-Time Object Detection with Directed Sparse Sampling, 2017 IEEE International Conference on Computer Vision (ICCV), pp. 428-436.

[214]. Gu, Y., Lu, X., Yang, L., Zhang, B., Yu, D., Zhao, Y., Gao, L., Wu, L., Zhou, T., 2018. Automatic lung nodule detection using a 3D deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs. Computers in Biology and Medicine 103, 220-231.

[215]. Xie, H., Yang, D., Sun, N., Chen, Z., Zhang, Y., 2019b. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. Pattern recognition 85, 109-119.

[216]. Akselrod-Ballin, A., Karlinsky, L., Hazan, A., Bakalo, R., Horesh, A.B., Shoshan, Y., Barkan, E., 2017. Deep Learning for Automatic Detection of Abnormal Findings in Breast Mammography, In: Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares, J.M.R.S., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., Nascimento, J.C., Cardoso, J.S., Belagiannis, V., Lu, Z. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer International Publishing, Cham, pp. 321-329.

[217]. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2018. Detecting and classifying lesions in mammograms with deep learning. Scientific reports 8, 1-7.

[218]. Lin, H., Chen, H., Graham, S., Dou, Q., Rajpoot, N., Heng, P.A., 2019. Fast ScanNet: Fast and Dense Analysis of Multi-Gigapixel Whole-Slide Images for Cancer Metastasis Detection. IEEE Transactions on Medical Imaging 38, 1948-1958.

[219]. Zhu, Z., Jin, D., Yan, K., Ho, T.-Y., Ye, X., Guo, D., Chao, C.-H., Xiao, J., Yuille, A., Lu, L., 2020b. Lymph Node Gross Tumor Volume Detection and Segmentation via Distance-Based Gating Using 3D CT/PET Imaging in Radiotherapy, In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer International Publishing, Cham, pp. 753-762.

[220]. Tao, Q., Ge, Z., Cai, J., Yin, S., See, S., 2019. Improving Deep Lesion Detection Using 3D Contextual and Spatial Attention, In: Shen, D., Liu, T., Peters, T.M., Staub, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 185-193.

[221]. Tang, Y., Yan, K., Tang, Y., Liu, J., Xiao, J., Summers, R.M., 2019. Uldor: A Universal Lesion Detector For Ct Scans With Pseudo Masks And Hard Negative Example Mining, 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 833-836.

[222]. Roth, H.R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., Kim, L., Summers, R.M., 2016. Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation. IEEE Transactions on Medical Imaging 35, 1170-1181.
[223]. Dou, Q., Chen, H., Yu, L., Qin, J., Heng, P., 2017. Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection. IEEE Transactions on Biomedical Engineering 64, 1558-1567.

[224]. Yan, K., Bagheri, M., Summers, R.M., 2018a. 3D Context Enhanced Region-Based Convolutional Neural Network for End-to-End Lesion Detection, In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 511-519.

[225]. Liao, F., Liang, M., Li, Z., Hu, X., Song, S., 2019. Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network. IEEE transactions on neural networks and learning systems 30, 3484-3495.

[226]. Ding, J., Li, A., Hu, Z., Wang, L., 2017. Accurate Pulmonary Nodule Detection in Computed Tomography Images Using Deep Convolutional Neural Networks, In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. Springer International Publishing, Cham, pp. 559-567.

[227]. Liu, S., Deng, W., 2015. Very deep convolutional neural network based image classification using small training sample size, 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 730-734.

[228]. Zhu, W., Liu, C., Fan, W., Xie, X., 2018. DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 673-681.

[229]. Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J., 2017b. Dual path networks, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 4470-4478.

[230]. Wang, X., Girshick, R., Gupta, A., He, K., 2018a. Non-local neural networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803.

[231]. Swiderska-Chadaj, Z., Pinckaers, H., van Rijthoven, M., Balkenhol, M., Melnikova, M., Geessink, O., Manson, Q., Sherman, M., Polonia, A., Parry, J., Abubakar, M., Litjens, G., van der Laak, J., Ciompi, F., 2019. Learning to detect lymphocytes in immunohistochemistry with deep learning. Medical Image Analysis 58, 101547.

[232]. Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., Yu, S., 2021. A survey on incorporating domain knowledge into deep learning for medical image analysis. Medical Image Analysis 69, 101985.

[233]. Gao, Z., Puttapirat, P., Shi, J., Li, C., 2020. Renal Cell Carcinoma Detection and Subtyping with Minimal Point-Based Annotation in Whole-Slide Images, In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer International Publishing, Cham, pp. 439-448.

[234]. Qi, H., Collins, S., Noble, J.A., 2020. Knowledge-guided Pretext Learning for Utero-placental Interface Detection. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention 12261, 582-593.

[235]. Wang, D., Zhang, Y., Zhang, K., Wang, L., 2020b. FocalMix: Semi-Supervised Learning for 3D Medical Image Detection, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3950-3959.

[236]. Liu, J., Cao, L., Akin, O., Tian, Y., 2019. 3D Feature Pyramid Network Based High Sensitivity and Specificity Pulmonary Nodule Detection, In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 513-521.

[237]. Ozdemir, O., Woodward, B., Berlin, A.A., 2017. Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection. arXiv preprint arXiv:1712.00497.

[238]. Yan, K., Tang, Y., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M., 2019. MULAN: Multitask Universal Lesion Analysis Network for Joint Lesion Detection, Tagging, and Segmentation, In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 194-202.

[239]. Yan, K., Cai, J., Zheng, Y., Harrison, A.P., Jin, D., Tang, Y.b., Tang, Y.X., Huang, L., Xiao, J., Lu, L., 2020. Learning from Multiple Datasets with Heterogeneous and Partial Labels for Universal Lesion Detection in CT. IEEE Transactions on Medical Imaging, 1-1.

[240]. Cai, J., Yan, K., Cheng, C.-T., Xiao, J., Liao, C.-H., Lu, L., Harrison, A.P., 2020. Deep Volumetric Universal Lesion Detection Using Light-Weight Pseudo 3D Convolution and Surface Point Regression, In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer International Publishing, Cham, pp. 3-13.
[264]. Zhao, S., Lau, T., Luo, J., Eric, I., Chang, C., Xu, Y., 2019b. Unsupervised 3D end-to-end medical image registration with volume tweening network. IEEE journal of biomedical and health informatics 24, 1394-1404.
[265]. Kim, B., Kim, J., Lee, J.-G., Kim, D.H., Park, S.H., Ye, J.C., 2019. Unsupervised deformable image registration using cycle-consistent cnn, International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 166-174.
[266]. Wu, G., Kim, M., Wang, Q., Munsell, B.C., Shen, D., 2016. Scalable High-Performance Image Registration Framework by Unsupervised Deep Feature Representations Learning. IEEE Transactions on Biomedical Engineering 63, 1505-1516.
[267]. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y., 2011. Unsupervised learning of hierarchical representations with convolutional deep belief networks. Communications of the ACM 54, 95-103.
[268]. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks, Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. MIT Press, Montreal, Canada, pp. 2017–2025.
[269]. Marino, J., Yue, Y., Mandt, S., 2018. Iterative Amortized Inference, In: Jennifer, D., Andreas, K. (Eds.), Proceedings of the 35th International Conference on Machine Learning. PMLR, Proceedings of Machine Learning Research, pp. 3403–3412.
[270]. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis 12, 26-41.
[271]. Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonnati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., Ourselin, S., Noble, J.A., Barratt, D.C., Vercauteren, T., 2018b. Weakly-supervised convolutional neural networks for multimodal image registration. Medical Image Analysis 49, 1-13.
[272]. Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonnati, E., Moore, C.M., Emberton, M., Noble, J.A., Barratt, D.C., Vercauteren, T., 2018c. Label-driven weakly-supervised learning for multimodal deformable image registration, 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1070-1074.
[273]. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I., 2019. A deep learning framework for unsupervised affine and deformable image registration. Med Image Anal 52, 128-143.
[274]. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I., 2017. End-to-end unsupervised deformable image registration with a convolutional neural network. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 204-212.
[275]. Wu, E., Wu, K., Cox, D., Lotter, W., 2018c. Conditional infilling GANs for data augmentation in mammogram classification, Image Analysis for Moving Organ, Breast, and Thoracic Images. Springer, pp. 98-106.
[276]. Dai, Z., Yang, Z., Yang, F., Cohen, W.W., Salakhutdinov, R., 2017. Good semi-supervised learning that requires a bad GAN. Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 6513–6523.
[277]. Li, C., Xu, K., Zhu, J., Zhang, B., 2017b. Triple generative adversarial nets, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 4091–4101.
[278]. Wang, Y., Wu, C., Herranz, L., van de Weijer, J., Gonzalez-Garcia, A., Raducanu, B., 2018b. Transferring GANs: Generating Images from Limited Data, In: Ferrari, V., Hebert, M., Smichisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp. 220-236.
[279]. Zhao, S., Liu, Z., Lin, J., Zhu, J., Han, S., 2020. Differentiable Augmentation for Data-Efficient GAN Training, Advances in Neural Information Processing Systems, pp. 1-23.
[280]. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020. Training Generative Adversarial Networks with Limited Data, Advances in Neural Information Processing Systems, pp. 1-37.
[281]. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J., 2021. Models Genesis. Medical Image Analysis 67, 101840.
[282]. Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019b. Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis, In: Shen, D., Liu, T., Peters, T.M., Staab, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 384-393.
[283]. Zhang, P., Wang, F., Zheng, Y., 2017. Self supervised deep representation learning for fine-grained body part recognition, 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 578-582.
[285]. Reed, C.J., Yue, X., Nrusimha, A., Ebrahimi, S., Vijaykumar, V., Mao, R., Li, B., Zhang, S., Guillery, D., Metzger, S., 2021. Self-Supervised Pretraining Improves Self-Supervised Pretraining. arXiv preprint arXiv:2103.12718.

[286]. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C., 2019b. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. CoRR abs/1911.09785.

[287]. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E., Kurakin, A., Li, C., 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence, Advances in Neural Information Processing Systems, pp. 1-13.

[288]. Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V., 2019. AutoAugment: Learning Augmentation Strategies From Data, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 113-123.

[289]. Hadad, O., Bakalo, R., Ben-Ari, R., Hashoul, S., Amit, G., 2017. Classification of breast lesions using cross-modal deep learning, 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 109-112.

[290]. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.L., Etmann, C., McCague, C., Beer, L., 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence 3, 199-217.