Foundation X: Integrating Classification, Localization, and Segmentation through Lock-Release Pretraining Strategy for Chest X-ray Analysis

Nahid Ul Islam[1] DongAo Ma[1] Jiaxuan Pang[1] Shivasakthi Senthil Velan[1]
Michael Gotway[2] Jianming Liang[1]
[1]Arizona State University, USA [2]Mayo Clinic, USA

Abstract

Developing robust and versatile deep-learning models is essential for enhancing diagnostic accuracy and guiding clinical interventions in medical imaging, but it requires a large amount of annotated data. The advancement of deep learning has facilitated the creation of numerous medical datasets with diverse expert-level annotations. Aggregating these datasets can maximize data utilization and address the inadequacy of labeled data. However, the heterogeneity of expert-level annotations across tasks such as classification, localization, and segmentation presents a significant challenge for learning from these datasets. To this end, we introduce ***Foundation X****, an end-to-end framework that utilizes diverse expert-level annotations from numerous public datasets to train a foundation model capable of multiple tasks including classification, localization, and segmentation. To address the challenges of annotation and task heterogeneity, we propose a Lock-Release pretraining strategy to enhance the cyclic learning from multiple datasets, combined with the student-teacher learning paradigm, ensuring the model retains general knowledge for all tasks while preventing overfitting to any single task. To demonstrate the effectiveness of Foundation X, we trained a model using 11 chest X-ray datasets, covering annotations for classification, localization, and segmentation tasks. Our experimental results show that Foundation X achieves notable performance gains through extensive annotation utilization, excels in cross-dataset and cross-task learning, and further enhances performance in organ localization and segmentation tasks. All code and pretrained models are publicly accessible at GitHub.com/JLiangLab/Foundation_X.*

1. Introduction

In computer vision, tasks like classification, localization, and segmentation are often handled independently. This approach can lead to inefficiencies and limit performance in complex, real-world applications. Isolated models miss the opportunity to leverage the rich, diverse information available when these tasks are integrated [6, 39]. In medical imaging, datasets often contain different diseases annotated with image-level labels, disease-specific bounding boxes, and segmentation masks. For example, CheXpert [9] dataset has only image-level labels for classification, while TBX11K [19] has both image-level labels and bounding boxes, and CANDID-PTX [3] provides annotations for all three tasks. By integrating these types of annotations into a single model, we can achieve a deeper understanding of each dataset. We hypothesize that combining classification (to identify diseases), localization (to generate bounding boxes), and segmentation (to delineate boundaries) tasks within the same framework will improve image analysis. This, in turn, will lead to better diagnostic accuracy and more informed medical decisions. Therefore, developing an end-to-end framework that handles all tasks simultaneously would boost performance and enhance robustness by taking advantage of the semantic relationships between tasks [6, 8, 10]. However, this integration poses a significant challenge, as model tend to overfit to a single task during training, hindering generalization across multiple tasks. To overcome this issue, our research proposes a framework that tackles these tasks concurrently and serves as a foundation model. By training it on large-scale, diverse datasets and tasks, we aim to build a system capable of handling a broad range of real-world applications, improving both task-specific performance and generalizability. Such a model leverages the synergy between classification, localization, and segmentation, creating a versatile, robust system while maximizing annotation use, reducing costs, and enhancing efficiency in medical image analysis. This leads us to our central research question: *How can we integrate classification, localization, and segmentation tasks within a single model to improve its performance and generalization ability, specifically in Chest X-ray image analysis?* In our research, we have chosen Chest X-rays (CXRs) because they are one of the most frequently used imaging modalities, and the availability of CXR data is substantial (see Table 1).

To this extent, we have developed **Foundation X**, an

Dataset	Classification (Head Id)	Localization (Decoder Id)	Segmentation (Head Id)
1. CheXpert [9]	✓(C_1)	-	-
2. NIH ChestX-ray14 [35]	✓(C_2)	-	-
3. VinDr-CXR [25]	✓(C_3)	-	-
4. NIH Shenzhen CXR [11]	✓(C_4)	-	-
5. MIMIC-II [12]	✓(C_5)	-	-
6. TBX11k [19]	✓(C_6)	✓(L_1)	-
7. NODE21 [31]	✓(C_7)	✓(L_2)	-
8. CANDID-PTX [3]	✓(C_8)	✓(L_3)	✓(S_1)
9. RSNA Pneumonia [26]	✓(C_9)	✓(L_4)	-
10. ChestX-Det [16]	✓(C_{10})	✓(L_5)	✓(S_2)
11. SIIM-ACR [1]	✓(C_{11})	✓(L_6)	✓(S_3)
12. CheXmask VinDr-CXR [5]	-	✓	✓
13. VinDr-RibCXR [24]	-	-	✓
14. NIH Montgomery [11]	-	-	✓
15. JSRT [33]	-	-	✓
16. VinDr-CXR [25]	-	✓	-

Table 1. We pretrain our Foundation X model using 11 publicly available chest X-ray datasets, as shown in the first 11 datasets in the table. Although not every dataset contains all three types of annotations—classification, localization, and segmentation—we leverage all available annotations to maximize the model's learning potential. Among these datasets, all include classification ground truths, 6 provide localization bounding box annotations, and 3 offer segmentation masks for diseases. Furthermore, we utilize organ localization and segmentation datasets from VinDr-CXR, VinDr-RibCXR, NIH Montgomery, and JSRT for target task finetuning. Here, the organ segmentation masks for VinDr-CXR were sourced from the CheXmask database.We also finetuned VinDr-CXR with local labels for disease localization task.

end-to-end model that integrates classification, localization, and segmentation tasks for medical imaging (illustrated in Figure 1). We hypothesize that sharing learned representations across tasks equips the model to better capture intricate patterns in medical images, leading to more reliable diagnostics. However, integrating classification, localization, and segmentation into a single model risks overfitting, especially during large-scale training. Foundation X tackles this by using a shared backbone and innovative pretraining strategies, ensuring balanced performance across diverse tasks. To demonstrate the capability of Foundation X, we train the model on 11 datasets (Table 1) and further finetune it on additional target tasks, showcasing the model's potential.

In summary, our analysis shows that Foundation X achieves significant performance gains through extensive annotation usage (Table 4), excels in cross-dataset and cross-task learning (Figure 2), and further enhances organ localization and segmentation (Table 3, 5, 6). Through this work, we have made the following contributions:

1. We develop and implement Foundation X, an integrated model for classification, localization, and segmentation tasks in Chest X-ray images;
2. We propose a Lock-Release pretraining strategy to enhance the cyclic learning from multiple datasets, preventing task overfitting and ensuring balanced learning across tasks and datasets;
3. We provide comprehensive experimental results to demonstrate Foundation X's improved performance and generalization ability.

Foundation X provides a versatile, end-to-end framework for handling classification, localization, and segmentation tasks in medical imaging. By leveraging shared knowledge across tasks, Foundation X enhances generalization, reduces overfitting, and maximizes annotation use, leading to more efficient data utilization. This approach allows the model to adapt to new tasks and datasets, making it valuable for continuous learning and real-world medical use.

2. Related Work

Multitask learning mimics the human brain by performing tasks simultaneously with minimal supervision and simplifying cross-learning from different tasks. Significant attention has been given to multitask learning [2] within the research community. In recent years, researchers have focused on utilizing a single backbone with one or more decoders for multitask learning [7, 14, 23, 28, 40]. However, the information shared between these unified or separate decoders was often ineffective, limiting performance and being restricted to either single-task multi-source or single-source multitask objectives. Researchers have explored various approaches to address and improve these learning methods. One approach is OmniFM-DR [37], which uses the text encoder in parallel with a backbone and along with multiple unified decoders to perform multitask learning. However, the expert annotations are modified to create the input text features for the encoder and the unified multitask decoder architecture. Other learning methods [15, 21, 34] involve using an image backbone and context encoder, then

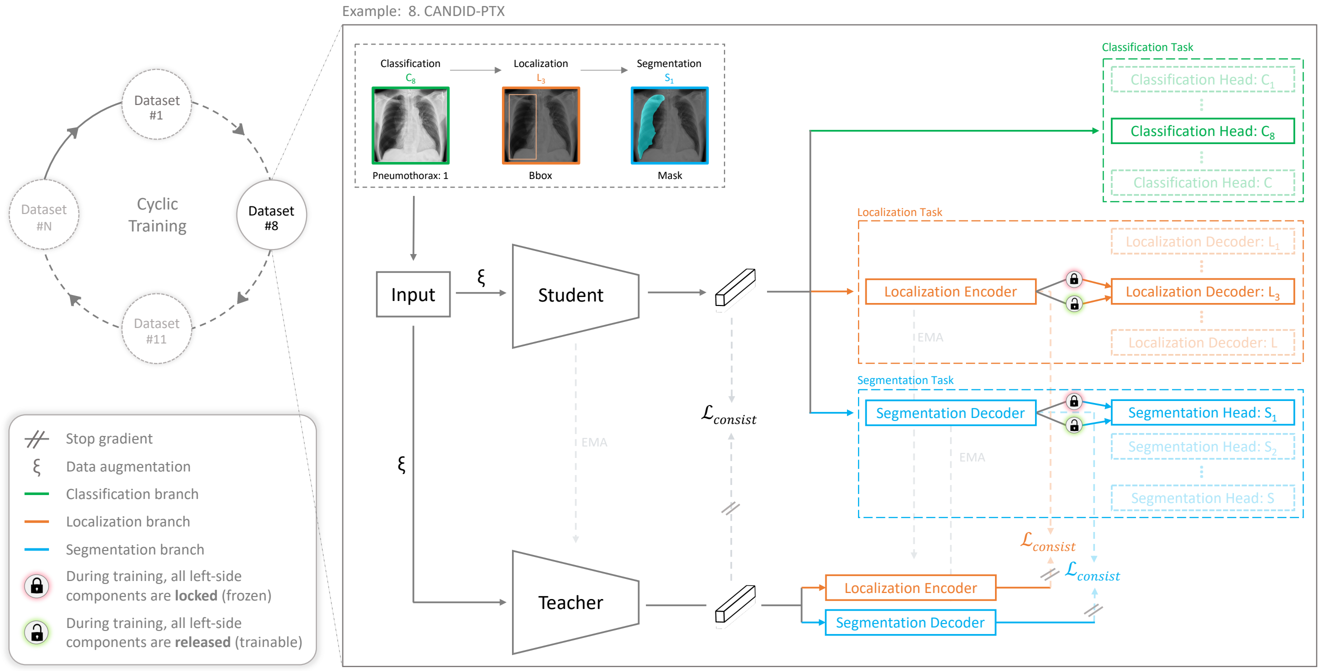

Figure 1. The proposed Foundation X model (detailed in Sec. 3) can utilize multiple datasets (Dataset #1 to #11) for pretraining and can also incorporate additional datasets (Dataset #N) dynamically into the pretraining process. The model is trained cyclically, processing each dataset sequentially. Each dataset may include one, two, or all three tasks: classification, localization, and segmentation. The figure illustrates the process with a dataset (e.g., 8. CANDID-PTX [3]) containing all three types of ground truths. The process begins with the student model (Swin-B) extracting relevant features from the input dataset, which are then directed sequentially to the appropriate branch. First, for classification, features are processed through the classification head (C_8). Second, for localization, features pass through the localization encoder and corresponding localization decoder (L_3). Third, for segmentation, features are handled by the segmentation decoder and segmentation head (S_1). The model undergoes two-phase training for each task: lock mode with most layers frozen, followed by release mode with all layers trainable. Additionally, the model uses a student-teacher learning paradigm. The teacher model, an identical copy of the student model, is updated after each epoch using an exponential moving average (EMA). We calculate the consistency loss (L_{const}) in three areas: extracted features from the backbone, features from the localization encoder, and features from the segmentation decoder. If a dataset contains only one or two types of ground truths, the model will skip the branch without the corresponding ground truth. The Foundation X model uses the Cyclic and Lock-Release pretraining strategies to enhance performance across tasks while preventing forgetting and avoiding task-specific overfitting.

feeding these features into a visual-lingual encoder and decoder. However, these one-stage model uses multiple heads for multitask learning, limited to either a text encoder or multiple backbones for the encoder-decoder architecture.

On the other hand, the framework X-Learner [8] consists of two stages: Expansion, where multiple sub-backbones are learned and interconnected to reduce task interference, and Squeeze, where the expanded backbone is condensed into a normal-sized one for effective downstream transfer. However, the multiple sub-backbones may not efficiently capture information between tasks, potentially leading to suboptimal performance on some tasks. Additionally, the lack of an open code-base limits opportunities for comparative studies and comprehensive evaluations.

For techniques to aggregate dataset annotations, Universal Object Detection with Large Vision Model [17] proposes to modify the existing hierarchy of the dataset labels to create a unified label space and incorporate a hierarchical loss suppression technique to efficiently calculate losses for the taxonomies in the labels. Our Foundation X framework differs from [17] by aggregating multiple datasets without any modifications to the label taxonomies or the classes themselves through the use of multiple task-heads. Each task-head in Foundation X is used for a single dataset and vision task. The work, ViM (Vision Middleware for Unified Downstream Transferring) [4], takes on accommodating multiple different tasks in a single model by introducing a new paradigm. The design proposed by [4] involves using a fully frozen backbone pretrained on a large dataset for an upstream task. It then combines various ViM modules (adapters) trained on datasets such as COCO [18], Objects365 [29], and others, to finally finetune on the target downstream dataset. Our method contrasts [4] by having a simpler but effective architecture that inherently learns generalizable representations from multiple datasets and tasks.

Foundation Ark [22] first introduced the idea of accruing and reusing knowledge embedded in the expert annotations from numerous datasets cyclically, but it focused solely on the classification task. In contrast, Foundation X focuses on encompassing diverse tasks, including classification, localization, and segmentation, via a Lock-Release strategy.

3. Method

3.1. Foundation X: Integrating Classification, Localization and Segmentation

Our Foundation X aims to integrate classification, localization, and segmentation tasks to train a powerful, robust, and versatile model. This integrated approach allows the model to develop a comprehensive understanding of images, enhancing its performance across all tasks. By leveraging the synergistic effect of these tasks, we can improve overall diagnostic accuracy and maximize the utilization of available annotated data, thereby saving the annotation costs. To achieve this, we designed Foundation X with a shared backbone to learn general and complementary knowledge across tasks, and separate branches to address the specific needs of classification, localization, and segmentation. This separation enables independent optimization and finetuning of each branch, giving Foundation X the flexibility and scalability to add new tasks and datasets while maintaining high performance across diverse applications.

As shown in Figure 1, the backbone of Foundation X serves as a knowledge encoder, extracting common features for various tasks performed by three branches: the classification branch with multiple heads, the localization branch with an encoder and multiple decoders, and the segmentation branch with a decoder and multiple heads. Each head or localization decoder corresponds to a specific dataset, as listed in Table 1. For the classification branch, we implement the classification head with a linear classifier. For the localization branch, we utilize the localization encoder and decoder from DINO (DETR with Improved deNoising anchOr boxes) [38], with modifications to accommodate multiple datasets by using multiple localization decoders. This design ensures that each localization dataset has a dedicated decoder, allowing the model to effectively handle multiple tasks. For the segmentation branch, we utilize the UperNet decoder [36] and add multiple segmentation heads atop it to handle different segmentation datasets. The design of multiple heads and localization decoders enhances the flexibility of Foundation X, allowing it to seamlessly add new tasks and making it scalable for task expansion.

Configuration	**Utilization of Branch**
Foundation X-C	Only **C**lassification
Foundation X-L	Only **L**ocalization
Foundation X-S	Only **S**egmentation
Foundation X-CL	**C**lassification and **L**ocalization
Foundation X-LS	**L**ocalization and **S**egmentation
Foundation X-CLS	**C**lassification, **L**ocalization & **S**egmentation

Table 2. Differences between Foundation X model configurations.

3.2. Cyclic Pretraining Strategy

Training a single model using diverse datasets with inconsistent or heterogeneous annotations is challenging. The model must learn from various types of information and integrate them into a cohesive understanding. For example, the model needs to extract high-level representations of the entire image for classification, identify specific regions of interest for localization, and delineate the precise boundaries of these regions for segmentation. Even for the same type of task, datasets created at different institutions tend to be annotated differently, further complicating the learning process [22].

To address these challenges, we employ the Cyclic pretraining strategy from Foundation Ark [22]. This strategy enables the model to learn from multiple tasks by revisiting each one in every training round, thereby reinforcing the learning process and preventing the model from forgetting previously acquired knowledge. Benefiting from the Cyclic pretraining, Foundation X avoids the issue of catastrophic forgetting [13, 22], where performance on earlier tasks significantly degrades when new tasks are introduced. This approach enables the model to achieve more robust and generalized performance, enhancing its effectiveness across multiple datasets and tasks.

3.3. Lock-Release Pretraining Strategy

Another challenge when training a model on diverse datasets and tasks is ensuring it maintains good generalizability across all, without overfitting towards any single dataset or task. The model must be capable of learning from various tasks, understanding heterogeneous annotations, and integrating sophisticated domain knowledge into a cohesive framework that performs well across different tasks. However, due to the varying number of training samples and the differing difficulty levels of each task, the model's learning speed can vary.

To balance the learning process for each task, we have developed a Lock-Release pretraining strategy for localization and segmentation tasks. Initially, the model is trained in the *lock* mode, where most early layers are frozen and only a few upper layers are trainable. Specifically, for localization and segmentation tasks, only the localization decoder and segmentation head are trainable, respectively. This mode focuses on finetuning higher-level features specific to the task while preserving general features learned from other datasets. Subsequently, training switches to the *release* mode, where all layers are made trainable, allowing for full adaptation and refinement. During the *lock* mode training, only half of the dataset is used to prevent early overfitting, while the full dataset is utilized during the *release* mode training to ensure comprehensive learning. This Lock-Release strategy helps prevent the model from overfitting too heavily towards one task when exposed to multiple tasks and datasets, ensuring a more balanced learning process.

We tried the Lock-Release strategy for classification but found it ineffective, so it was not applied. Unlike localization and segmentation, which have more parameters to tune, classification relies on lightweight heads, making Lock-Release less impactful.

3.4. Student-Teacher Learning Paradigm

To further mitigate the issue of forgetting and prevent overfitting towards any single task, thereby balancing and stabilizing the learning process across diverse tasks and datasets, we introduce the student-teacher learning paradigm in Foundation X. The teacher model uses the same architecture as the student model, and both are initialized with the same weights. The student model is updated through standard training processes, while the teacher model is updated using an exponential moving average (EMA) [32] based on the student's learning at the end of each epoch. Furthermore, as shown in Figure 1, We incorporate a consistency loss for the features from the backbone, the localization encoder, and the segmentation decoder between the student and teacher models. This consistency loss ensures that the features learned by the student model remain aligned with those of the teacher model, promoting stability and improved performance during training. The student-teacher learning paradigm enhances generalization across classification, localization, and segmentation tasks, resulting in improved performance and robustness on various tasks. After pretraining, the teacher model from Foundation X will be used for the downstream tasks.

4. Experiments and Results

Pretraining Foundation X. Foundation X is pretrained on 11 datasets (see Table 1) encompassing three common medical tasks: *disease* classification, localization, and segmentation. Among these datasets, all provide disease classification, two include localization and three offer segmentation mask annotations. Since CANDID-PTX [3] and SIIM-ACR Pneumothorax [1] are annotated only with disease masks, we derive localization bounding boxes from them, resulting in a total of four localization tasks. We use the official dataset splits when available and perform random splits (70% train, 10% val, 20% test) for those without.

Finetuning Foundation X. We collect an additional 5 publicly available chest X-ray datasets to finetune Foundation X on *organ* segmentation and localization tasks. CheXmask [5] offers a comprehensive collection of uniformly annotated chest radiographs compiled from five public sources: ChestX-ray8, Chexpert, MIMIC-CXR-JPG, Padchest, and VinDr-CXR. It includes segmentation masks for the heart, left lung, and right lung. From these provided segmentation masks, we derive localization bounding boxes for the same organs. We specifically utilize the VinDr-CXR portion of the dataset to localize and segment the organs. Furthermore, we use the NIH Montgomery [11], VinDr-RibCXR [24], and JSRT [30] datasets for organ segmentation in our Foundation X model. NIH Montgomery provides lung masks, VinDr-RibCXR provides rib masks, and JSRT provides heart, lung, and clavicle masks.

Dataset	Baseline	Baseline	Foundation X-LS	
VinDr-CXR	Loc.	Seg.		
	[mAP40%]	[Dice%]	[mAP40%]	[Dice%]
Heart	80.17	95.82	**88.41** ↑ 8.24	**96.15** ↑ 0.33
Left Lung	90.72	97.46	**95.58** ↑ 4.86	**97.57** ↑ 0.11
Right Lung	92.42	98.03	**96.78** ↑ 4.36	**98.13** ↑ 0.10

Table 3. Baseline localization and segmentation models are trained separately using DINO [38] and UperNet [36] to localize and segment the heart, left lung, and right lung in the VinDr-CXR dataset, each employing a single head for three classes. In contrast, our Foundation X-LS model handles both tasks together in pretraining, showing enhanced performance. The green arrow highlights Foundation X's performance improvements over the baselines.

4.1. Foundation X achieves performance gains through extensive annotation utilization

Experimental Setup: We pretrain Foundation X on a large-scale dataset collection, utilizing the first 11 datasets from Table 1, which encompass 20 tasks (C_1-C_{11} for classification, L_1-L_6 for localization, and S_1-S_3 for segmentation). We initialize the Swin-B backbone with Ark-6 [22] pretrained weights. For each dataset, we pretrain Foundation X on all tasks sequentially. For example, epoch x is dedicated to classification, $x + 1$ to localization, and $x + 2$ to segmentation. If a dataset lacks annotations for a specific task, that task is simply skipped. We define one *cycle* as the model completing training on all 20 tasks listed in Table 1. During pretraining, we employ Student-Teacher learning paradigm, Cyclic and Lock-Release pretraining strategies. This approach ensures comprehensive exposure to all tasks and datasets, promoting better generalization and robust performance in medical imaging.

Results and Analysis: Using Cyclic and Lock-Release pretraining strategies, Foundation X achieves significantly better performance on most tasks with large-scale pretraining.

Dataset	Baseline Cls.†	Baseline Loc.†	Baseline Seg.†	Foundation X-CLS		
	[AUC%]	[mAP40%]	[Dice%]	[AUC%]	[mAP40%]	[Dice%]
CheXpert	90.03±0.48	-	-	**90.64** ↑ 0.61	-	-
NIH ChestX-ray14	83.05±0.09	-	-	82.95 (**83.35***↑ 0.30)	-	-
VinDr-CXR	95.07±0.15	-	-	**95.85** ↑ 0.78	-	-
NIH Shenzhen CXR	98.99±0.16	-	-	**99.64** ↑ 0.65	-	-
MIMIC-II	79.12±0.16	-	-	78.94 ↓ 0.18	-	-
TBX11K	99.89±0.06	78.08±0.81	-	**99.95** ↑ 0.06	78.38 (**81.80***↑ 3.72)	-
NODE21	99.35±0.45	37.78±2.67	-	**99.68** ↑ 0.33	**46.57** ↑ 8.79	-
CANDID-PTX	72.61±0.57	50.51±1.36	86.36	**73.86** ↑ 1.25	**54.14** ↑ 3.63	**89.81** ↑ 3.45
RSNA Pneumonia	88.87±0.21	20.83±0.54	-	**89.88** ↑ 1.01	**27.44** ↑ 6.61	-
ChestX-Det	88.17±0.33	38.12±0.50	**79.33**	85.07 (**89.89***↑ 1.72)	37.77 (**43.98***↑ 5.86)	64.49 (79.17*↓ 0.16)
SIIM-ACR	95.01±0.16	28.56±0.94	81.92	**96.44** ↑ 1.43	**34.59** ↑ 6.03	**83.65** ↑ 1.73

† All baselines use Swin-B [20] as the backbone with Ark-6 [22] pretrained weights. The classification baseline uses only Swin-B, the localization baseline uses Swin-B + DINO [38], and the segmentation baseline uses Swin-B + UperNet [36].

* Values inside parentheses indicate the finetuning results of Foundation X, while the preceding values represent the pretraining results.

Table 4. Performance Comparison of Foundation X and Baseline Models. The table presents the results from pretraining Foundation X, compared to baseline models, across 11 datasets encompassing 20 tasks (see Table 1). Foundation X is trained sequentially on these tasks using the Cyclic and Lock-Release pretraining strategies, which helps it generalize efficiently and retain knowledge of previous tasks. The results indicate that Foundation X outperforms most of the baselines, which are trained individually on specific datasets and tasks. This highlights the effectiveness of the integrated multitask learning approach in improving model performance and generalization ability. The arrow shows Foundation X's performance gain/loss compared with the baseline performance.

As shown in Table 4, it outperforms baseline models that are independently trained on specific datasets and tasks. Notably, due to our superior pretraining strategy, Foundation X does not exhibit overfitting toward a single task, resulting in consistent improvement across various datasets and tasks. In contrast, baseline models are fully finetuned on individual datasets, giving them an advantage in single-task performance. Consequently, during Foundation X's large-scale pretraining, it is expected that some tasks may face difficulties due to the need to balance learning across diverse datasets and tasks. Despite this challenge, Foundation X achieves comparable or better results than the baselines in most cases. To address underperforming tasks during pretraining, we further finetune Foundation X from the latest checkpoint, achieving better results than the baselines. As shown in Table 4, the NIH ChestX-ray14 disease classification task improves to 83.35%, the localization performance on TBX11K increases to 81.80%, the ChestX-Det classification raises to 89.89%, and ChestX-Det localization raises to 43.98%. However, finetuning for the ChestX-Det segmentation task (79.17%) showed limited improvement, likely due to the low class-to-data ratio.

4.2. Foundation X Enhances Performance for Organ Loc. and Seg.

Experimental Setup: We pretrain Foundation X exclusively on organ localization and segmentation tasks using the VinDr-CXR dataset. We first evenly divide the official training split, resulting in 7,500 non-overlapping images for both localization and segmentation pretraining tasks. Additionally, we further divide the dataset into three non-overlapping subsets for three specific organs (heart, left lung, and right lung). This careful partitioning ensures a rigorous assessment of Foundation X's performance in localizing and segmenting these three organs independently. To establish baseline performance, we train DINO [38] for localization and UperNet [36] for segmentation on the aforementioned tasks separately. To demonstrate Foundation X's superior pretraining strategy on localization and segmentation tasks, we pretrain Foundation X using both localization and segmentation branches along with the Cyclic and Lock-Release pretraining strategies. We refer to the resulting model as *Foundation X-LS* (Table 3) since it is trained exclusively on **L**ocalization and **S**egmentation tasks. To demonstrate that Foundation X-LS provides superior fine-grained features, we further finetune the model on three organ segmentation tasks: JSRT, NIH Montgomery, and

Dataset	Ark† [22]	POPAR‡ [27]	Foundation X-LS
	[Dice%]	[Dice%]	[Dice%]
JSRT-Heart	94.62±0.16	94.64±0.21	**95.42** ±0.02↑ 0.78
JSRT-Lung	97.48±0.06	97.71±0.07	**98.04** ±0.04↑ 0.33
JSRT-Clavicle	90.05±0.15	90.18±0.18	**91.17** ±0.34↑ 0.99
NIH Montgomery	97.68±0.03	97.78±0.05	**98.29** ±0.02↑ 0.51
VinDr-RibCXR	63.96±0.30	61.17±0.40	**71.12** ±0.56↑ 7.16

†We adopt this performance reported by the original authors [22].

‡ POPAR [27] is finetuned for the baselines.

Table 5. We initialize Foundation X-S with weights from the Foundation X-LS model, trained on VinDr-CXR for organ localization and segmentation. After finetuning on other target tasks, Foundation X-S shows notable performance gains over Ark and POPAR. The green arrow indicates its improvements over the second-best method (underlined).

VinDr-RibCXR. Lastly, we evaluate the model's effectiveness in few-shot learning setups using the JSRT-clavicle dataset. Here, we compare our results against two pretraining baselines: Ark [22] and POPAR [27]. Ark utilizes the Swin-B backbone pretrained on 6 chest X-ray datasets in a supervised setup using the cyclic pretraining strategy. POPAR, a self-supervised method, also employs Swin-B pretrained on NIH ChestX-ray14 images, leveraging consistent and recurrent anatomical patterns in medical images to learn patch-level spatial relationships and fine-grained appearance features. For comparison on the segmentation task, we build the baseline using Swin-B + UperNet, where the Swin-B backbone is initialized with Ark-6 or POPAR pretrained weights.

Results and Analysis: As shown in Table 3, compared to the two baseline methods, DINO and UperNet, which are independently trained for specific tasks, the Foundation X-LS model—pretrained by sequentially incorporating both localization and segmentation—achieves the best performance. Specifically, during pretraining, we observe substantial gains in localization, with performance increases of 8.24% for the heart, 4.86% for the left lung, and 4.36% for the right lung. Additionally, segmentation tasks show improvements of 0.33%, 0.11%, and 0.10% for the heart, left lung, and right lung, respectively. Table 5 demonstrates the Foundation X-LS model's finetuning performance on the individual task. When finetuning on the JSRT dataset, we observe performance gains of 0.78% for heart, 0.33% for lung, and 0.99% for clavicle segmentation. Additionally, we note gains of 0.51% for the NIH Montgomery and 7.16% for the VinDr-RibCXR segmentation task. In a few-shot setup for clavicle segmentation on the JSRT dataset (Table 6), Foundation X consistently outperforms Ark and POPAR across all training sample sizes.

Training Samples	Ark [22] [Dice%]	POPAR [27] [Dice%]	Foundation X-LS [Dice%]
24	86.32	86.14	**88.81** ↑ 2.49
20	84.87	86.27	**88.23** ↑ 1.96
15	84.73	83.23	**86.65** ↑ 1.92
12	80.82	81.46	**85.89** ↑ 4.43
6	82.71	79.03	**83.03** ↑ 0.32
3	74.98	70.68	**78.18** ↑ 3.20

Table 6. We finetune Foundation X-LS which is trained on VinDr-CXR for organ localization and segmentation. We then finetune it for JSRT clavicle segmentation in a few-shot learning setup. Our results consistently show that Foundation X-S outperforms Ark and POPAR, with the green arrow indicating performance boosts over the second-best method (underlined).

4.3. Foundation X maximizes performance with cross-dataset and cross-task learning

Experimental Setup: We assess the generalizability and effectiveness of the Foundation X model across datasets and tasks. Specifically, We evaluate 1) how well the Foundation X model pretrained on one dataset performs on another, and 2) how the Foundation X model performs on two other tasks when pretrained by the third task within the same dataset. During the pretraining of Foundation X, we evaluate the model on the testing sets of all tasks across all datasets after each epoch, generating a series of Epoch V.S. Performance graphs. Our goal is to observe how testing performance changes over time when the model is trained on the same dataset (focused training) compared to when it is trained on other datasets (unfocused training). Figure 2 presents the results from focused training, unfocused training, and the best results from focused training for three selected datasets (RSNA Pneumonia, SIIM-ACR, and CANDID-PTX). Comprehensive plots for all datasets with multiple tasks are included in the supplementary document.

Results and Analysis: Figure 2 demonstrates positive performance trends across the datasets for both focused and unfocused training on each dataset. This indicates that the Foundation X model can effectively generalize and improve performance even without explicit training on specific tasks of the evaluated dataset. During unfocused training (light-colored line), the performance dips are common initially, but improvement is typically observed over time. In all cases shown in Figure 2, the results from unfocused training do not drift away from the task, indicating that the model can generalize efficiently and retain knowledge of previous tasks due to the Student-Teacher learning paradigm, Cyclic and Lock-Release pretraining strategies. The model performs consistently across all datasets and tasks. In some cases, unfocused training even outperforms focused training, highlighting the benefits of cross-task and cross-dataset learning in Foundation X.

5. Discussion

Foundation X aims to develop a model that outperforms task-specific models by collaboratively learning from multiple datasets and tasks. Moreover, finetuning is required if a task-specific model needs to adapt to other tasks, such as adapting a classification pretrained model to a localization task. Our Foundation X model addresses evolving diagnostic task requirements through easy and quick finetuning. As observed in Figure 3, finetuning Foundation X that is pretrained only on the classification task of the VinDr-CXR [25] dataset, amongst other pretraining datasets and tasks (Refer to Table 1), achieves superior localization performance compared to the baseline model for the same dataset. We attribute this performance gain to the knowledge gathered from the Cyclic and Lock-Release pretraining strategies across all tasks and datasets.

To further assess our Foundation X, we finetune the latest checkpoint from Foundation X-CLS using full finetun-

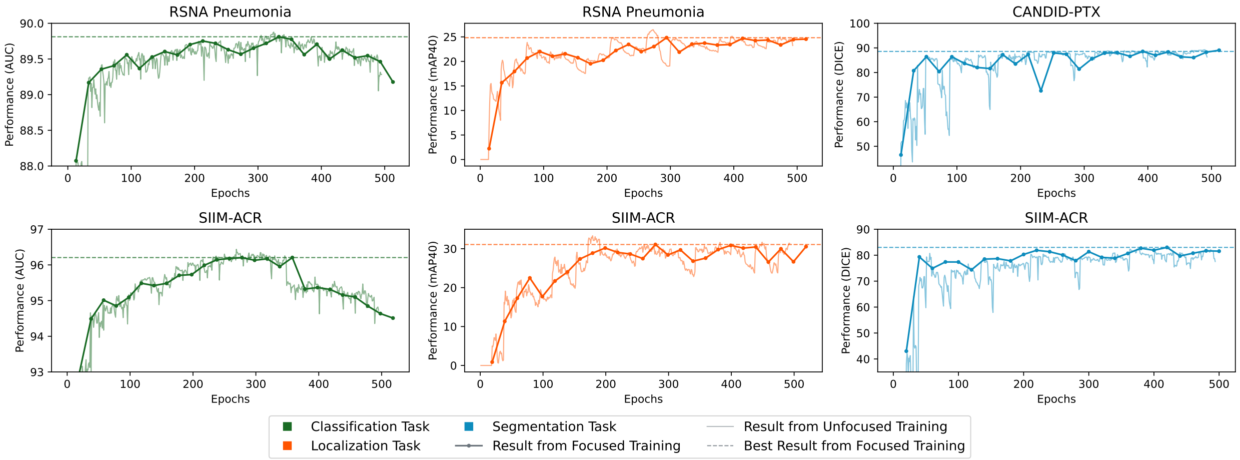

Figure 2. Cross-Dataset & Cross-Task Learning Analysis. The figure demonstrates the performance trends of Foundation X across multiple datasets for both focused and unfocused training scenarios. Focused training refers to scenarios where the model is explicitly trained on the specific dataset being evaluated. In contrast, unfocused training refers to scenarios where the model is trained on other datasets and not directly on the one being evaluated. The green, orange, and blue lines represent classification, localization, and segmentation tasks, respectively. Dark-colored lines indicate focused training results, while light-colored lines show unfocused training results. Dashed lines represent the best testing outcomes from focused training. In some cases, unfocused training surpasses focused training, highlighting the benefits of cross-task and cross-dataset learning in enhancing Foundation X's capabilities. The model efficiently generalizes, retains knowledge of previous tasks, and avoids overfitting during pretraining.

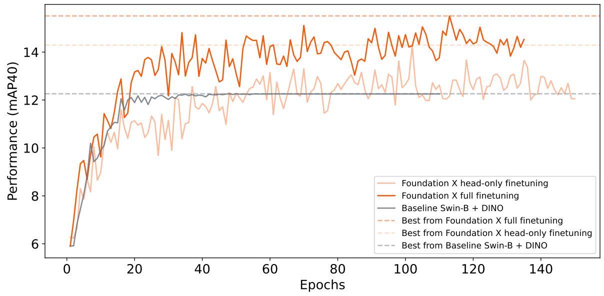

Figure 3. Full finetuning of Foundation X outperforms both head-only finetuning Foundation X and the baseline Swin-B + DINO model with Ark-6 [22] initialized backbone weights. All three settings followed the same hyperparameters as mentioned in the supplementary material.

ing, head-only finetuning and compare it with the localization baseline for the VinDr-CXR [25] localization dataset. In full finetuning, the model is trained on the VinDr-CXR localization dataset with a randomly initialized localization decoder. In head-only finetuning, the backbone and localization encoder are frozen, leaving only the new task head (localization decoder) trainable, with $\approx$ 9.6 million parameters. While it is challenging to train a model with most of its layers frozen, head-only finetuning achieved 12.27% mAP40 at the 49th epoch, where the baseline achieved its best result of 12.26% mAP40 (see Figure 3). The head-only finetuning of Foundation X, moreover, has significantly fewer trainable parameters than the baseline model (107M) but outperforms the latter by achieving its best result of 14.29% mAP40 at the 101st epoch. Full finetuning of Foundation X, on the other hand, surpassed both head-only finetuning and baseline at epoch 49 with 13.10% mAP40 and went on to achieve the best result of 15.51% mAP40 at the 112th epoch. This reinforces that Foundation X, by leveraging knowledge from diverse datasets and utilizing Cyclic and Lock-Release pretraining, boosts performance, even with minimal (head-only) finetuning.

6. Conclusion

In this study, we introduce Foundation X, an advanced model for chest X-ray analysis designed to handle classification, localization, and segmentation tasks with a shared backbone. By leveraging the Cyclic and Lock-Release pretraining strategies, Foundation X achieves significant performance improvements across diverse datasets, confirming its capability for combined task learning. Foundation X surpasses baselines across various datasets and tasks while maximizing the utilization of all available annotations. This efficiency reduces annotation costs and enhances the effectiveness of data analysis and processing in medical image analysis. Overall, Foundation X is a robust and versatile solution for advancing medical imaging technology. Currently, our work focuses on Chest X-ray images, and we plan to extend it to other imaging modalities as future work.

References

[1] Siim-acr pneumothorax segmentation data. https://www.kaggle.com/jesperdramsch/siimacrpneumothorax-segmentation-data, 2019. 2, 5, 6

[2] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997. 2

[3] Sijing Feng, Qixiu Liu, Aakash Patel, Sibghat Ullah Bazai, Cheng-Kai Jin, Ji Soo Kim, Mikal Sarrafzadeh, Damian Azzollini, Jason Yeoh, Eve Kim, et al. Automated pneumothorax triaging in chest x-rays in the new zealand population using deep-learning algorithms. *Journal of medical imaging and radiation oncology*, 66(8):1035–1043, 2022. 1, 2, 3, 5, 6

[4] Yutong Feng, Biao Gong, Jianwen Jiang, Yiliang Lv, Yujun Shen, Deli Zhao, and Jingren Zhou. ViM: Vision Middleware for Unified Downstream Transferring. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11662–11673, Paris, France, Oct. 2023. IEEE. 3, 4

[5] Nicolás Gaggion, Candelaria Mosquera, Lucas Mansilla, Julia Mariel Saidman, Martina Aineseder, Diego H Milone, and Enzo Ferrante. Chexmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. *Scientific Data*, 11(1):511, 2024. 2, 5, 6

[6] Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8856–8865, 2021. 1

[7] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[8] Yinan He, Gengshi Huang, Siyu Chen, Jianing Teng, Kun Wang, Zhenfei Yin, Lu Sheng, Ziwei Liu, Yu Qiao, and Jing Shao. X-learner: Learning cross sources and tasks for universal visual representation. In *European Conference on Computer Vision*, pages 509–528. Springer, 2022. 1, 3

[9] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 1, 2, 6

[10] Nahid Ul Islam, Zongwei Zhou, Shiv Gehlot, Michael B Gotway, and Jianming Liang. Seeking an optimal approach for computer-aided diagnosis of pulmonary embolism. *Medical image analysis*, 91:102988, 2024. 1

[11] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014. 2, 5, 6

[12] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a deidentified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 2, 6

[13] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 4

[14] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3041–3050, June 2023. 2

[15] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021. 2

[16] Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021. 2, 6

[17] Feng Lin, Wenze Hu, Yaowei Wang, Yonghong Tian, Guangming Lu, Fanglin Chen, Yong Xu, and Xiaoyu Wang. Universal Object Detection with Large Vision Model. *International Journal of Computer Vision*, 132(4):1258–1276, Apr. 2024. 3

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[19] Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2646–2655, 2020. 1, 2, 6

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6, 1

[21] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[22] DongAo Ma, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Foundation ark: Accruing and reusing knowledge for superior and robust performance. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 651–662, Cham, 2023. Springer Nature Switzerland. 4, 5, 6, 7, 8, 1

[23] Aakarsh Malhotra, Surbhi Mittal, Puspita Majumdar, Saheb Chhabra, Kartik Thakral, Mayank Vatsa, Richa Singh, Santanu Chaudhury, Ashwin Pudrod, and Anjali Agrawal. Multi-task driven explainable diagnosis of covid-19 using chest x-ray images. *Pattern Recognition*, 122:108243, 2022. 2

[24] Hoang C Nguyen, Tung T Le, Hieu H Pham, and Ha Q Nguyen. Vindr-ribcxr: A benchmark dataset for automatic segmentation and labeling of individual ribs on chest x-rays. *arXiv preprint arXiv:2107.01327*, 2021. 2, 5, 6

[25] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022. 2, 7, 8, 6

[26] Radiological Society of North America. Rsna pneumonia detection challenge. *Kaggle*, 2018. 2, 6

[27] Jiaxuan Pang, Fatemeh Haghighi, DongAo Ma, Nahid Ul Islam, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Popar: Patch order prediction and appearance recovery for self-supervised medical image analysis. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 77–87. Springer, 2022. 6, 7

[28] Clément Playout, Renaud Duval, and Farida Cheriet. A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images. *IEEE transactions on medical imaging*, 38(10):2434–2444, 2019. 2

[29] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, Seoul, Korea (South), Oct. 2019. IEEE. 3

[30] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American journal of roentgenology*, 174(1):71–74, 2000. 5

[31] Ecem Sogancioglu, Bram van Ginneken, Finn Behrendt, Marcel Bengs, Alexander Schlaefer, Miron Radu, Di Xu, Ke Sheng, Fabien Scalzo, Eric Marcus, et al. Nodule detection and generation on chest x-rays: Node21 challenge. *arXiv preprint arXiv:2401.02192*, 2024. 2, 6

[32] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 5, 1

[33] Bram Van Ginneken, Mikkel B Stegmann, and Marco Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis*, 10(1):19–40, 2006. 2, 6

[34] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022. 2

[35] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 2, 6

[36] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 4, 5, 6, 1

[37] Lijian Xu, Ziyu Ni, Xinglong Liu, Xiaosong Wang, Hongsheng Li, and Shaoting Zhang. Learning a multi-task transformer via unified and customized instruction tuning for chest radiograph interpretation. *arXiv preprint arXiv:2311.01092*, 2023. 2

[38] H Zhang, F Li, S Liu, L Zhang, H Su, J Zhu, LM Ni, and HY Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arxiv 2022. *arXiv preprint arXiv:2203.03605*, 5, 2022. 4, 5, 6, 1

[39] Yan Zhao, Xiuying Wang, Tongtong Che, Guoqing Bao, and Shuyu Li. Multi-task deep learning for medical image computing and analysis: A review. *Computers in Biology and Medicine*, 153:106496, 2023. 1

[40] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021. 2