










# *MTPret*: Improving X-Ray Image Analytics With Multitask Pretraining

Weibin Liao , Qingzhong Wang , *Member, IEEE*, Xuhong Li , *Member, IEEE*, Yi Liu , Zeyu Chen ,  
Siyu Huang , Dejing Dou , *Senior Member, IEEE*, Yanwu Xu , *Senior Member, IEEE*,  
and Haoyi Xiong , *Senior Member, IEEE*

**Abstract**—While deep neural networks (DNNs) have been widely used in various X-ray image analytics tasks such as classification, segmentation, detection, etc., there frequently needs to collect and annotate a huge amount of training data to train a model for every single task. In this work, we proposed a multitask self-supervised pretraining strategy *MTPret* to improve the performance of DNNs in various X-ray analytics tasks. *MTPret* first trains the backbone to learn visual representations from multiple datasets of different tasks through contrastive learning, then *MTPret* leverages a multitask continual learning to learn discriminative features from various downstream tasks. To evaluate the performance of *MTPret*, we collected eleven X-ray image datasets from different body parts, such as heads, chest, lungs, bones, and etc., for various tasks to pretrain backbones, and fine-tuned the networks on seven of the tasks. The evaluation results on top of the seven tasks showed *MTPret* outperformed a large number of baseline methods, including other initialization strategies, pretrained models, and task-specific algorithms in recent studies. In addition, we also performed experiments based on two external tasks, where the datasets of external tasks have not been used in pretraining. The excellent performance of *MTPret* further confirmed the generalizability and superiority of the proposed multitask self-supervised pretraining.

**Impact Statement**—This work has pushed back the frontiers of AI-enabled medical image analytics, which is the need for a large amount of annotated data to train models for different tasks. The proposed *MTPret* strategy reduces this need by leveraging self-supervised pretraining and multitask learning to improve the performance of DNNs across multiple tasks. The proposal of using multidataset contrastive learning and multitask continual learning to pretrain the backbone on multiple datasets of different tasks is particularly innovative. This approach allows the network to learn more generalizable features that can be

applied to a range of downstream tasks, and it also tests the feasibility of multitask pretraining for large foundational models of broader medical or clinical interests.

**Index Terms**—AI for medical image analytics, continual learning (CL), multitask learning, self-supervised learning.

## I. INTRODUCTION

MEDICAL imaging technologies, including computed tomography (CT), magnetic resonance imaging (MRI), and X-ray, have greatly enhanced our ability to detect, diagnose, and treat diseases at earlier stages [1]. Benefit from the development of computer-assisted interventions and machine learning, the analysis of medical images is no longer limited to interpretation by human experts such as radiologists and physicians, but adopts computational medical image analytics technology such as deep neural networks (DNNs) algorithm, which is conducive to alleviate the potential fatigue of human experts. A large number of works have indicated that DNN solution can reach a level similar to that of experienced medical professionals in a variety of tasks related to analyzing medical images, such as *COVID-19 detection*, *skeletal abnormality classification*, *lung segmentation*, *tuberculosis (TB) detection*, and so on [2], [3].

Several recent works have shown self-supervised learning (SSL) [4], [5] and multitask learning (MTL) [6], [7] approaches can effectively improve the performance of DNN in X-ray image analytics tasks. SSL proposes various pretext tasks that facilitate feature learning through pseudolabels and utilizes unlabeled data to acquire underlying representations. On the other hand, MTL aims to extract valuable information from multiple related tasks in order to enhance the generalization performance of all tasks. The combination of these two machine learning paradigms can further improve the performance of representation learning [8]. We follow this line of research and intend to investigate the contribution and significance of using multitask self-supervised learning algorithms based on both labeled/unlabeled datasets to improve the performance of DNNs for X-ray image analytics.

Dong et al. [5] developed a framework for self-supervised multitask representation learning in sequential 2-D medical images. The framework utilizes multiple pretext tasks to exploit underlying structures and improve cardiac segmentation. Similarly, [9] also combined SSL and MTL to optimize the performance of quantifying CT image quality. While we have

Manuscript received 24 February 2024; accepted 1 May 2024. Date of publication 15 May 2024; date of current version 10 September 2024. An earlier version of this paper was presented at the MICCAI-22 [DOI: 10.1007/978-3-031-16452-1\_15]. This article was recommended for publication by Associate Editor Alejandro F. Frangi upon evaluation of the reviewers' comments. (*Corresponding author: Haoyi Xiong.*)

Weibin Liao, Qingzhong Wang, Xuhong Li, Yi Liu, Zeyu Chen, Dejing Dou, and Haoyi Xiong are with Baidu Inc., Beijing 100193, China (e-mail: liaoweibin@baidu.com; qingzhang@outlook.com; lixuhong@baidu.com; liuyi22@baidu.com; chenzeyu01@baidu.com; doudejing@baidu.com; haoyi.xiong.fr@ieee.org).

Siyu Huang is with Harvard University, Cambridge, MA 02138 USA (e-mail: huang@seas.harvard.edu).

Yanwu Xu is with Baidu Inc., Beijing 100193, China, also with the School of Future Technology, South China University of Technology, Guangzhou 510641, China, and also with Pazhou Lab, Guangzhou 510005, China (e-mail: ywxu@ieee.org).

Digital Object Identifier 10.1109/TAI.2024.3400750

seen the power of SSL and MTL for medical image analytics tasks, they are limited to the same type of task [5] are limited to segmentation, while [9] are limited to regression) and tasks exist in the learning phase. Especially, one could pretrain a powerful backbone neural networks with multiple datasets through self-supervised multitask learning, and then transfer the pretrained backbone to various tasks [10].

In this work, we proposed a promising pretraining strategy, namely *MTPret*—a *multitask pretraining* pipeline based on self-supervised representation learning for various analytical tasks in X-ray images, including classification, segmentation and bounding box (bbox) detection. It pretrained a backbone on a collection of *eleven X-ray image datasets* and validated performance on *seven medical analytics tasks*. We made contributions as follows.

- 1) In this work, we investigate the potential of utilizing self-supervised representation learning to pretrain X-ray models using multiple datasets and tasks. To the best of our knowledge, few studies have explored this area [4], [7], especially by addressing nontrivial technical challenges in multidataset contrastive learning (e.g., *domain divergence between different X-ray datasets*) and multitask continual learning (e.g., *overfitting* and *catastrophic forgetting* [11]).
- 2) We present *MTPret* uses multiple datasets to train a backbone neural network using self-supervised learning and multitask continual learning techniques. Specifically, on top of different tasks/datasets *MTPret* pretrains the backbone using task-specific heads, which include fully connected (FC) layers, DeepLab-V3 [12], and FasterRCNN [13], for the tasks of classification, segmentation, and abnormality detection, respectively. Generally, *MTPret* is with a three-step approach: Given a DNN as a backbone neural network, *MTPret* leverages so-called a) MD-MoCo to preprocess multiple datasets and pretrain the backbone using all datasets in self-supervised manner. Then, *MTPret* b) pretrains the backbone to learn discriminative features with continual learning (CL) subject to multiple tasks, while avoiding over-fitting and “catastrophic forgetting” using advanced transfer learning techniques [11]. *MTPret* independently c) fine-tunes the pretrained backbone on each individual task to obtain the learning outcomes for the task.
- 3) We have designed several experiments in detail to validate this hypothesis, where *MTPret* pretrains and fine-tunes the network to adapt all tasks. The evaluation results demonstrate that *MTPret* achieved superior performance compared to backbones pretrained on ImageNet/MoCo [4] when using ResNet-18 and ResNet-50 as backbones. The comparison results confirms the advantage of *MTPret*. On the other hand, *MTPret* does not only work well on the datasets used for pretraining, it also deliver good performance for tasks that are out of distribution. With two external validation experiment, *MTPret* demonstrates decent generalization capabilities and can be transferred to other datasets not involved in pretraining. Please be advised that the motivation of *MTPret* is not to provide

a solution to all these X-ray image analysis tasks with best accuracy, but to study the “proof-of-concept” of self-supervised multitask learning to improve the performance of X-ray image analysis through leverage multiple tasks derived from multiple datasets.

## II. *MTPRET*: FRAMEWORK DESIGN AND LEARNING ALGORITHMS

In this section, we introduce *MTPret*, a self-supervised multitask pretraining pipeline. It is constructed from scratch by effectively utilizing ample unlabeled data for improving X-ray image representation learning of backbone, then it enables the backbone to learn discriminative features that are beneficial to tasks from scarce labeled data.

### A. Framework Designs of *MTPret*

We aim to construct a pretraining pipeline for the deep learning backbone, to generate feature representations of X-rays for medical analytics tasks from a large amount of unlabeled images. While supervised models generally exhibit higher accuracy compared to unsupervised models, they often require significantly more labeled data. This is particularly challenging in clinical settings where there may be limited annotated data but a substantial amount of unlabeled data available. In addition, we aim to the pretrained backbone learns the similarity and specificity of representations between different datasets to adapt to images from different body parts, and learns more knowledge from different tasks such as classification, segmentation, and bbox detection to better match those unknown future tasks.

*MTPret* contributes to generate this solution to achieve the above goals. It leverages a MoCo paradigm and a large number of unlabeled image datasets from different body parts to make the backbone learn the feature representation of X-rays. To break the limits of the pretext task and contrast self-supervised loss, and allow the backbone learn those task-related knowledge, *MTPret* adds a subpipeline for multitask continual learning. Specifically, it utilizes limited labeled data to train task-specific heads, such as FC layer, DeepLab-V3 [12], and FasterRCNN [14], to learn discriminative features for classification, segmentation, and bbox detection tasks, respectively. *MTPret* finally performs independent and separate fine-tuning of the pretrained backbone for each task.

### B. Multidataset Contrastive Learning

As shown in Fig. 1(a), *MTPret* adopts momentum contrastive learning (MoCo) algorithm on aggregated datasets to obtain an underlying pretrained backbone network, namely *MD-MoCo*. Given  $N$  X-ray datasets  $\{\mathcal{S} = \mathcal{S}_1, \dots, \mathcal{S}_N\}$ ,  $x_i \in \mathbb{R}^d$ , the goal of the MoCo task is to find a mapping function  $F: \mathbb{R}^d \mapsto \mathbb{R}^a$ ,  $a \ll d$  that satisfies

$$s(F(x), F(x^+)) \gg s(F(x), F(x^-)) \quad (1)$$

where the function  $s(\cdot, \cdot)$  measures image similarity,  $F$  is responsible for representation learning and dimension reduction.

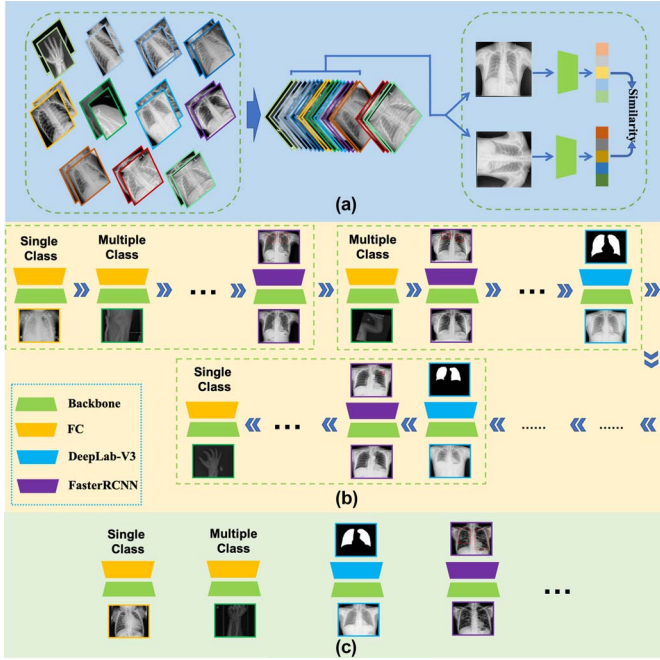


Fig. 1. Overview of *MTPret*, consists of three parts: (a) Multidataset momentum contrastive learning (MD-MoCo). (b) Multitask continual learning. (c) Fine-tuning on tasks.

Positive and negative samples are denoted as  $x^+$  and  $x^-$ , respectively, with  $x^+$  indicating similarity to  $x$  and  $x^-$  indicating dissimilarity. Especially, the model learns representations by maximizing agreement between differently augmented views  $x_i$  and  $x_j$  of the same example  $x$  using a contrastive loss in the latent space, and the augmented views  $x_i$  and  $x_j$  are generated from data augmentation  $\mathcal{DA}$ .

1) *Dictionary as a Queue*: MoCo trains an encoder to perform a dictionary lookup task, where a query  $q$  and encoded samples  $x_1, \dots, x_k$  serve as the keys in the dictionary. Therefore there is a match if the query  $q$  is similar to the positive sample  $x^+$ , and there is no match in the dictionary for those negative sample  $x^-$ . Specifically, MoCo employs two visual encoders, denoted as  $f_q$  and  $f_k$ , to learn query representations  $q = f_q(x_q)$  and key representations  $k = f_k(x_k)$ . Here,  $x_q$  represents the query sample, and  $x_k$  represents the key sample. To enable the encoder to reuse the previously encoded samples, MoCo uses the dictionary as a queue. The pretrained model is trained using a loss function defined as follows:

$$\mathcal{L}_{\text{Contra}} = -\log \frac{\exp(q, k^+) / \tau}{\exp(q, k^+) / \tau + \sum_{k^-} \exp(q, k^-) / \tau} \quad (2)$$

where the  $\tau$  is a temperature hyperparameter per [15], and  $k^+$  and  $k^-$  denote positive and negative samples, respectively.

2) *Momentum Update*: MoCo utilizes a momentum update strategy to update parameters of visual encoders. Denoting the parameters of  $f_q$  as  $\theta_q$  and those of  $f_k$  as  $\theta_k$ , MoCo updates  $\theta_q$  by back-propagation using contrastive loss  $\mathcal{L}_{\text{Contra}}$  in (2) and updates  $\theta_k$  by

$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q \quad (3)$$

where  $m \in [0, 1)$  is a momentum coefficient. This updating strategy makes the evolution of  $\theta_k$  smoother compared to  $\theta_q$  [16].

While MoCo has been optimized for natural images and has been extended for various tasks, there are two problems when transferring contrastive learning from natural images to multiple data sources X-ray images, including the *difference between natural images and X-ray images*, and the *inconsistency of X-ray images from different data sources*.

- 1) Compared to natural images, X-rays have larger gray scale and similar spatial structures across image, which are always either anterior-posterior, posterior-anterior, or lateral. To utilize contrastive learning for X-ray pretraining, *MTPret* optimized data augmentation strategies in contrastive learning. Specifically, certain augmentation techniques such as random cropping and Gaussian blurring may lead to a change in the disease label or cause confusion between different diseases. Similarly, color jittering and random grayscale do not offer meaningful enhancements for grayscale X-rays and hence are disabled by *MTPret*. This allows for the preservation of the semantic information in X-rays.
- 2) For medical images, the imaging process is different from that of natural images. Imaging parameters in this imaging process directly affect the quality of imaging, and result for the inconsistency of X-ray images from different data sources. To overcome these issues, *MTPret* employs certain image preprocessing techniques, such as normalizing the gray-scale distribution of these datasets using the Z-score method with the mean of 122.786 and a standard deviation of 18.390, and resizing the images to a consistent resolution of  $800 \times 500$ .

Algorithm 1 provides the pseudocode of multidataset contrastive learning (MD-MoCo) of *MTPret*.

### C. Multitask Continual Learning

As shown in Fig. 1(b), given the backbone pretrained by MD-MoCo, *MTPret* continues to learn discriminative features from different types of medical image analytics tasks based on a training pipeline of CL. In order to perform multitasks continual learning on the X-rays effectively, *MTPret* adopts the following training procedure and various task-specific training losses.

1) *Multitask Training Procedure*: Given a set of training tasks, including classification, segmentation, and detection, with X-rays collected from different body parts, *MTPret* trains a shared DNN backbone with various task-specific heads using the following training procedure to avoid *overfitting* and *catastrophic forgetting* in CL.

- 1) Repeating learning multiple times on a single task tends to causes the backbone overfits to the current task and thus lose its ability on other tasks. To avoid it, *MTPret* utilizes a strategy of shuffling the task order in each learning round and employs a cosine annealing learning rate schedule

$$\eta_t = \eta_{\min}^i + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left( 1 + \cos \left( \frac{t}{T} \cdot 2\pi \right) \right) \quad (4)$$

**Algorithm 1 Multi-Dataset MoCo**


---

**Require:** The X-ray datasets  $\mathcal{S}$ , the batch size  $N$ , backbone modules  $f_k$  and  $f_q$ , a head module  $g$ , the set of data augmentation functions  $\mathcal{DA}$ , the dictionary queue  $q_k$ , the temperature hyper-parameter  $\tau$ , the image resize parameters  $W, H$ , the parameters of mean value  $E[\mathcal{S}]$  and standard deviation  $\sigma(\mathcal{S})$  for Z-score normalization;

**Ensure:**  $q_k := k^+ \cup k^-$ ;

- 1: Normalize  $\mathcal{S}$  using Z-score:  $Z = \frac{\mathcal{S} - E[\mathcal{S}]}{\sigma(\mathcal{S})}$ ;
- 2: Re-size image resolution with  $W$  and  $H$ ;
- 3: **for** sampled mini-batch  $\{x_k\}_{k=1}^N$  from Datasets  $\mathcal{S}$  **do**
- 4:   **for**  $k \in \{1, \dots, N\}$  **do**
- 5:     Randomly select  $DA_1, DA_2$  from  $\mathcal{DA}$ ;
- 6:      $\hat{x}_{2k-1} = DA_1(x_k), \hat{x}_{2k} = DA_2(x_k)$ ;
- 7:      $h_{2k-1} = f_k(\hat{x}_{2k-1}), h_{2k} = f_q(\hat{x}_{2k})$ ;
- 8:      $z_{2k-1} = g(h_{2k-1}), z_{2k} = g(h_{2k})$ ;
- 9:   **end for**
- 10: **for**  $i \in \{1, \dots, 2N\}, j \in \{1, \dots, 2N\}$  **do**
- 11:    Calculate Similarity  $s_{i,j} = z_i^\top z_j / (\|z_i\| \|z_j\|)$ ;
- 12: **end for**
- 13: Update  $f_k$  to minimize Eq 2;
- 14: Momentum update  $f_q$  using Eq. 3;
- 15: Enqueue  $q_k$  with  $z_{2k-1}$ ;
- 16: Dequeue  $q_k$ ;
- 17: **end for**
- 18: **return**  $f_k$ ;

---

where  $\eta_t$  represents the learning rate at the  $t$ th iteration,  $\eta_{\max}$  and  $\eta_{\min}$  denote the maximum and minimum learning rates, and  $T$  is the total number of iterations in one cycle of the cyclic learning rate schedule.

- 2) Multitask continual learning faces the challenge of ‘‘catastrophic forgetting’’, where the backbone may ‘‘forget’’ the knowledge learned in previous iterations. To address this issue, *MTPret* employs a knowledge transfer regularization technique based on  $L^2$ -SP [11]

$$\Omega(\mathbf{w}) = \frac{\alpha}{2} \|\mathbf{w}_S - \mathbf{w}_S^0\|_2^2 + \frac{\beta}{2} \|\mathbf{w}_{\bar{S}}\|_2^2 \quad (5)$$

where the parameter vector  $\mathbf{w} \in \mathbb{R}^n$  consists of all the network parameters that need to be adjusted for the target task. Especially,  $\mathbf{w}^0$  is the parameter vector of the model pretrained on the source problem,  $\mathbf{w}_S$  is one for the part of the target network that shares the architecture of the source network, and  $\mathbf{w}_{\bar{S}}$  is the other one for the part that only exists in the target network. In addition,  $\alpha$  and  $\beta$  are the regularization parameter setting the strength of the penalty,  $\|\cdot\|_2$  is the  $\ell_2$ -norm of a vector. In each iteration, *MTPret* adopts regularization to constrain the distance between the current learning outcome and the feature extractor that was trained in previous iterations.

Algorithm 2 shows the pseudocode of multitask continual learning of *MTPret*.

2) *Task-Specific Training Losses:* Given multiple medical image analytics tasks denoted as  $\{\mathcal{T} = \mathcal{T}_1, \dots, \mathcal{T}_N\}$ , *MTPret*

**Algorithm 2 Multi-Task Continual Learning**


---

**Require:** The number of epochs  $eps$ , a queue of training tasks  $\mathcal{T}$ , a backbone module  $f_k$ , head modules  $g = \{g_{Clas}, g_{Seg}, g_{Det}\}$ , the schedule of learning rates  $lr = \{lr_1, \dots, lr_M\}$ , loss functions  $\mathcal{L} = \{\mathcal{L}_{clas}, \mathcal{L}_{seg}, \mathcal{L}_{det}\}$

- 1: **for**  $ep \in \{1, \dots, eps\}$  **do**
- 2:   Reshuffle task queue  $\mathcal{T}$ ;
- 3:   **for**  $\mathcal{T}_i \in \mathcal{T}$  **do**
- 4:      $h_{x^{(i)}} = f_k(x^{(i)})$
- 5:     Select head  $g_j$  subject to task  $\mathcal{T}_i$ ;
- 6:      $z_{x^{(i)}} = g_j(h_{x^{(i)}})$
- 7:     Select loss  $\mathcal{L}_j$  according to task  $\mathcal{T}_i$ ;
- 8:     Update backbone  $f_k$  to minimize  $\mathcal{L}_j$  and Eq 2
- 9:     Update learning rate  $lr_i$  using Eq 1
- 10:   **end for**
- 11: **end for**
- 12: **return**  $f_k$ ;

---

makes feature encoding with the same backbone pretrained by Section II-B, but makes decoding with different head for specific tasks, e.g., FC Layer for classification task, DeepLabV3 [12] for segmentation task and FasterRCNN [14] for bbox detection task. It allows the pretrained backbone learn these  $M$  tasks one-by-one, and adopts task-specific losses to optimize the backbone network.

a) *Classification task:* *MTPret* adopts *CrossEntropyLoss* as the loss function of classification task, which is defined as

$$\mathcal{L}_{Clas} = -\frac{1}{N_{Clas}} \sum_{i=1} (y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)) \quad (6)$$

where  $N_{Clas}$  is the mini-batch size,  $y$  and  $\hat{y}$  denote the ground-truth and the result that model predicted, respectively. In other experimental details, *MTPret* adopts *Sigmoid* as the activation function for the *binary classification task* and the *multilabel classification task*, but adopts *Softmax* for the *multiclass classification task*.

b) *Segmentation task:* *MTPret* adopts *DiceLoss* to processing segmentation tasks, this loss is defined as

$$\mathcal{L}_{Seg} = \frac{1}{N_{Seg}} \sum_{i=1} \left( 1 - \frac{2|\hat{Y}_i \cap Y_i| + \text{smooth}}{|\hat{Y}_i| + |Y_i| + \text{smooth}} \right) \quad (7)$$

where  $Y$  and  $\hat{Y}$  denote the actual ground-truth and the predicted mask by the model, respectively, and *smooth* is a smoothing hyperparameter.

c) *Bbox detection task:* For this task, *MTPret* adopts a compound loss consisting of two items, a classification loss  $\mathcal{L}_{clas}$  and a regression loss  $\mathcal{L}_{reg}$ . Where the  $\mathcal{L}_{clas}$  is similar to (6), and the  $\mathcal{L}_{reg}$  is a robust loss function (smooth  $L_1$ ) that defined as

$$\mathcal{L}_{reg} = |y_i - \hat{y}_i| \quad (8)$$

where  $\hat{y}$  represents the predicted bounding box’s four parameterized coordinates, and  $y$  represents those of the ground-truth

TABLE I  
OVERVIEW OF ELEVEN PUBLICLY AVAILABLE X-RAY IMAGE DATASETS

Datasets	Body Part	Task	Train	Validation	Test	Total
Only used for multidataset contrastive learning of <i>MTPret</i>						
China-set-CXR [17]	Chest	N/A	661	N/A	N/A	661
Montgomery-set-CXR [17]	Chest	N/A	138	N/A	N/A	138
Indiana-CXR [18]	Chest	N/A	7470	N/A	N/A	7470
RSNA bone age [19]	Hand	N/A	10 811	N/A	N/A	10 811
Used for multidataset contrastive learning or model validation of <i>MTPret</i>						
NIHCC [20]	Chest	N/A	112 120	N/A	N/A	112 120
	Chest	Multilabel classification of pathology	78 484	11 212	22 424	112 120
Used for all phase of <i>MTPret</i>						
Pneumonia [21]	Chest	Binary classification of pneumonia	4686	585	585	5856
MURA [22]	Various Bones	Binary classification of abnormal skeleton	32 013	3997	3995	40 005
Chest X-ray... [17]	Chest	Segmentation of lung	718	89	89	896
TBX [23]	Chest	Detection of tuberculosis	640	80	80	800
Only used for external validation of <i>MTPret</i>						
CheXpert [24]	Chest	Multilabel classification of pathology	223 414	234	N/A	223 648
DeepCovid [2]	Chest	Binary classification of COVID	2084	3100	N/A	5184
<b>Total</b>	N/A	N/A	N/A	N/A	N/A	424 746

box associated with a positive anchor. Finally the detection loss is defined as follows:

$$\mathcal{L}_{\text{Det}} = \frac{1}{N_{\text{clas}}} \sum_{i=1} L_{\text{clas}}(p_i, \hat{p}_i) + \lambda \frac{1}{N_{\text{reg}}} \sum_{i=1} p_i L_{\text{reg}}(y_i, \hat{y}_i) \quad (9)$$

where  $\hat{p}_i$  represents the predicted probability that anchor  $i$  corresponds to an object. The true label  $p_i$  for the anchor is 1 if it is positive and 0 if it is negative. The regression loss term  $p_i L_{\text{reg}}$  is applied exclusively to positive anchors ( $p_i = 1$ ), while it is disregarded for negative anchors ( $p_i = 0$ ).

#### D. Fine-Tuning on Tasks

Finally, after pretraining the backbone using the two stages mentioned above, *MTPret* combines the backbone with FC Layer for classification tasks, DeepLab-V3 [12] for segmentation tasks, and FasterRCNN [14] for bbox detection tasks, then fine-tunes the whole model on each task independent and separately, as shown in Fig. 1(c).

### III. EXPERIMENT AND EVALUATION RESULTS

In this section, we detail the design of our experiments, including data collection and data distribution, some baselines algorithm and other initialization settings, detailed setup of the experiment and comparison with recent works, then we present the results of our algorithms in comparisons with other baselines.

#### A. Datasets Collection and Preparation

Table I reports the datasets collection used in this study. A large set of X-ray images are collected from eleven open source datasets, including China-Set-CXR [17], Montgomery-Set-CXR [17], Indiana-CXR [18], RSNA Bone Age [19], NIHCC [20], Pneumonia [21], MURA [22], Chest X-ray Masks

and Labels [17], TBX [23], CheXpert [24], and Deep-Covid [2]. A total of 424 746 X-rays cover several parts of the human body, including the chest, hand, elbow, finger, forearm, humerus, shoulder, and wrist. Based on these datasets, *MTPret* developed a rigorous experimental plan to use these data for learning and validation.

*MTPret* first puts all data from China-Set-CXR [17], Montgomery-Set-CXR [17], Indiana-CXR [18], RSNA Bone Age [19], NIHCC [20] and partial data from Pneumonia [21], MURA [22], Chest X-ray Masks and Labels [17], TBX [23] to the Section II-B. It is worth mentioning that contrastive learning, as an unsupervised network, does not require any labeling information at that phase. Then *MTPret* carefully selects some X-ray analytics tasks from these datasets to put in the Section II-C, these task include *pneumonia classification task* on Pneumonia [21], *skeletal abnormality classification task* on MURA [22], *lung segmentation task* on Chest X-ray Masks and Labels [17], and *tuberculosis bbox detection task* on TBX [23]. To avoid data leakage, for these four datasets, *MTPret* uses the same data and data corresponding label information in this phase as in the previous phase. (Note: Some of X-rays from these four datasets were also used in the first phase, and both phases have remaining data that have not been learned by the model and are available for subsequent model validation.) Finally, *MTPret* fine-tunes the backbone by utilizing the pre-trained weights as the initial point, and adapting its own task-specific head to separately fit each of the seven tasks, including four tasks that appeared in the previous Section II-C, *pulmonary disease classification tasks* on NIHCC [20] and CheXpert [24], and *COVID classification task* on Deep-Covid [2].

#### B. Baselines Algorithms

We evaluated the performance of *MTPret* on above seven tasks and compare with initialization strategies as follows.

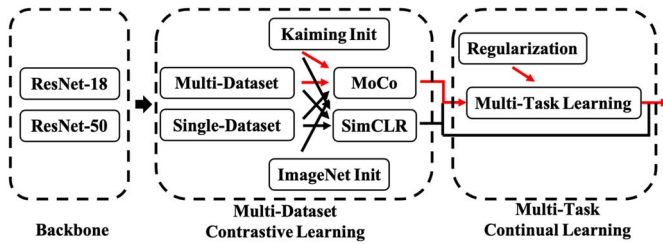


Fig. 2. Pipeline shared by *MTPret* and baselines, where the red arrows indicate *MTPret*.

- 1) *Scratch*: the models are all initialized using Kaiming’s random initialization [25] and fine-tuned on the target datasets.
- 2) *ImageNet*: the models are initialized with the weights pretrained on the ImageNet [26] dataset that are officially released, and then fine-tuned on the target datasets.

In addition to the above initialization strategies, we also use each individual step of *MTPret* and their variants/alternatives as baseline algorithms. These baseline models undergo a detailed validation process by progressively modifying or removing relevant modules of *MTPret*. This stepwise approach aims to assess the significance of each individual step in *MTPret*. Fig. 2 presents the pipelines shared by all baseline algorithms and *MTPret*, where the combinations of different modules configure the above algorithms.

- 1) *MD-MoCo*: The models undergo pretraining using *multi-dataset MoCo* with initialization from Kaiming’s method [25], and then fine-tuned accordingly;
- 2) *SD-MoCo*: It changes *multidataset MoCo* to *single-dataset MoCo* in *MD-MoCo* to provide proof that the backbone is able to learn more favorable feature representations of X-rays from multiple data sources;
- 3) *MD-MoCo<sup>++</sup>*: It changes Kaiming’s initialization [25] to ImageNet initialization [26] in *MD-MoCo*, this trick observes the proposal of [4];
- 4) *SD-SimCLR* and *MD-SimCLR*: It changes MoCo [16] to SimCLR [27], another contrastive self-supervised solution that has performed well on ImageNet [26]. On the one hand, this set of control experiments can search a more superior contrastive learning strategy on X-ray images, and on the other hand, it can reaffirm the superiority of multidatasets pretraining; and
- 5) *MTPret<sup>−−</sup>*: All models are pretrained and fine-tuned with *MTPret* but without the use of *cross-task memorization* and *cyclic and reshuffled learning schedule*.

To ensure a fair comparison, all baseline models employ identical training/evaluation code and experimental hyperparameters. During the pretraining phase, only SD-SimCLR and SD-MoCo exclusively utilize the NIHCC dataset, which constitutes the largest portion of the pretraining dataset, accounting for 85.46% of the overall pretraining images. All other baseline models utilize the entire pretraining dataset. For a more detailed overview of the pretraining dataset utilization plan, refer to Table. I.

### C. Experiment Setup

All of the above experimental designs evaluated *MTPret* on two backbone including *ResNet-18* and *ResNet-50*. To compare those baseline pretrain algorithms or initialization settings, *MTPret* adopts corresponding metrics for different tasks, including *Area under the Curve* (AUC) and *Accuracy* (Acc.) for binary classification tasks, *mean Area under the Curve* (mAUC) and *Area under the Curve of Single Label* (AUC) for multilabel classification tasks, *Dice Similarity Coefficient* (Dice) and *mean Intersection over Union* (mIoU) for segmentation task, *mean Average Precision* (mAP), and *Average Precision of Single Target* (AP, at the IoU threshold of 0.5) for bbox detection (TBX). In this experiment, *MTPret* tunes optimal hyperparameters by evaluating performance on validation datasets using a *main metric* indicated in bold, and presents the results on the testing datasets.

### D. Overall Results

We compare *MTPret* with models trained using every single dataset in an end-to-end manner, as well as other pretraining/initialization strategies. Table II presents the performance of *MTPret* on seven datasets with the major performance metrics. From it we can see that *MTPret* achieves the best performance in most results on these tasks, which is proven by both two backbone networks (*ResNet-18* and *ResNet-50*). Only on the Deep-Covid [2] dataset, with setting of *ResNet-18*, *MTPret* is only 0.01% lower than MD-MoCo. This result demonstrates that *MTPret* perform robustly and can be adapted to various medical image analytics tasks with decent performance.

### E. Ablation Studies

1) *Effect of Multidataset Pretraining*: By comparing SD-SimCLR and MD-SimCLR, as well as SD-MoCo and MD-MoCo, we observe that our multidataset pretraining strategy outperforms the single-dataset pretraining approach. For instance, in all classification tasks using *ResNet-50*, SD-SimCLR achieves an average AUC of 86.72%, while SD-MoCo achieves 90.32%. On the other hand, MD-SimCLR achieves an average AUC of 86.99% (**0.27%↑**), and MD-MoCo achieves 90.75% (**0.43%↑**). This demonstrates the efficacy of introducing image representations from various body parts in improving the model’s expressive capability during multidataset pretraining. Similar conclusions are drawn in experiments with *ResNet-18* setting. In segmentation and detection tasks, specifically when utilizing *ResNet-18* as the backbone model, MD-SimCLR exhibits a performance decrease in segmentation compared to MD-MoCo, with a DiCe score of 95.01% (**0.30%↓**).

2) *Comparison Between SimCLR and MoCo*: By comparing SD-SimCLR with SD-MoCo and MD-SimCLR with MD-MoCo, we analyze the performance differences between the MoCo and SimCLR contrastive learning strategies in the context of medical image pretraining tasks. Across all classification tasks using *ResNet-50*, SD-MoCo achieves a performance improvement of 3.60% AUC compared to SD-SimCLR. Similarly,

TABLE II  
OVERVIEW RESULTS FOR SEVEN TASKS USING VARIOUS PRETRAINING ALGORITHMS

Datasets	Metrics	Pretrain								
		Scratch	ImageNet	SD-SimCLR	MD-SimCLR	SD-MoCo	MD-MoCo	MD-MoCo <sup>++</sup>	<i>MTPret</i> <sup>--</sup>	<i>MTPret</i>
ResNet-18										
NIHCC [20]	mAUC	71.00	70.84	74.68	75.58	76.52	77.97	78.02	75.27	<b>79.05</b>
Pneumonia [21]	AUC	96.59	96.16	98.44	96.73	98.46	98.49	99.39	99.51	<b>99.60</b>
MURA [22]	AUC	86.66	88.10	87.19	87.37	88.06	88.29	88.15	88.40	<b>88.52</b>
Chest X-ray... [17]	Dice	95.24	95.26	95.18	95.01	95.29	95.31	95.25	95.14	<b>95.37</b>
TBX [23]	mAP	30.71	29.46	31.92	36.36	34.27	36.00	35.95	34.70	<b>36.71</b>
CheXpert [24]	AUC	82.90	82.79	85.76	85.81	88.11	88.18	87.43	87.45	<b>88.52</b>
DeepCovid [2]	AUC	96.80	98.98	99.11	99.71	99.64	<b>99.95</b>	99.92	99.89	99.94
ResNet-50										
NIHCC [20]	mAUC	66.05	72.36	77.62	78.05	77.83	79.31	79.74	77.20	<b>80.06</b>
Pneumonia [21]	AUC	96.58	98.75	91.80	92.14	99.29	99.52	99.49	99.58	<b>99.72</b>
MURA [22]	AUC	86.24	87.92	87.32	86.64	87.70	87.95	87.99	87.15	<b>88.41</b>
Chest X-ray... [17]	Dice	93.52	94.08	94.26	94.38	94.26	94.33	95.05	95.04	<b>95.27</b>
TBX [23]	mAP	23.93	35.61	33.26	35.25	33.28	36.78	36.52	35.14	<b>37.83</b>
CheXpert [24]	AUC	77.57	79.26	77.95	79.66	87.40	87.23	88.73	88.61	<b>89.22</b>
DeepCovid [2]	AUC	98.13	98.91	98.93	98.44	99.38	99.76	99.75	99.71	<b>99.91</b>

Note: Bold indicate the best results among all methods.

MD-MoCo outperforms MD-SimCLR by 3.77% AUC. This indicates that MoCo exhibits a stronger capability for visual feature representation compared to SimCLR in this medical image pretraining task. Similar experimental results are observed in segmentation or detection tasks. Likewise, the conclusion holds true when ResNet-18 is used as the backbone model.

3) *Effect of Multitask Pretraining and Cross-Task Memorization*: By observing the experimental results of MD-MoCo and *MTPret*<sup>--</sup>, we can discern the effectiveness of multitask learning. We summarize the experimental results based on the average performance in classification tasks, and similar conclusions can be observed in segmentation or detection tasks. In fact, after introducing multitask learning, *MTPret*<sup>--</sup> shows a performance decrease in most cases. For instance, based on ResNet-18, *MTPret*<sup>--</sup> experiences an average AUC of **0.48%↓**, and based on ResNet-50, it encounters an average AUC of **0.30%↓**. This decline is attributed to the lack of effective handling of catastrophic forgetting during the introduction of cross-task learning, leading to a decrease in the model’s domain generalization capability.

However, when we introduce *cross-task memorization* and a *cyclic and reshuffled learning schedule* to address catastrophic forgetting, representing the model as *MTPret*, we observe that compared to MD-MoCo, *MTPret* achieves an average AUC of 91.13% (**0.55%↑**) on ResNet-18 and an average AUC of 91.46% (**0.71%↑**) on ResNet-50. This underscores the potential brought by multitask learning and validates the effectiveness of our advanced techniques in addressing catastrophic forgetting.

4) *Comparison of Different Parameter Initialization*: We compare the experimental results of MD-MoCo and MD-MoCo<sup>++</sup> to examine the impact of different parameter initialization strategies on pretraining. The results show that in the classification task, based on ResNet-18, MD-MoCo and MD-MoCo<sup>++</sup> achieve identical performance with an average AUC of 90.58%. However, based on ResNet-50, MD-MoCo<sup>++</sup> surpasses MD-MoCo, achieving an average AUC of 91.14% (**0.39%↑**). Nevertheless, in segmentation and detection tasks, such as based on ResNet-18, MD-MoCo<sup>++</sup> performs worse than MD-MoCo, achieving a Dice of 95.25% (**0.06%↓**) and mAP of 35.95% (**0.05%↓**). This suggests that the performance

differences brought about by different initialization strategies are often unpredictable, and although they may lead to performance improvements in certain tasks, the improvement is extremely limited. This phenomenon has also been observed in recent studies [28], supporting similar conclusions. We attribute this to the absence of any catastrophic forgetting strategy during the contrastive learning process, resulting in the model completely forgetting the rich image representations learned from ImageNet.

#### F. Case Studies

In this section, we show in detail the performance of *MTPret* in every dataset/task. In Section III-F1 we present the performance of *MTPret* on multilabel classification tasks, which details its performance of all classification and performance of single classification. In Section III-F2 we analyze the performance of *MTPret* on X-rays from different body parts. In addition, using these two tasks as examples (similar conclusions can be drawn from other tasks), we explore the contribution of each module of *MTPret* in detail. In Section III-F3 we show the performance of *MTPret* on various tasks, specifically in addition to the previous classification task, we analyze the performance of *MTPret* on segmentation task and bbox detection task. Finally in Section III-F4 we perform two external validations of the *MTPret* to evaluate its performance on tasks that are not learned during the pretrain phase. Besides, we explore the possibility of combining *MTPret* with other models.

1) *MTPret on NIHCC*: *MTPret* was initially developed on the NIHCC [20], the dataset contains chest X-ray images of patients suffering from one or more of 14 diseases. *MTPret* was evaluated on testing set as well as some baselines and performances of the ablation experiments mentioned above. All results are summarized below and an overview is given in Fig. 3, which shows the performance on a single disease and the average performance on all diseases with various pre-trained algorithms.

As we can see in Fig. 3, the performance of *MTPret* is excellent and robust both on average and in terms of single-class disease prediction. Specifically, *MTPret* not only achieved the

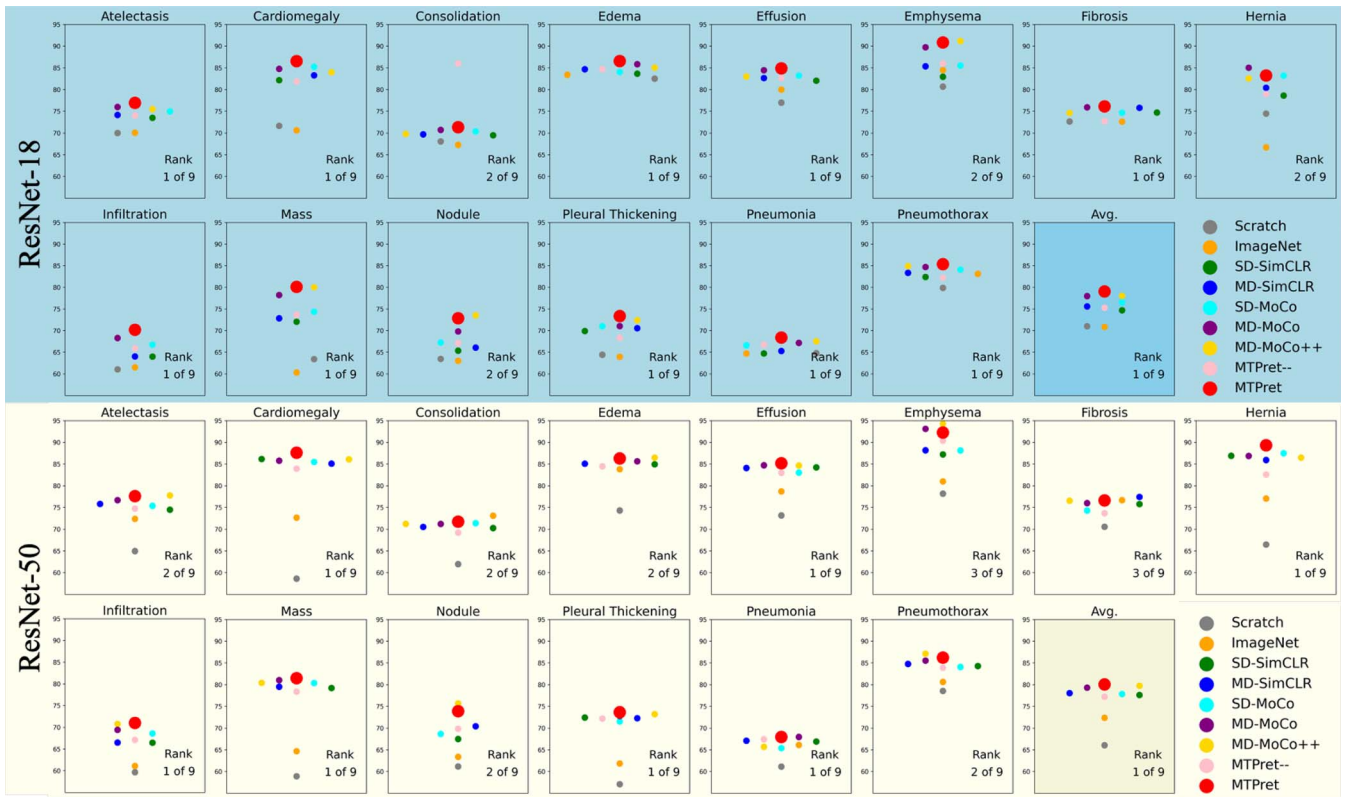


Fig. 3. Summary of *MTPret* performance on the testing sets on NIHCC [20]. AUC values are used as metric. The performance of all baselines and *MTPret* based on two backbones are plotted.

first place in average AUC on both backbone, but it also excelled in the prediction of single diseases. For example, on ResNet-18, *MTPret* achieved the best performance on all ten single disease predictions, finishing second on only four diseases (consolidation, emphysema, hernia, and nodule). In addition, *MTPret* achieved a significant performance improvement compared to other pretraining algorithms, such as Scratch on ResNet-18 (14.01%↓) and on ResNet-50 (7.70%↓), ImageNet on ResNet-18 (8.21%↓), and on ResNet-50 (7.70%↓).

2) *MTPret* on Pneumonia and MURA: The Pneumonia dataset [21] contains chest X-ray images of patients diagnosed with pneumonia and those of normal individuals, and MURA [22] is a bone X-rays dataset consisting of skeletal abnormalities and normal bones. For this two binary classification task with various body parts, the performances of *MTPret* and other solutions on testing set are shown in Fig. 4.

As shown in Fig. 4 we can see improvement of *MTPret* at each step. From Scratch to SD-MoCo, to MD-MoCo, to *MTPret*, all results show the contribution of contrastive learning with multidataset and continual learning with multitasks. For contrastive learning algorithm, all four sets of experiments show that MoCo outperforms SimCLR for both SD and MD. For single or multiple dataset contrastive learning, all experiment result show that MD outperforms SD using MoCo solution. For parameter initialization tips of model, compare MD-MoCo and MD-MoCo<sup>++</sup>, we can see similar scores, which means that the parameters of the backbone network are updated substantially in the contrastive learning to the extent that they are independent of the initial parameter settings. For multitask learning,

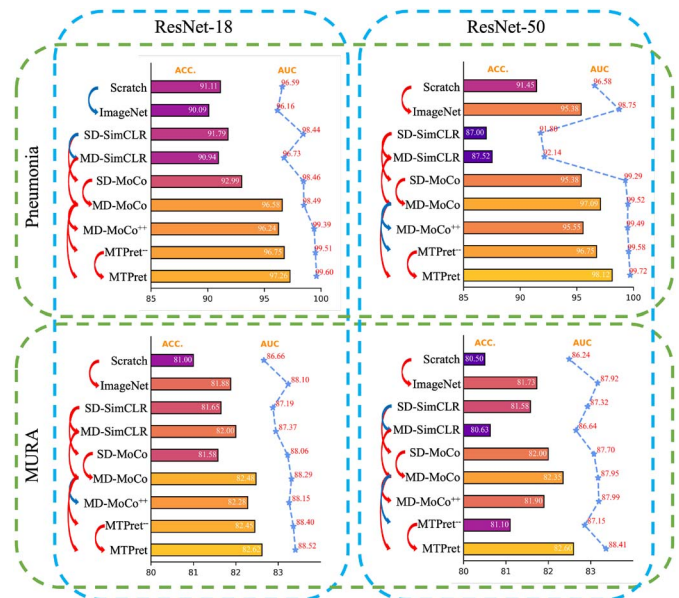


Fig. 4. Ablation Studies of *MTPret* performance on the testing sets on Pneumonia [21] and MURA [22], where Acc. and AUC values are used as metric. Performance of *MTPret* and all baselines on two backbone networks are plotted for ablation studies. On the vertical axis of each subplot, red arrows indicate AUC improvements from one to another and arrows in blue show performance degradation.

*MTPret* optimized with regularization constraints and learning rate scheduler performs consistently and outperforms MD-MoCo and *MTPret*<sup>++</sup> in all experiments. Finally, for multiple



Original Image	ImageNet	MTPret	Original Image	ImageNet	MTPret
	 91.99	 99.80		 92.66	 99.29
	 80.40	 39.16		 86.68	 33.88
	 45.29	 98.81		 47.40	 98.45

Fig. 5. Six X-ray examples with abnormal skeletons, each including the original image and the attention maps extracted from the ImageNet and *MTPret* pretrained DNNs. The number below each attention map is the probability that the model predicts the presence of abnormal bones in that image. The blue numbers indicate incorrect prediction and red numbers indicate correct prediction.

body parts, *MTPret* exceeds those baselines, such as scratch and ImageNet, on any body parts, which is proven by both two backbone networks.

In addition to quantitative analysis, we also visualize the attention maps of DNN under various settings for the MURA [22] task, as an example to explain how DNN models identify pathologies in X-ray images and interpret misclassification. Examples of identified cases are shown in Fig. 5, which includes the original input images, output heatmaps highlighting regions of high importance, detected high-risk areas, and probability scores for abnormal skeletons. As shown in the first row of Fig. 5, *MTPret* can correctly identify whether the bone is abnormal in X-ray with higher confidence scores than ImageNet. The detection results are also consistent with the location of the skeletal abnormality. Some inaccurate cases are shown in the second row (*MTPret* recognition error) and the third row (ImageNet recognition error) of Fig. 5. The cause of recognition error is often that the model does not find the lesion or the lesion that model found is offset. Based on the observed phenomenon in the heatmaps, it is possible that the model incorrectly identified the entire lung region as a potentially abnormal area, resulting in a failed identification.

Some lung segmentation results of cases are plotted in Fig. 6, from it we see that, on the lung segmentation task, *MTPret* progressively optimizes the results of segmentation by each proposed module, especially on the top and bottom boundaries of the lung.

3) *MTPret* on Lung and TBX: In the previous section, we discussed the performance of *MTPret* on multi body part X-rays. But both of them are limited to classification tasks, next, we invited *MTPret* to evaluate on others, such as lung segmentation task on Chest X-ray Masks and Labels [17] and tuberculosis bbox detection task on TBX [23].

Table III presents the results obtained with different pretraining algorithms. On the lung segmentation task, *MTPret* obtained the highest scores for both the Dice and mIoU metrics, and on the TB detection task, *MTPret* achieved the first place on

mAP. An interesting finding is that MD-MoCo achieved the best results for  $AP_{Active}$  on both backbones, while SimCLR achieved the best results for  $AP_{Latent}$ . Although *MTPret* did not obtain the best detection results for individual targets, it demonstrates stability and reliability in global senses.

4) *MTPret* on *MTPret* for *CheXpert* and *Deep-Covid*: We here invited *CheXpert* [24] and *Deep-Covid* [24] as external validation, which are out of distribution of pretraining datasets. *CheXpert* [24] is a large chest X-rays dataset including 14 diseases and *Deep-Covid* [24] is composed of COVID-19 X-rays. Actually, for all above experiments, either all or part of training data has been used for pretraining the backbones. In this experiments, we hope to use DNN backbones pretrained by *MTPret* to handle external tasks, where these datasets were not used in pretraining.

Figs. 7 and 8 show the results of various pretraining algorithms for Acc. and AUC on *Deep-Covid*, respectively. From them we can see that on the Acc. metric, compared to other methods, *MTPret* achieves better performance on both backbone, and on the AUC metric with ResNet-50, *MTPret* outperforms other pretrained algorithms. On the AUC with ResNet-18, *MTPret* achieves the second place and is only 0.01 lower than the first place MD-MoCo. These results indicate that *MTPret* is robust and can transfer to a new task that were not learned during the pretraining phase easily.

Table IV lists the AUC and average AUC results for the five major lung diseases on the *CheXpert* [24] dataset for various pretraining algorithms. The first and second place pretraining algorithms are marked in red font and green font. Compared to other pretraining algorithms, *MTPret* still achieves optimal performance on the overall average results, and for individual diseases, it can also achieve a competitive result. In addition, we can learn that MoCo-based solutions always achieve competitive results on this task, and they always achieve first or second place results for several cases. This result indicates that MoCo-based solution can provide robust performance for transfer of pretrained parameters, affirming the correctness of choosing MoCo in the context of *MTPret*.

### G. More Comparisons With Results of Other Literature

1) *Comparisons With Recent Works Based on the Same Datasets*: While *MTPret* was not specifically designed for any particular medical imaging task, it outperforms many recent works that use the same datasets in terms of overall performance. For NIHCC datasets, [29] also proposed novel pretraining algorithms based on grayscale ImageNet using InceptionV3 as the backbone, and they reported a 77.06% (2.94%↓) average AUC for the same task. For other tasks, [30] reported a 93.73% (4.42%↓) accuracy for the same pneumonia classification task, [31] reported a 94.64% Dice (0.73%↓) for the lung segmentation. [32] reported an AUC of 82.45% (6.07%↓) for MURA [22] dataset, and [23] reported a 58.70% (4.76%↓)  $AP_{Active}$  and a 9.60% (2.61%↓)  $AP_{Latent}$  for TBX [23] based on FasterRNN. For *Deep-Covid* datasets, [2] similarly employed ResNet-18 and ResNet-50 for the Covid classification task, reporting an AUC value of 98.90% (1.04%↓) for ResNet-18 and 99.00%

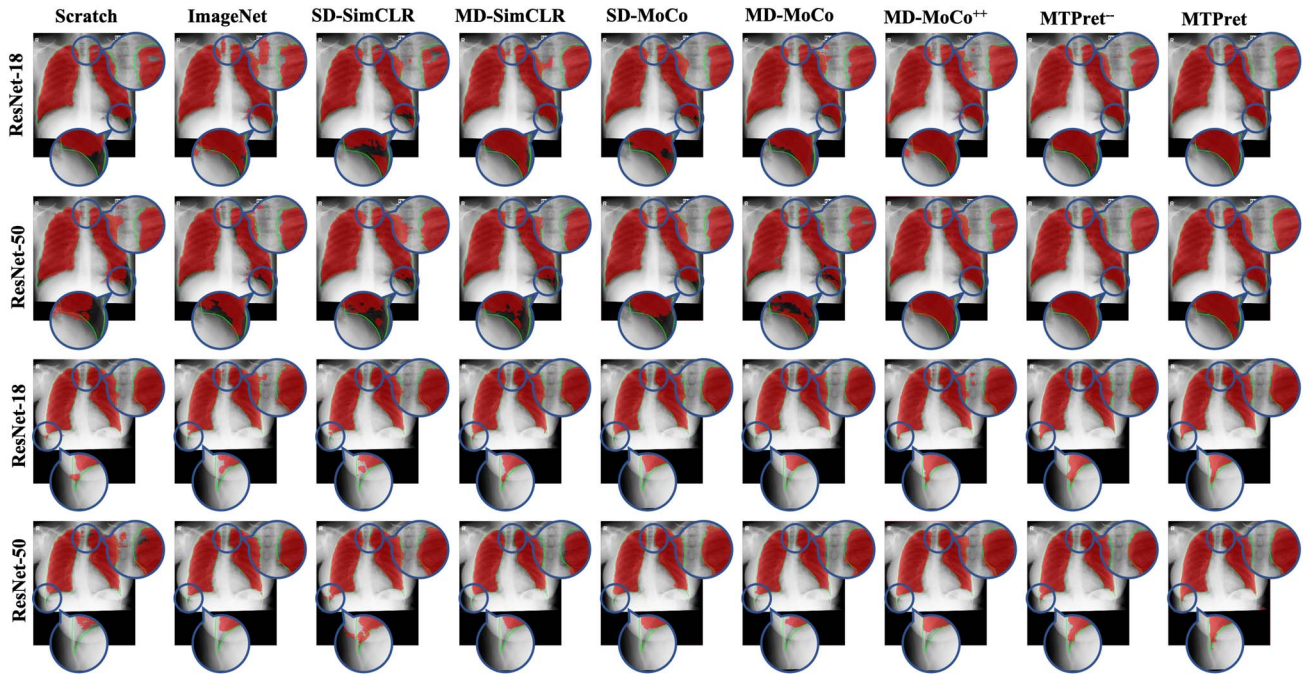


Fig. 6. Two cases of the segmentation results of various pretraining algorithms are shown. The green lines plot the ground truth and the red area is the model prediction results, and each result is zoomed in on the main differences between the models using the blue circles.

TABLE III  
PERFORMANCE COMPARISONS FOR LUNG SEGMENTATION (LUNG)  
AND TB DETECTION (TBX) USING VARIOUS PRETRAINING  
ALGORITHMS

Pretrain	Lung		TBX		
	Dice	mIoU	mAP	AP <sub>Active</sub>	AP <sub>Latent</sub>
ResNet-18					
Scratch	95.24	94.00	30.71	56.71	4.72
ImageNet	95.26	94.10	29.46	56.27	2.66
SD-SimCLR	95.18	93.97	31.92	50.69	<b>13.14</b>
MD-SimCLR	95.01	93.77	36.36	66.40	6.31
SD-MoCo	95.29	94.10	34.27	56.29	12.24
MD-MoCo	95.31	94.14	36.00	<b>67.17</b>	4.84
MD-MoCo <sup>++</sup>	95.25	94.03	35.95	65.19	6.70
MTPret <sup>--</sup>	95.14	93.90	34.70	63.43	5.97
MTPret	<b>95.37</b>	<b>94.22</b>	<b>36.71</b>	64.84	8.59
ResNet-50					
Scratch	93.52	92.03	23.93	44.85	3.01
ImageNet	94.08	92.65	35.61	58.81	12.42
SD-SimCLR	94.26	92.88	33.26	57.89	8.62
MD-SimCLR	94.38	93.00	35.25	57.15	<b>13.36</b>
SD-MoCo	94.26	92.86	33.28	62.23	4.32
MD-MoCo	94.33	93.04	36.78	<b>64.37</b>	9.18
MD-MoCo <sup>++</sup>	95.05	93.79	36.52	62.22	10.82
MTPret <sup>--</sup>	95.04	93.82	35.14	57.32	12.97
MTPret	<b>95.27</b>	<b>94.10</b>	<b>37.83</b>	63.46	12.21

Note: Bold indicate the best results among all methods.

(0.91%↓) for ResNet-50. In addition, for CheXpert datasets, [33] proposed deep AUC maximization (DAM) to achieve the best performance on this dataset. We conducted detailed experiments to substantiate the superiority of *MTPret*, and the experimental results are showcased in Table V. The findings reveal that *MTPret* surpasses DAM [33], demonstrating its robust model domain generalization capability. The benefits of *MTPret* showcase the possibility of utilizing SSL and CL for pretraining the backbone with multiple tasks.

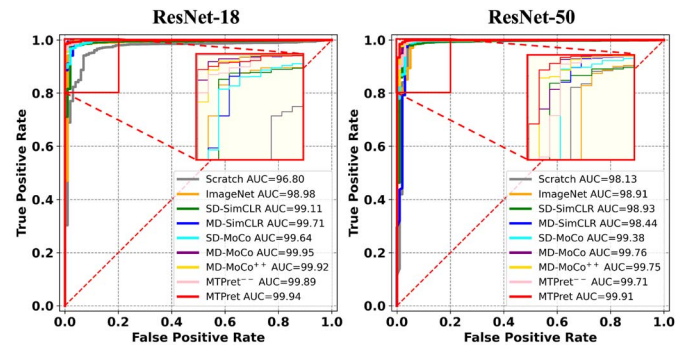


Fig. 7. Performance (AUC) comparisons for Deep-Covid using various pretraining algorithm, the most important areas of the ROC curve are shown in the black rectangle to see the differences in the results of each algorithm.

2) *Comparisons With Self-Supervised Learning*: To further demonstrate domain generalization ability of *MTPret* in self-supervised learning, we selected SimCLR [27], MoCo [16], SwAV [34], and MoCo-CXR [4] as baseline models for comparison. All these baseline models utilized ResNet-50 as the backbone for self-supervised pretraining on visual images. Some works extended the efforts to ResNet-18, such as SimCLR [27] and MoCo-CXR [4]. We conducted experiments on two datasets, CheXpert [24] and Deep-Covid [2], as these datasets were not used in the pretraining of *MTPret*. The experimental results are presented in Tables VI and VII. The results show that *MTPret* achieved the best performance on all four metrics in Deep-Covid. On the CheXpert dataset, *MTPret* outperformed other self-supervised learning algorithms on average and demonstrated superior performance for specific diseases in most cases. In this experimental outcome, MoCo-CXR [4],

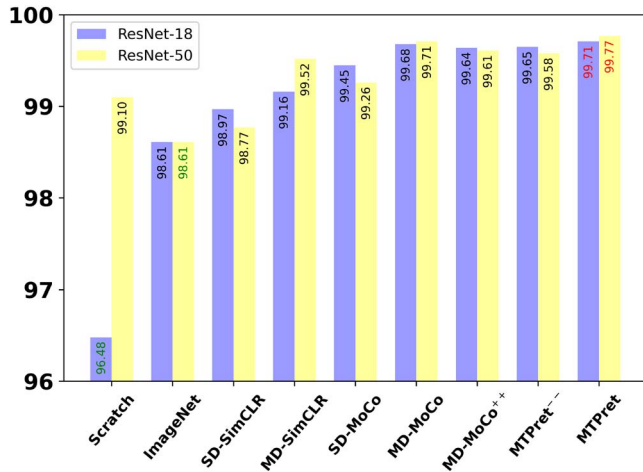


Fig. 8. Performance (Acc.) comparisons for Deep-Covid using various pretraining algorithm. the red numbers and green numbers indicate the results of the best and worst.

TABLE IV  
PERFORMANCE (AUC) COMPARISONS FOR CHEXPert USING VARIOUS PRETRAINING ALGORITHMS

Pre-train	CheXpert					
	Atel.	Cardi.	Consol.	Edema	Pleural.	Average
ResNet-18						
Scratch	79.73	77.12	87.98	86.29	83.40	82.90
ImageNet	77.55	77.93	89.28	86.76	82.45	82.79
SD-SimCLR	81.73	82.84	89.65	87.66	86.91	85.76
MD-SimCLR	81.69	82.91	89.45	88.13	86.90	85.81
SD-MoCo	<b>83.07</b>	83.91	<b>92.11</b>	91.47	90.00	88.11
MD-MoCo	81.74	<b>84.35</b>	90.90	<b>92.59</b>	<b>91.30</b>	<b>88.18</b>
MD-MoCo <sup>++</sup>	79.94	80.85	91.60	<b>93.48</b>	<b>91.26</b>	87.43
MTPret <sup>--</sup>	82.24	82.43	91.31	91.77	89.50	87.45
MTPret	<b>82.37</b>	<b>84.92</b>	<b>93.22</b>	91.68	90.42	<b>88.52</b>
ResNet-50						
Scratch	73.43	71.39	83.99	82.90	76.15	77.57
ImageNet	73.38	75.47	87.67	80.42	79.39	79.26
SD-SimCLR	73.68	73.92	83.13	83.48	75.54	77.95
MD-SimCLR	71.49	75.08	88.46	82.47	80.80	79.66
SD-MoCo	<b>83.34</b>	<b>87.30</b>	89.06	88.65	88.65	87.40
MD-MoCo	<b>82.78</b>	84.77	90.48	89.48	88.64	87.23
MD-MoCo <sup>++</sup>	81.28	85.71	90.85	<b>93.94</b>	<b>91.88</b>	<b>88.73</b>
MTPret <sup>--</sup>	82.04	<b>86.25</b>	<b>92.30</b>	91.68	90.76	88.61
MTPret	82.57	84.89	<b>93.38</b>	<b>92.10</b>	<b>93.15</b>	<b>89.22</b>

Note: Atel.: Atelectasis, Cardi.: Cardiomegaly, Consol.: Consolidation, and Pleural.: Pleural Effusion.

based on the MoCo algorithm for pretraining on X-ray images, exhibited inferior performance compared to *MTPret*. This is attributed to *MTPret*'s further extension of multitask pretraining on top of MoCo, leading to enhanced representation learning for X-ray images.

#### IV. RELATED WORKS AND DISCUSSIONS

In this section, we present the studies that are relevant to our work, and discuss several open issues in this work.

##### A. Deep Learning for Medical Image Analysis

Deep learning [21] has made tremendous success in medical image analytics, such as image classification and segmentation tasks for diseases in bones, lungs and heads with X-rays. It

TABLE V  
PERFORMANCE (AUC) COMPARISONS FOR CHEXPert OF *MTPRET* AND DAM [33]

Methods	CheXpert					
	Atel.	Cardi.	Consol.	Edema	Pleural.	Average
DenseNet-121						
DAM [33]	82.52	<b>87.51</b>	88.27	<b>92.87</b>	91.51	88.54
ResNet-18						
DAM [33]	77.55	77.93	89.28	86.76	82.45	82.79
MTPret	82.37	84.92	93.22	91.68	90.42	88.52
ResNet-50						
DAM [33]	73.38	75.47	87.67	80.42	79.39	79.26
MTPret	<b>82.57</b>	84.89	<b>93.38</b>	92.10	<b>93.15</b>	<b>89.22</b>

Note: Atel.: Atelectasis, Cardi.: Cardiomegaly, Consol.: Consolidation, and Pleural.: Pleural Effusion. Bold indicate the best results among all methods.

TABLE VI  
PERFORMANCE COMPARISONS FOR DEEP-COVID OF *MTPRET* AND OTHER SELF-SUPERVISED LEARNING (SSL) ALGORITHMS

SSL	Deep-Covid			
	Accuracy	Specificity	Sensitivity	AUC
ResNet-18				
SimCLR [27]	99.67	89.23	<b>99.90</b>	99.02
MoCo-CXR [4]	97.81	93.85	97.90	99.48
MTPret	<b>99.71</b>	<b>94.00</b>	<b>99.90</b>	<b>99.94</b>
ResNet-50				
SimCLR [27]	98.43	89.23	98.63	99.37
MoCo [16]	<b>99.77</b>	89.23	<b>100.00</b>	99.42
SwAV [34]	99.74	87.69	<b>100.00</b>	98.90
MoCo-CXR [4]	<b>99.77</b>	89.23	<b>100.00</b>	99.70
MTPret	<b>99.77</b>	<b>93.00</b>	<b>100.00</b>	<b>99.91</b>

Note: Bold indicate the best results among all methods.

TABLE VII  
PERFORMANCE COMPARISONS FOR CHEXPert OF *MTPRET* AND OTHER SELF-SUPERVISED LEARNING (SSL) ALGORITHMS

SSL	CheXpert					
	Atel.	Cardi.	Consol.	Edema	Pleural.	Average
ResNet-18						
SimCLR [27]	80.25	84.48	92.40	86.23	90.17	86.71
MoCo-CXR [4]	<b>84.20</b>	81.42	92.61	89.02	<b>91.54</b>	87.76
MTPret	82.37	<b>84.92</b>	<b>93.22</b>	<b>91.68</b>	90.42	<b>88.52</b>
ResNet-50						
SimCLR [27]	79.59	<b>86.47</b>	91.74	86.92	90.57	87.07
MoCo [16]	83.86	79.79	89.08	87.36	90.12	86.04
SwAV [34]	<b>85.62</b>	85.73	92.12	88.56	91.09	88.62
MoCo-CXR [4]	81.75	83.56	90.98	85.49	90.03	86.36
MTPret	82.57	84.89	<b>93.38</b>	<b>92.10</b>	<b>93.15</b>	<b>89.22</b>

Note: Atel.: Atelectasis, Cardi.: Cardiomegaly, Consol.: Consolidation, and Pleural.: Pleural Effusion. Bold indicate the best results among all methods.

usually requires an extremely large number of images with fine-grained annotations to train the DNN models and deliver decent performance [8] in a supervised learning manner.

1) *Self-Supervised Learning for Medical Images*: To lower the size of annotated samples required, the self-supervised pretraining paradigm has been recently proposed to boost the performance of DNN models through learning visual features from images [16], [27] without the use of labels. Among a wide range of self-supervised pretraining methods, contrastive learning (CL) [27] algorithms use a similarity-based metric to

measure the distance between two embeddings derived from two different views of a single image, where the views of image are generated through data augmentation, e.g., rotation, clip, and shift, and embeddings are extracted from the DNN with learnable parameters. In particular, for computer vision tasks, the contrastive loss is computed using the feature representations of the images extracted from the encoder network, resulting in the clustering of similar samples together and the dispersal of different samples. Recent methods such as SwAV [34], SimCLR [27], MoCo [16], and PILR [35] have been proposed to outperform supervised learning methods on natural images. While contrastive learning methods have demonstrated promising results on natural image classification tasks, the attempts to leverage them in medical image analysis are often limited [36], [37]. Most recently [4] proposed MoCo-CXR that can produce models with better representations for the detection of pathologies in chest X-rays using MoCo [16]. Surveys on self-supervised pretraining could be found [38], [39].

2) *Multitask Learning for Medical Images*: In addition to self-supervised learning, multitask learning [8] is yet another paradigm to boost the performance of DNNs using datasets of various tasks. In recent days, a number of studies have applied multitask learning to medical image analytics. For example, [5] proposed a multitask representation learning framework for sequential medical images via self-supervision. [9] adopted multitask learning on CT images for quality assessment without the use of reference data. Furthermore, [6] proposed to combine multitask learning and contrastive learning for the diagnosis of COVID-19 with CT and X-ray data. More recently, [7] leveraged multitask vision transformer (ViT) and low-level chest X-ray feature corpus to determine the severity of COVID-19 infection.

3) *Catastrophic Forgetting for Multitask Learning*: A common phenomenon observed in DNN is the deterioration of performance on previous tasks when the networks are continually updated on new tasks or different data distributions. This phenomenon, known as “catastrophic forgetting”, becomes particularly pronounced in the context of medical images [40], which are generated through continuously changing policies, protocols, scanner hardware, or settings. Recent efforts have addressed this issue by employing dynamic memory for optimized data replay [40], [41], aiming to mitigate forgetting. In this work, we use the use of  $L^2$ -SP to address catastrophic forgetting in multitask continual learning.  $L^2$ -SP, functioning as a method of parameter regularization, introduces penalties for forgetting to retain the memory of features learned on source tasks. Through this approach, we seek to enhance the robustness of DNNs in the face of continuous updates for multitask learning, specifically in the domain of medical image processing.

## B. Comparisons With Relevant Works

The most relevant works to our study are [4], [6] from the problem and solution perspectives, respectively. Compared to [4], we both intend to pretrain DNNs using X-ray images

through self-supervisions. However, our work intend to pre-train the DNNs using multiple datasets with various tasks. Thus, MD-MoCo has been proposed by *MTPret* to incorporate multiple datasets in MoCo-based representation learning stage. Furthermore, *MTPret* leveraged multitask continual learning to fully utilize the task-specific information in all pretraining datasets so as to learn discriminative features better. Compared to [6], both of us intended to take self-supervised learning in multitask settings. However, *MTPret* used multiple datasets to pretrain DNNs for various transfer learning applications, while [6] only improved COVID-19 diagnosis using multimodal datasets (i.e., X-ray and CT). In this work, we only evaluate *MTPret* with two convolutional backbones ResNet-18 and ResNet-50. For future work, we intend to study novel ViTs [42] with even more self-supervised learning tasks [43] to further improve the pretraining and fine-tuning performance of *MTPret*.

Compared to the conference version [44] of this work, we have made nontrivial extension from following four aspects: 1) We have revised the whole manuscript with more informative materials in introduction section. We summarize the goal of this research as the answer to three research questions. Refer to Section Section I for details. 2) We have included new materials in methodology section, where we significantly extended and improved methodology section with elaborations, pseudocodes, and analysis on detailed design. Refer to Section III for details. 3) We have included new results in experiment section. Specifically, we included two new external tasks derived from two external datasets for detailed evaluation. We also provided new results in ablation studies, and interpretation studies with neural network dissection and visualization. These new results further confirmed the generalizability and superiority of the proposed method. Refer to Section IV for details. 4) We have included additional discussions on the limitation of this work and the related works. We compared this work more comprehensively with more baseline algorithms under various setting. Refer to Section V for details.

## V. LIMITATIONS

While *MTPret* has achieved significant performance gains, the work still has the following limitations. The first limitation is that this work only evaluate *MTPret* with two convolutional backbones ResNet-18 and ResNet-50, without extending *MTPret* to ViTs [42]. ViTs has been widely used in medical image-related analysis tasks over the past two years, including classification, segmentation, registration, etc. For future work, we intend to study novel ViTs with even more self-supervised learning tasks [43] to further improve the pretraining and fine-tuning performance of *MTPret*.

Another limitation of this work is that, compared to traditional self-supervised learning methods such as SimCLR [27] and MoCo [16], *MTPret* further introduces multitask learning to enhance representation learning, potentially introducing some supervised information. Although we mitigate the risk of the model learning from the test set data during experimental design by partitioning the dataset, there is still a risk of the model

learning the dataset distribution prematurely during the pre-training stage, such as the NIHCC [20] dataset. However, we still assert that *MTPret* is a robust algorithm, as evidenced by our evaluation of *MTPret* on Deep-Covid [2] and CheXpert [24] datasets, neither of which were used for any stage of *MTPret* pretraining. This experiment further demonstrates the domain generalization ability of *MTPret*, showcasing its robust performance on unseen datasets.

## VI. CONCLUSION

This work presents a “proof-of-concept” study on multitask/multidataset self-supervised representation learning for X-ray images. It introduces *MTPret*, a pretraining pipeline designed to enhance the performance of DNNs in various X-ray analytic tasks. The approach involves collecting and aggregating multiple X-ray image datasets from different body parts to address data inconsistency issues between tasks and datasets. Additionally, it incorporates MD-MoCo and multitask continual learning for pretraining the backbone DNNs in a self-supervised CL manner. The study demonstrates the performance of *MTPret* using eleven X-ray image datasets and evaluates its effectiveness on seven tasks, including pneumonia classification, skeletal anomaly classification, lung segmentation, tuberculosis bbox detection, chest disease diagnosis, and COVID-19 classification. The results show significant performance improvements, particularly in ResNet-18 (8.21%↑) and ResNet-50 (7.70%↑) on the NIHCC dataset. Moreover, *MTPret* displays robustness and generalizability in ubiquitous X-ray analytic tasks, as evidenced by its performance in external tasks such as chest disease diagnosis on CheXpert and COVID-19 classification on Deep-Covid. It is noted that *MTPret* is considered a proof-of-concept study, leveraging self-supervised representation learning and multitask continual learning to achieve performance improvement in numerous tasks.

## REFERENCES

- [1] H. Brody, “Medical imaging,” *Nature*, vol. 502, no. 7473, 2013.
- [2] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Soufi, “Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning,” *Med. Image Anal.*, vol. 65, 2020, Art. no. 101794.
- [3] E. Calli, E. Sogancioglu, B. Ginneken, K. Leeuwen, and K. Murphy, “Deep learning for chest X-ray analysis: A survey,” *Med. Image Anal.*, vol. 72, 2021, Art. no. 102125.
- [4] H. Sowrirajan, J. Yang, A. Ng, and P. Rajpurkar, “MoCo pretraining improves representation and transferability of chest X-ray models,” in *Proc. Med. Imag. Deep Learn.*, PMLR, 2021, pp. 728–744.
- [5] N. Dong, M. Kampffmeyer, and I. Voiculescu, “Self-supervised multi-task representation learning for sequential medical images,” in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Cham, Switzerland: Springer-Verlag, 2021, pp. 779–794.
- [6] J. Li et al., “Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19,” *Pattern Recognit.*, vol. 114, 2021, Art. no. 107848.
- [7] S. Park et al., “Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification,” *Med. Image Anal.*, vol. 75, 2022, Art. no. 102299.
- [8] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022.
- [9] D. Pal, B. Patel, and A. Wang, “SSIQA: Multi-task learning for non-reference CT image quality assessment with self-supervised noise level prediction,” in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 1962–1965.
- [10] Z. Zhou, V. Sodha, J. Pang, M. Gotway, and J. Liang, “Models genesis,” *Med. Image Anal.*, vol. 67, 2021, Art. no. 101840.
- [11] X. Li, Y. Grandvalet, and F. Davoine, “Explicit inductive bias for transfer learning with convolutional networks,” in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2018, pp. 2825–2834.
- [12] L. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, *arXiv:1706.05587*.
- [13] R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [14] R. Faster, “Towards real-time object detection with region proposal networks,” *Adv. Neural Inf. Process. Syst.*, vol. 9199, no. 10.5555, pp. 2969239–2969250, 2015.
- [15] Z. Wu, Y. Xiong, S. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.
- [16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
- [17] S. Jaeger, S. Candemir, S. Antani, Y. Wang, P. Lu, and G. Thoma, “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases,” *Quantitative Imag. Med. Surgery*, vol. 4, no. 6, pp. 475–477, 2014.
- [18] D. Demner-Fushman et al., “Preparing a collection of radiology examinations for distribution and retrieval,” *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, 2016.
- [19] S. Halabi et al., “The RSNA pediatric bone age machine learning challenge,” *Radiology*, vol. 290, no. 2, pp. 498–503, 2019.
- [20] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, “ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2097–2106.
- [21] D. Kermany et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
- [22] P. Rajpurkar et al., “MURA: Large dataset for abnormality detection in musculoskeletal radiographs,” 2017, *arXiv:1712.06957*.
- [23] Y. Liu, Y. Wu, Y. Ban, H. Wang, and M. Cheng, “Rethinking computer-aided tuberculosis diagnosis,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2646–2655.
- [24] J. Irvin et al., “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 590–597.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
- [26] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, 2009, pp. 248–255.
- [27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 1597–1607.
- [28] A. Kirillov et al., “Segment anything,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.
- [29] Y. Xie and D. Richmond, “Pre-training on grayscale ImageNet improves medical image classification,” in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 476–484.
- [30] O. Stephen, M. Sain, U. Maduh, and D. Jeong, “An efficient deep learning approach to pneumonia classification in healthcare,” *J. Healthcare Eng.*, 2019.
- [31] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, “Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8300–8311.
- [32] B. Bozorgtabar, D. Mahapatra, G. Vray, and J. Thiran, “SALAD: Self-supervised aggregation learning for anomaly detection on X-rays,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Cham, Switzerland: Springer-Verlag, 2020, pp. 468–478.
- [33] Z. Yuan, Y. Yan, M. Sonka, and T. Yang, “Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3040–3049.
- [34] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting

- cluster assignments,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9912–9924, Jan. 2020.
- [35] I. Misra and L. Maaten, “Self-supervised learning of pretext-invariant representations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6707–6717.
- [36] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 32.
- [37] V. Cheplygina, M. Bruijne, and J. Pluim, “Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.
- [38] A. Jaiswal, A. Babu, M. Zadeh, D. Banerjee, and F. Makedon, “A survey on contrastive self-supervised learning,” *Technologies*, vol. 9, no. 1, p. 2, 2020.
- [39] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, 2020.
- [40] J. Hofmanninger, M. Perkonig, J. A. Brink, O. Pinykh, C. Herold, and G. Langs, “Dynamic memory to alleviate catastrophic forgetting in continuous learning settings,” in *Proc. Med. Image Comput. Comput. Assisted Intervention (MICCAI)*, Lima, Peru, Cham, Switzerland: Springer-Verlag, 2020, pp. 4–8.
- [41] M. Perkonig et al., “Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging,” *Nature Commun.*, vol. 12, no. 1, pp. 5678–5678, 2021.
- [42] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–12.
- [43] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
- [44] W. Liao et al., “MUSCLE: Multi-task self-supervised continual learning to pre-train deep models for X-ray images of multiple body parts,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Cham, Switzerland: Springer-Verlag, 2022, pp. 151–161.