Data-Assemble: Leveraging Multiple Datasets with Partial Labels

Mintong Kang\(^1\) Yongyi Lu\(^2\) Alan L. Yuille\(^2\) Zongwei Zhou\(^2,\ast\)
\(^1\)Zhejiang University \(^2\)Johns Hopkins University

Abstract

The success of deep learning relies heavily on large and diverse datasets with extensive labels, but we often only have access to several small datasets associated with partial labels. In this paper, we start a new initiative, “Data-Assemble”, that aims to unleash the full potential of partially labeled data from an assembly of public datasets. Specifically, we introduce a new dynamic adapter to encode different visual tasks, which addresses the challenges of incomparable, heterogeneous, or even conflicting labeling protocols. We also employ pseudo-labeling and consistency constraints to harness data with missing labels and to mitigate the domain gap across datasets. From rigorous evaluations on three natural imaging and six medical imaging tasks, we discover that learning from “negative examples” facilitates both classification and segmentation of classes of interest. This sheds new light on the computer-aided diagnosis of rare diseases and emerging pandemics, wherein “positive examples” are hard to collect, yet “negative examples” are relatively easier to assemble. Apart from exceeding prior arts in the ChestXray benchmark, our model is particularly strong in identifying diseases of minority classes, yielding over 3-point improvement on average. Remarkably, when using existing partial labels, our model performance is on-par with that using full labels, eliminating the need for an additional 40% of annotation costs. Code will be made available at https://github.com/MrGiovanni/DataAssemble.

1. Introduction

Recent years have witnessed an increasing number of datasets becoming publicly available thanks to the collective efforts of imaging data archives and international competitions [2, 6, 14, 22, 66]. These datasets are collected, organized, and annotated differently, and are often limited in their own scope. Very few studies have been done to unleash the full potential of an assembly of multiple datasets with partial labels. The challenge is that labels in those datasets are often incomparable, heterogeneous, or even conflicting [50, 73]. Therefore, in this paper, we ponder the question: How can we assemble and exploit such a great number of isolated datasets?

To address this question, we start with probing the principal hypothesis: a dataset that is labeled with various classes can foster more powerful models than one that is only labeled with the class of interest. Consequently, we start a new initiative, “Data-Assemble”, for leveraging an assembly of datasets with partial labels. Specifically, we introduce a new dynamic adapter to encode different visual tasks, which can dynamically integrate partial labels across different datasets. Studies in §5.5 presents four unique advantages of our dynamic adapter: (i) discovering novel classes, (ii) benefiting multi-label classification, and (iii) reducing computational costs, and (iv) establishing inter-class relationships. Furthermore, pseudo-labeling and consistency constraints are used for the missing part of labels and to mitigate the domain gap across different datasets.

We have conducted extensive experiments on three natural imaging datasets and six medical benchmarks. Our results in §5 demonstrate that (i) assembling data with strictly non-overlapping labels improves the performance;
(ii) learning from a mixture of partial labels performs on par with that from full labels; (iii) annotating “negative examples” facilitates long-tail and rare disease classification; and (iv) exceeding prior arts on NIH ChestXray. These results are attributable to our simple yet powerful observation (illustrated in Figure 1): learning from the classes of “negative examples” can better delimit the decision boundary of the class of interest. This suggests that besides chasing for the class of interest.

To our best knowledge, we are among the first to systematically examine the rationale of assembling multiple datasets and fully exploit the potential of partial labels—especially if the interested object is of a minority class—e.g. rare diseases and emerging pandemics.

To our best knowledge, we are among the first to systematically examine the rationale of assembling multiple datasets and fully exploit the potential of partial labels—the latest attempts [18, 50, 68] built models on the labeled part of the data only. In summary, we make the following four contributions:

1. We devise a novel framework that consists of (i) a dynamic adapter with learnable task encoding to harness partially labeled data and (ii) pseudo labeling & consistency constraints to exploit unlabeled data, while mitigating the domain gap across datasets (see §3).

2. We start a new initiative, namely “Data-Assemble”, and establish the feasibility and effectiveness of assembling multiple datasets with partial labels (see §5.1).

3. We demonstrate that using existing partial labels, our model can achieve comparable performance to that using a fully curated dataset with exhaustive labels, eliminating the need for an additional 40% of annotation costs in thorax disease classification (see §5.2).

4. We present a unique annotation scheme for computer-aided diagnosis of rare diseases and emerging pandemics, assembling existing labeled of related diseases rather than narrowly pursuing extensive labels for the class of interest (see §5.3).

2. Preliminary

2.1. Principal Hypothesis

We hypothesize that a dataset that is labeled with various classes can foster more powerful models than one that is only labeled with the class of interest. To validate this point, we begin with a toy example: MNIST-zero is created from the original MNIST dataset [33], wherein the images with the number “0” are labeled as positives, and the rest images are negatives; the goal is to train a model that can recognize the images with “0”. Note that the total numbers of images are the same between MNIST-zero and MNIST—the only variation is that the makeup of negatives is unknown in MNIST-zero, yet it is known in MNIST. Figure 1 shows that MNIST-zero decreases the classification of images with “0”. This is attributed to the lack of fine-grained labels (e.g. 1–9) in negative examples, which potentially causes the confusion between zero-like “6” and “0”. In Table 1, similar experiments have been conducted on the tasks of segmentation and classification, revealing a consistent observation over the seven datasets. Taking semantic segmentation as another example, the model can segment “bus” from the image more precisely if it is exposed to all 19 classes in Cityscapes in contrast with that exposed to five partial classes only. Now, we reach the conclusion that learning from the classes of “negative examples” can better delimit the decision boundary of the class of interest, verified in both segmentation and classification tasks. This is the foundation of the “Data-Assemble” initiative, underlining the necessity of combining multiple datasets with diverse (yet partial) labels.

2.2. Problem Definition

We consider a dataset to be partially labeled if at least one data point that belongs to the set of labeled classes has not been annotated. Formally, given a dataset of $D = \{ (x_1, y_1), (x_2, y_2), ..., (x_N, y_N) \}$, there are a total of $M (M > 0)$ unique classes labeled in the ground truth $\{ y_1, y_2, ..., y_N \}$. For $\forall x_i \in D$, if the presence of $\forall M$ classes in $y_i$ is annotated, $D$ is a fully labeled dataset; otherwise, $D$ is a partially labeled dataset. Under this circumstance,

Table 1. A model can better segment “foreground” pixels if it recognizes classes in the “background”; similarly, a model can better classify “positive” images if it identifies classes in the “negative”. The gaps in color are statistically significant.

| Toy example of segmentation | IoU of “Bus”(%) |
|-----------------------------|------------------|
| Cityscapes [15]             |                  |
| Zero in 5-way               |                  |
| 5-way                      |                  |
| Bus in 5-way                |                  |
| Bus in 19-way†              |                  |

† 5-way includes: Bus, Road, Sidewalk, Building, and Wall.

| Toy examples of classification | mAUC of interest (%) |
|-------------------------------|----------------------|
| MNIST [33]                    |                      |
| Zero vs. Others               | 98.8                 |
| 10-way                        | 99.7 (± 0.9)         |
| CIFAR-10 [30]                 |                      |
| Cat vs. Others                | 86.3                 |
| 10-way                        | 87.5 (± 1.2)         |
| DermaMNIST [65]               |                      |
| Mel.† vs. Others              | 81.1                 |
| 7-way                         | 86.8 (± 5.7)         |
| TissueMNIST [65]              |                      |
| DCT.† vs. Others              | 79.2                 |
| 8-way                         | 82.9 (± 3.7)         |
| ChestXray [57]                |                      |
| 5-way**                       | 80.9                 |
| 14-way†                       | 82.1 (± 1.2)         |
| CheXpert [26]                 |                      |
| 5-way**                       | 82.6                 |
| 13-way†                       | 87.2 (± 4.9)         |

*“Mel.” denotes Melanoma, “DCT.” denotes Distal Convoluted Tabule
** 5-way includes: Atelectasis, Cardiomegaly, Consolidation, Edema, and Effusion
• An unlabeled dataset is neither partially labeled or fully labeled \((M = 0)\).

• CheXpert\(^2\) is fully labeled \((M = 13)\) because for \(\forall x_i\), the presence of 13 classes has been annotated.

• ChestXray\(^3\) is fully labeled \((M = 14)\) because for \(\forall x_i\), the presence of 14 classes has been annotated.

• The union of CheXpert and ChestXray is, however, partially labeled \((M = 20)\) because for \(\forall x_i\) in CheXpert, the presence of six classes in ChestXray has not been annotated; and for \(\forall x_i\) in ChestXray, the presence of seven classes in CheXpert has not been annotated.

3. Data-Assemble

Dynamic Adapter with Learnable Task Encoding: To learn from partially-labeled datasets, we develop a dynamic adapter with learnable task encoding and train it in a question-and-answer manner. Similar to the controller in Zhang et al. [68], our task encoding is a numeral vector, but ours can be learnable during the training phase. As a result, the learned encoding will become soft-label vectors and establish an inter-class relationship, which is much more informative than conventional one-hot vectors (all classes are orthogonal among each other). Our dynamic adapter consists of a convolutional layer that generates parameters according to the task encoding, followed by a linear classification layer to produce the output answer. As shown in Figure 2, given task encoding \((q)\) and input image \((x)\), our dynamic adapter can compute the answer \((a)\) as follows,

\[
a = w(q; \theta_w) \ast f(x),
\]

where \(*\) is the inner product operation, \(w\) denotes the convolutional layer of the adapter, and \(f\) represents the feature extractor. Subsequently, binary cross entropy loss is used if the label is provided, i.e.

\[
L_{bce} = - (y \cdot \log(a) + (1 - y) \cdot \log(1 - a)),
\]

where \(y\) is the right answer based on the ground truth.

Pseudo Labeling & Consistency Constraints: To unleash the full potential of unannotated labels, we introduce a sharpening operator to generate pseudo-labels, i.e.

\[
\tilde{a} = \begin{cases} 
  a + (1 - a) / t, & a > \tau \\
  a - a / t, & a \leq \tau 
\end{cases}
\]

where \(\tilde{a}\) is the pseudo-label of the answer, \(t\) is the sharpen temperature, and \(\tau\) is the threshold \((\tau = 0.5\) in our experiments). The prediction beyond \((\text{below})\) the threshold \(\tau\) can be assigned to a higher \((\text{lower})\) score controlled by \(t\). If \(t = \infty\), there is no pseudo-labeling; if \(t = 1\), the model converts a soft label to a completely hard label (either 1 or 0, equivalent to FixMatch et al. [51]).

Overall Loss Functions: The overall loss function consists of binary cross-entropy regularization for annotated labels as well as pseudo-labeling & consistency constraints for unlabeled ones, i.e.

\[
L_{\text{total}} = L_{bce} + L_{\text{pseudo}} + L_{\text{consist}},
\]
Algorithm 1 PyTorch pseudocode

```python
# f: feature extractor
# q: task encoding (nxn)
# n: number of total labels
# m: dimension of features extracted by f
# aug_w, aug_s: weak and strong augmentation
# q: task encoding (nxn)
# f: feature extractor

val = torch.eye(n).cuda()  # initialize q
w = conv2d(n, m, kernel_size=1, stride=1)  # parameters (nxm)
# aug_w, aug_s: weak and strong augmentation
# m: dimension of features extracted by f
# n: number of total labels

# Lbce, Lpseudo and Lconsist are formulated in Eq. 2, Eq. 4, and Eq. 5. The PyTorch pseudocode of our proposed framework has been provided in Alg. 1.

4. Experiment

4.1. Dataset & Metric

We examine our framework on six medical imaging datasets and three natural imaging datasets: CheXpert2 [26], ChestXray1 [57], DermaMNIST, TissueMNIST, OrganAMNIST, RetinaMNIST, MNIST [33], CIFAR10 [30], and Cityscapes [15]. Appendix A–C provide training recipe along with further details of each dataset. As defined in §2.2, the union of ChestXray and CheXpert can be considered as a partially-labeled dataset, as they are derived from two different institutes and offer labels for 21 classes in total, with only seven overlapping classes. Considering the limited number of images in the CheXpert official val set, we separate the training set to 166,739 images and 56,675 images for a new training set and a test set (the proportion is the same as ChestXray official split).

Following prior metrics for benchmarking, we evaluate the performance using Area Under the Curve (AUC) for thorax disease classification and Intersection over Union (IoU) for semantic segmentation. For all experiments, we present statistical analysis based on an independent two-sample t-test. Code will be available at https://github.com/MrGiovanni/DataAssemble.

4.2. Baseline & Implementation

For comparison, we must re-implement existing methods for the task of multi-label classification because most methods [17, 19, 68] were proposed for the task of multi-organ segmentation. We have compared our method with DoDNet [68] (more precisely, with its controller) since it was one of the most recent works in this area. Predominant semi-supervised methods such as FixMatch [51] were compared as well. We have also justified the superiority of our method to the multi-network strategy [37] (one-model-one-task) in §2.1. In addition, we set an upper-bound performance using fully labeled datasets as a strong reference in Tables 3–4. Finally, we benchmark our method against the state-of-the-art methods [7, 20, 21, 27, 39, 48, 56] on ChestXray (§5.4 and Appendix §A).

We choose DenseNet121 [25] as the backbone. All experiments run 64 epochs and utilize Adam optimizer [28] with an initial learning rate of 2e-4. We reduce the learning rate by a factor of 2.0 on the plateau with 5 steps of patience. Early stopping patience is set to be 10 epochs. The pseudo-label threshold γ and the sharpen temperature t are 0.5 and 4.0, respectively. Ablation study of the sharpen temperature t is provided in Table 11.

5. Result

This section summarizes four major findings in our extensive experiments, followed by four unique properties of our framework.

5.1. Assembling Data with Strictly Non-overlapping Labels Improves the Performance

In medical imaging, labels in different datasets are often non-overlapping, e.g. KiTS [23] and LiTS [9]. To demonstrate the effectiveness of our framework in this scenario, we conduct experiments on 15,062 images from ChestXray and CheXpert; each includes 7,531 images. We split the seven overlapping classes of ChestXray and CheXpert into two separate label sets, so that these two datasets do not have a single label in common. Detailed splits and their assembling results are provided in Table 2. We can see that by learning images with the other three diseases (i.e. Emphysema, Consolidation, and Pneumothorax) from ChestXray, the performance of identifying the four diseases (i.e. Cardiomegaly, Pneumonia, Atelectasis, Edema) on CheXpert is increased by a large margin (56.0%→73.5%). Similar observations hold on ChestXray, showing a mAUC improvement (69.9%→74.9%) for identifying the three diseases.

We also conduct experiments on CIFAR10 and MNIST to strengthen the generalizability of this observation. In CIFAR10, we split the ten classes into two separate label sets (i.e. animals and transportation), and in MNIST, all the digits are divided into two strictly non-overlapping sets (i.e. odd numbers and even numbers). We can see that by learning images with transportation (i.e. airplane, automobile, ship, truck), the performance of identifying the animals (i.e. bird, cat, deer, dog, frog, horse) is increased (mAUC is improved from 89.5% to 90.2%). We also observed similar results on transportation classes for identifying the six ani-
Table 2. The results on Chest X-rays, CIFAR10, and MNIST suggest that assembling data with strictly non-overlapping labels substantially improves the performance of each label. For Chest X-rays, we examine two subsets of chest diseases, i.e., $D_0 = \{\text{Cardiomegaly, Pneumonia, Atelectasis, Edema} \}$ and $D_1=\{\text{Effusion, Consolidation, Pneumothorax} \}$; note that $D_0$ and $D_1$ are taken from the CheXpert and ChestXray datasets, respectively; both of them include 7,531 images. For CIFAR10, we examine two subsets of animals and transportation, i.e., $D_0 = \{\text{Bird, Cat, Deer, Dog, Frog, Horse} \}$ and $D_1=\{\text{Airplane, Automobile, Ship, Truck} \}$; both of them contain 25,000 images. For MNIST, we examine two subsets of odd and even numbers, i.e., $D_0 = \{1, 3, 5, 7, 9\}$ and $D_1=\{0, 2, 4, 6, 8\}$; both of them contain 300 images. The gaps in color are statistically significant ($p < 0.05$).

| Dataset          | $D_0$ | $D_0 \& D_1$ (partial) | (full) |
|------------------|-------|------------------------|--------|
| # of labels      | 37,655| 75,310                 | 105,434|
| Strategy         | $L_{bce}$ | $L_{bce} + \text{adapter}$ | $L_{\text{pseudo}}$ | $L_{\text{consist}}$ | $L_{\text{pseudo}} + L_{\text{consist}}$ | $L_{\text{bce}}$ |
| mAUC (%)         | 62.6 (16.7) | 74.3 (14.0) | 74.6 (14.7) | 78.7 (10.9) | 78.3 (11.0) | 79.0 (10.9) | 79.5 (10.2) | 79.3$^f$ |
| Cardiomegaly     | 64.6 (19.8) | 75.0 (18.5) | 77.9 (16.6) | 83.5 (10.0) | 82.6 (10.9) | 83.6 (10.1) | 83.9 (10.4) | 83.5 |
| Pneumonia        | 48.1 (22.2) | 62.9 (15.4) | 59.8 (18.5) | 66.3 (14.2) | 65.6 (14.7) | 67.6 (14.7) | 67.9 (14.4) | 68.3 |
| Atelectasis      | 43.1 (26.8) | 66.3 (13.6) | 64.9 (10.0) | 70.0 (7.0) | 69.9 (10.0) | 69.5 (10.0) | 70.5 (8.0) | 69.9 |
| Edema            | 79.1 (17.4) | 83.9 (12.5) | 85.6 (9.9) | 85.8 (15.4) | 85.4 (14.1) | 85.8 (15.7) | 86.7 (7.3) | 86.4 |
| Effusion         | 80.0 (18.5) | 83.6 (14.9) | 84.7 (13.8) | 88.0 (10.5) | 88.0 (10.5) | 88.4 (10.1) | 88.6 (10.1) | 88.5 |

1) Training with $L_{bce}$, $L_{\text{pseudo}}$, and $L_{\text{consist}}$ on full labels can lead to a mAUC of 79.1%, which is comparable to $L_{bce}$ only (79.3%).

Table 3. The results on CheXpert (val) reveal that (1) learning from a mixture of 75,310 partial labels performs on par with that from 105,434 full labels, therefore eliminating the need of additional 40% annotation costs; (2) learning from alternative classes boosts the performance of classifying the interested disease, e.g., learning from Consolidation and Pneumothorax results in a noticeable improvement of Cardiomegaly (64.6%→83.9%) and Pneumonia (46.1%→67.9%) classification. In both Table 3 and Table 4, $D_0=\{\text{Cardiomegaly, Pneumonia, Atelectasis, Edema, Effusion} \}$ and $D_1=\{\text{Atelectasis, Edema, Effusion, Consolidation, Pneumothorax} \}$. Note that $D_0$ and $D_1$ are taken from the CheXpert and ChestXray datasets, respectively; both of them include 7,531 images. The gaps in color are statistically significant ($p < 0.05$).

5.2. Learning from a Mixture of Partial Labels Performs on Par with That from Full Labels

We explore the effectiveness of our framework on the datasets that share partial labels in common. Similar to §5.1, we as well utilize 15,062 images from ChestXray and CheXpert with seven diseases labeled; the difference lies in their label sets—three out of the seven diseases are shared across the two datasets in this setting. Table 3 shows the assembling results on CheXpert. Compared to the model trained only on CheXpert (with five labels), including more data from ChestXray in the training phase brings about 11.7% improvement (62.6%→74.3%) of mAUC. Note that the number of labels for Cardiomegaly and Pneumonia remains the same before and after assembling the two datasets, yet by learning from other disease in ChestXray, they achieve an remarkable improvement of 10.4% and 16.8%, respectively. Moreover, pseudo-labeling and consistency constraints enable the model to leverage unlabeled data (e.g., Consolidation and Pneumothorax are not labeled in CheXpert) and the model yields an mAUC improvement of 4.7 points (74.3%→79.0%). Finally, our dynamic adapter is demonstrated to produce another 0.5% improvement, leading to the final mAUC of 79.5%, which
is on par with the performance of the upper bound model trained with full labels (79.3%), while eliminating an additional annotation cost of 40% (75,310 labels vs. 105,434 labels). Similar performance gain is also observed on ChestXray (see Table 4). This observation (along with more experiments in Appendix Table 12) indicates that assembling multiple partially labeled datasets is not required to fulfill all in-completed labels. With our dynamic adapter and semi-supervised learning methods, learning from an assembly of partial labels can perform on par with that from full labels.

### 5.3. Annotating “Negative Examples” Facilitates Long-tail and Rare Disease Classification

We have shown that learning with additional “negative examples” improves the performance of the class(es) of interest, but how different classes of “negative examples” contribute to the performance remains unknown. We further studied this problem and found that the performance gain is positively related to the similarity between the class of interest and the added classes. Figure 3 displays the similarity between Nodule and other 13 diseases in ChestXray; Table 5 illustrates the improvements of classifying “Nodule” by learning these 13 diseases. The Pearson correlation coefficient between the similarity and performance gain is 0.83, which indicates a significant positive correlation (p = 4.93e-4). This means that co-learning with similar classes is more beneficial than dissimilar classes for the class of interest (i.e. Nodule in this setting). However, it is hard to obtain enough ground truth labels for training since rare diseases have limited positive examples. By incorporating similar diseases from other datasets, the model can better identify rare diseases, thus relieving the long-tail problem in computer-aided diagnosis.

### 5.4. Exceeding the Prior Art in NIH ChestXray

Table 6 demonstrate that our method is significantly superior to predominant methods on ChestXray classification.
Table 6. In comparison with the state-of-the-art methods on the ChestXray dataset (official split), our method achieves the best mean performance over all 14 thorax diseases and the best performance (mAUC) over eight diseases.

| Dataset          | Novelty      | Adapter | mAUC (%) |
|------------------|--------------|---------|----------|
| MNIST            | Zero         | ✗       | 72.4     |
|                  |              | ✓       | 77.8     |
| CIFAR10          | Cat          | ✗       | 83.0     |
|                  |              | ✓       | 84.6     |
| DermaMNIST       | Vascular lesion | ✗    | 75.2     |
|                  |              | ✓       | 78.6     |
| TissueMNIST      | Proximal tubule | ✗    | 79.6     |
|                  |              | ✓       | 81.0     |
| OrganAMNIST      | Heart        | ✗       | 97.2     |
|                  |              | ✓       | 98.6     |

Table 7. Our dynamic adapter is more robust to novel class attacks, knowing what it does not known. The novel class occurs in the test set only, not in the training or validation set. More results can be found in §B. The gaps in color are statistically significant.

| Dataset          | Adapter | mAUC (%) |
|------------------|---------|----------|
| MNIST-muti†      | ✗       | 88.2     |
|                  | ✓       | 90.9     |
| CheXpert         | ✗       | 87.4     |
|                  | ✓       | 88.0     |
| ChestXray        | ✗       | 81.2     |
|                  | ✓       | 81.6     |

†MNIST-muti: Multiple numbers can occur in one image (detailed in §C.4).

Table 8. Our dynamic adapter benefits multi-label classification in both medical and natural imaging datasets. Training with dynamic adapter can alleviate overfitting to the underlying pattern of one-image-one-label in some manually curated datasets and faithfully identify more than one labels, if present, in the unseen image.

5.5. Ablating Key Properties

(1) Discovering Novel Classes: Models trained with conventional strategies tend to be overconfident when encountering novel classes in the test set [29]. However, the model is expected to be aware of out-of-distribution data, rather than giving certain (but erroneous) predictions. Our dynamic adapter advantageously trains a model that knows what it does not know (see Table 7). For example, we train the model using images of numbers 1–9 and ask it to predict images of the number zero in the inference. Our dynamic adapter behaves more robustly to the novel class attacks than conventional training (↑5.4%). This property enables detecting the long tail of unseen conditions.

(2) Benefiting Multi-label Classification: Numerous medical imaging tasks require multi-label classification, which allows each image to belong to multiple classes [41]. Without extra parameters, we discover that our dynamic adapter—learning in a question-and-answer manner—is a simple yet effective strategy to manage multi-label classification, compared with previous state-of-the-art approaches without the adapter [20, 45] (see Table 8). We attribute this benefit to the flexibility of question-and-answer strategy: given an image of “0”, one can create a great number of questions for the model to answer (e.g. “is 0 in the image?”, “is 100 in the image?”, and many more); whereas the conventional training is largely tailored to the provided ground truth (essentially, only dealing with “is 0 in the image?”).

(3) Reducing Computational Costs: We have devised a multi-question training strategy that enables the model to predict answers for all labels at once. Then, $L_{bce}$ will be applied if the answer is given in the ground truth; otherwise, $L_{consist}$ and $L_{pseudo}$ will be applied. DoDNet [68], one of the latest works in efficient learning from partially labeled data, generated a single answer of one class of interest every time. Compared with DoDNet, our multi-question strategy not only reduces the model parameters by 13% (8.02 MB → 6.98 MB), elevates the performance by over 1-point (80.5% → 81.6% on ChestXray), but also substantially shortens the training time for each epoch by nearly 90% (150 min → 16 min).

(4) Establishing Inter-class Relationships: We have introduced learnable task encoding to discover the relationship
among classes (see Figure 3). Labels in the dataset are often formatted in one-hot vectors, therefore being mutually orthogonal, e.g. 14 diseases in ChestXRay are absolutely independent. In the actual scenario, however, thorax diseases retain an account of co-occurrence [57]. Consequently, our dynamic adapter is fundamentally different from [11] and [68], which encoded tasks as one-hot vectors—the former concatenates the one-hot vector with the input image, and the latter concatenates with the image feature. In addition, the established relationship provides important guidance for the study of rare disease classification (see Table 5). The similarity among classes are calculated by the mean feature of a single class, and we shift the origin point of the feature space to the mean vector of all features to make the similarity vary in a larger interval.

6. Related Work

Partially Labeled Medical Datasets: Most medical datasets are partially labeled, such as liver and liver tumors [35,69], kidney and kidney tumors [24,44], and CT images for lung cancer diagnosis [3]. Assembling multiple datasets to train a single model has been attempted in a line of research. Med3D [12] trains a shared encoder and task-specific decoders for eight domains. PIPO-FAN [18] employs a deep supervision mechanism to refine the outputs in different scales and designs an adaptive weighting layer to fuse the outputs. LENS [63] learns from multiple heterogeneous lesion datasets in a multi-task fashion and mines missing annotations through prior knowledge. Marginal loss and exclusion loss are proposed to make the model learn a union of datasets more efficiently [50]. DoD-Net [68] uses fixed task encoding concatenated with image features to dynamically generate the parameters of the classification heads. However, the underlying rationale behind the assembling strategy remains unjustified. In this work, we design a systematic framework to accommodate different datasets as well as empirically justifying the rationale of data assemble.

Dynamic Filter Learning: Dynamic filter learning has drawn considerable research attention in the computer vision community due to its adaptive nature. Some recent approaches dynamically adjust the filter neighborhoods by adaptive dilation factors [70], estimating the neighborhood sampling grid [16], or adapting the receptive fields [52]. Another line of research generates filter values dynamically based on input features. WeightNet [38], CondConv [64], DyNet [71], and DynamicConv [13], predict coefficients to aggregate template filters to generate the prediction. [10,55,59] generate dynamic filters for every pixel adaptively. Moreover, [68] uses the concatenation of image features and task encoding to dynamically generate parameters for classification heads. Our dynamic adapter differs from existing works. First, our adapter generates the parameters of the classification heads conditionally only on task encoding, thus reducing the computational cost. Second, our task encoding is learnable, which benefits establishing inter-class relationships and multi-label classification.

Semi-supervised Learning: Semi-supervised algorithms rely on the assumptions of smoothness and low-density separation in the feature space. One popular class of SSL methods is producing artificial pseudo-labels. The model can provide artificial labels for unlabeled images based on confidence and entropy. For instance, pseudo-labeling [34,60] (a.k.a. self-training) utilizes the model’s class predictions as labels to train the model. Similarly, consistency regularization [5,32,46,47,61] obtains an artificial label using the model’s predicted distribution after randomly modifying the input or model function. Moreover, Virtual Adversarial Training (VAT) [40,42,43] represents another line of research that seeks to utilize adversarial examples. It employs a regularization that trains the distribution of the target label (conditional on input features) to be isotropically smooth around each input data point. Another kind of method builds a graph in which each node represents a sample, while the graph edges connect nodes that are likely to have the same label. A learning function such as the label propagation [4,72] or a constraint such as manifold regularization [8] is applied over the graph to deduce the label for unlabeled samples. These works focus on image-level semi-supervised learning in which the pseudo-labels or consistency constraints are generated or applied on images. However, our work proposes to utilize label-wise semi-supervised learning to generate pseudo-labels and apply consistency constraints on each not annotated label, thus mitigating the domain gaps across datasets.

7. Conclusion

We start a new initiative, “Data-Assemble”, exploring the full potential of an assembly of datasets with partial labels. This work represents the first step towards creating large-scale, multi-center, fully-labeled medical datasets—one of the foundations of fostering future research in deep learning applied to medical images. A key contribution of ours is to empirically justify the rationale of data assemble: Learning from classes of “negative examples” can better delimit the decision boundary of the class of interest. We have also developed a novel framework that consists of (i) a dynamic adapter with learnable task encoding to harness partially labeled data and (ii) pseudo labeling & consistency constraints to exploit enormous unlabeled data, while mitigating the domain gap across datasets. Finally, our results shed new light on computer-aided diagnosis of rare diseases and emerging pandemics, underlining the role of an assembly of existing labels of related diseases, in contrast to narrowly pursuing exhaustive labels for the class of interest.
Acknowledgments

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research. We thank M. R. Hossein-zadeh Taher and F. Haghighi for surveying top solutions on the ChestXray benchmark (Tab. 6); thank R. Feng for re-producing baseline methods on ChestXray. We also thank the constructive suggestions from Dr. J. Liang, Dr. Z. Zhu, L. Chen, J. Chen, Y. Bai, G. Li, and X. Li.

References

[1] Imane Allaouzi and Mohamed Ben Ahmed. A novel approach for multi-label chest x-ray classification of common thorax diseases. *IEEE Access*, 7:64279–64288, 2019. [12]

[2] Michela Antonelli, Annika Reinke,Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021. [1]

[3] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wexing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019. [8]

[4] Angelica I. Aviles-Rivero, Nicolas Papadakis, Ruoteng Li, Philip Sellars, Qin Nan Fan, Robby T. Tan, and Carola-Bibiane Schönlieb. GraphX$^{NET}$– chest x-ray classification under extreme minimal supervision, 2020. [8]

[5] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles, 2014. [8]

[6] Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. [1]

[7] Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):1–10, 2019. [4, 7, 12]

[8] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006. [8]

[9] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. [4]

[10] Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks, 2016. [8]

[11] Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast image processing with fully-convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2497–2506, 2017. [8]

[12] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. [8]

[13] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels, 2020. [8]

[14] Errol Colak, Felipe C Kitamura, Stephen B Hobbs, Carol C Wu, Matthew P Lungren, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Robyn L Ball, George Shih, Anouk Stein, et al. The rsna pulmonary embolism ct dataset. *Radiology: Artificial Intelligence*, 3(2):e200254, 2021. [1]

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*, pages 3213–3223, 2016. [2, 4, 14]

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. [8]

[17] Konstantin Dmitriev and Ari E. Kaufman. Learning multi-class segmentations from single-class datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [4]

[18] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020. [2, 8]

[19] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction, 2020. [4]

[20] Qingji Guan and Yaping Huang. Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 2018. [2, 4, 7, 12, 13]

[21] Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Siqi Liu, Andreas Maier, and Dorin Comaniciu. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In *Iberoamerican Congress on Pattern Recognition*, pages 757–765, Springer, 2018. [4, 7, 12]

[22] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chummei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021. [1]

[23] Nicholas Heller, Niranan Sathianathan, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreicher, et al. The kits19 challenge data: 300 kidney tumor cases from labeled and unlabeled examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition of the IEEE conference on computer vision and pattern recognition (CVPR)*, June 2019. [4]

[24] Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Jiaping Wang, Chuanfeng Lv, Guotong Xie, and Yang Nan. A triple-stage self-guided network for kidney tumor segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 341–344. IEEE, 2020. [8]
[25] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017. 4, 13

[26] 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. 2, 3, 4, 12, 13

[27] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: Diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15719–15728, June 2021. 4, 7, 12

[28] D Kinga and J Ba Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, volume 5, 2015. 4, 13

[29] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International Conference on Machine Learning*, pages 5436–5446. PMLR, 2020. 7

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 4, 16, 17

[31] Pulkit Kumar, Monika Grewal, and Muktabh Mayank Srivastava. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs, 2017. 13

[32] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. 2, 3, 4, 12, 13

[33] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1378–1382. IEEE, 2019. 4, 7, 12, 13

[40] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, pages 1038–1042. IEEE, 2018. 8

[42] Andrei Margeloiu, Nikola Simidjievski, Mateja Jammik, and Adrian Weller. Improving interpretability in medical imaging diagnosis using adversarial training, 2020. 8

[43] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning, 2018. 8

[44] Andriy Myronenko and Ali Hatamizadeh. 3d kidneys and kidney tumor semantic segmentation using boundary-aware networks, 2019. 8

[45] Hieu H. Pham, Tung T. Le, Dat T. Ngo, Dat Q. Tran, and Ha Q. Nguyen. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels, 2020. 2, 7, 12

[46] Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations, 2020. 8

[47] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning, 2016. 8

[48] Laleh Seyyed-Kalantari, Guanxiang Liu, Matthew McDermott, Irene Y Chen, and Maryeh Ghassemi. Chexclution: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 232–243. World Scientific, 2020. 4, 7, 12

[49] Yan Shen and Mingchen Gao. Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization, 2018. 13

[50] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70:101979, 2021. 1, 2, 8

[51] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 3, 4, 12

[52] Domen Tabernik, Matej Kristan, and Aleš Leonardis. Spatially-adaptive filter units for compact and efficient deep neural networks. *International Journal of Computer Vision*, 128(8):2049–2067, 2020. 8

[53] Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. Attention-guided curriculum
learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pages 249–258. Springer, 2018. 7, 12, 13

[54] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation, 2020. 2

[55] F. Boray Tek, Ilker Çam, and Deniz Karlı. Adaptive convolution kernel for artificial neural networks, 2020. 8

[56] Hongyu Wang, Haozhe Jia, Le Lu, and Yong Xia. Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. *IEEE journal of biomedical and health informatics*, 24(2):475–485, 2019. 4, 7, 12

[57] 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, 3, 4, 8, 13

[58] 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. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 7, 12

[59] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation, 2020. 8

[60] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020. 8

[61] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 8

[62] Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. Efficient multiple organ localization in ct image using 3d region proposal network. *IEEE transactions on medical imaging*, 38(8):1885–1898, 2019. 13

[63] Ke Yan, Jinzheng Cai, Youjing Zheng, Adam P Harrison, Dakai Jin, You-bao Tang, Yu-Xing Tang, Lingyun Huang, Jing Xiao, and Le Lu. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct. *IEEE Transactions on Medical Imaging*, 2020. 8

[64] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *arXiv preprint arXiv:1904.04971*, 2019. 8

[65] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, Apr 2021. 2, 13

[66] Yang Yang, Xueyan Mei, Philip Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Katherine Link, Thomas Yang, Chendi Cao, et al. Radimagenet: A large-scale radiologic dataset for enhancing deep learning transfer learning research. 2021. 1

[67] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels, 2018. 13

[68] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1195–1204, 2021. 2, 3, 4, 7, 8

[69] Jianpeng Zhang, Yutong Xie, Pingping Zhang, Hao Chen, Yong Xia, and Chunhua Shen. Light-weight hybrid convolutional network for liver tumor segmentation. In *IJCAI*, volume 19, pages 4271–4277, 2019. 8

[70] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2031–2039, 2017. 8

[71] Yikang Zhang, Jian Zhang, Qiang Wang, and Zhao Zhong. Dynet: Dynamic convolution for accelerating convolutional neural networks, 2020. 8

[72] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, 2004. 8

[73] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10672–10681, 2019. 1
Table 11. Ablation study on the sharpen temperature \(\tau\) in Eq. 3. Compared with [51], the hyper-parameter \(t\) is introduced by our paper for the purpose of generating soft pseudo labels. To explore the impact of sharpen temperature on the performance of disease classification, we have conducted an ablation study to extensively search for the optimal value, spanning from 1.0 to 8.0, incremented by a factor of 2.0. For both CheXpert and ChestXray, \(t = 4.0\) achieves the best performance; therefore, we set \(t = 4.0\) across our experiments. Intuitively, for every prediction from the model, instead of directly converting it into a hard pseudo label (either 0 or 1), we generate the soft pseudo label by raising the predicted value towards 1.0 if the prediction is greater than the threshold \(\tau\) (for every prediction from the model, instead of directly converting it into a hard pseudo label (either 0 or 1), we generate the soft pseudo label by raising the predicted value towards 1.0 if the prediction is greater than the threshold \(\tau\)).

**Figure 4.** Class distribution of the ChestXray and CheXpert datasets. We discover that assembling heterogeneous and partially labeled datasets benefits more on the minority classes in the dataset. For the ChestXray dataset (left panel), a significant performance improvement lies in the top 3 minority classes, i.e., 6.4%, 1.5%, and 1.9% for Hern, Pneu1, and Fibr; whereas a relatively more minor improvement occurs in the top 3 majority classes, i.e., 0.5%, 0.3%, and 0.4% for Infi, Effu, and Atel. This observation enlightens computer-aided diagnosis of rare diseases and emerging pandemics, suggesting that instead of collecting sufficiently large amounts of labels for the disease of interest, assembling existing labels of related, common diseases can also be considered a practical approach.
A. Chest X-ray Benchmark

In [36], the authors perform disease classification and localization simultaneously by leveraging both class information as well as limited location annotation, showing that utilizing the location annotation improves the classification performance. Ma et al. [39] proposes a multi-attention network with cross-channel feature recalibration leveraging global information as well as local information. Tang et al. [53] uses an attention-guided curriculum learning method to recognize lesion localization and categories. Shen et al. [49] combines the routing-by agreement mechanism and the deep convolutional neural network. To model the relationships among diseases, Kumar et al. [31] designs a boosted cascaded convolutional network framework which is similar to the classifier chains. Yao et al. [67] uses Densenet [25] as an encoder and a Long-short Term Memory Network (LSTM) as a decoder to capture label correlations. In [20], a category-wise residual attention learning framework has been proposed to consider the correlation among relevant diseases and mitigate the interference of uncorrelated diseases for CXR classification. In contrast, we propose using datasets combination strategy and a dynamic adapter, which captures class relationships to deal with multi-label classification problems flexibly and effectively.

A.1. State of the Arts on NIH ChestXray

ChestXray dataset [57] contains 112,120 frontal-view chest x-rays of 30,805 unique patients, in which each radiography is labeled with one or multiple classes of 14 common thorax diseases. We resized the original images from size 1024×1024 into 224×224 for training a deep neural network. To ensure a fair comparison with previous methods, we adopt the official data split of ChestXray, which splits the entire dataset into 70% for training, 10% for validation, and 20% for testing. Figure 4 provides the statistics of classes and Table 9 summarizes the current state-of-the-art results on the ChestXray dataset\(^4\).

A.2. State of the Arts on Stanford CheXpert

CheXpert dataset [26] is a large dataset with 224,316 chest radiographs of 65,240 patients. 14 observations in radiology reports exist in the dataset and five frequently seen chest x-rays of 30,805 unique patients, in which each radiography is labeled with one or multiple classes of 14 common thorax diseases. We resized the images into 224×224 and the test is done on the validation set. Figure 4 provides the statistics of classes and Table 10 summarizes the current state-of-the-art results on the CheXpert dataset\(^5\).

A.3. Implementation Details

We choose DenseNet121 [25] as our backbone. All experiments run 64 epochs and utilize Adam optimizer [28] with an initial learning rate of 2e-4. We reduce the learning rate by a factor of 2.0 on the plateau with 5 steps of patience. Early stopping patience is set to be 10 epochs. The pseudo-label threshold \(\tau\) and the sharpen temperature \(t\) are 0.5 and 4.0, respectively. Ablation study of the sharpen temperature \(t\) is provided in Table 11. Considering the limited number of images in the valid set of CheXpert, we separate the training set to 166,739 images and 56,675 images for a new training set and a test set\(^6\). In the experiment of Table 3 and Table 4, we randomly sample 7,531 images in each dataset by setting the seed as 1.

B. MedMNIST Benchmark

B.1. DermaMNIST

The DermaMNIST is a large collection of multi-source dermatoscopic images of common pigmented skin lesions. The dataset consists of 10,015 dermatoscopic images categorized as seven different diseases, formalized as a multi-class classification task. The images are split into training, validation and test set with a ratio of 7:1:2. The source images of \(3 \times 600 \times 450\) are resized into \(3 \times 28 \times 28\).

B.2. TissueMNIST

The dataset contains 236,386 human kidney cortex cells, segmented from three reference tissue specimens and organized into eight categories. The source dataset is split with a ratio of 7:1:2 into training, validation and test set. Each gray-scale image is \(32 \times 32 \times 7\) pixels, where 7 denotes 7 slices. Maximum values are taken across the slices and resize them into \(28 \times 28\) gray-scale images.

B.3. OrganAMNIST

The OrganAMNIST is based on 3D computed tomography (CT) images from Liver Tumor Segmentation Benchmark (LiTS). MedMNIST [65] uses bounding-box annotations of 11 body organs from another study [62] to obtain the organ labels. Hounsfield-Unit (HU) of the 3D images are transformed into gray-scale with an abdominal window. 2D images are cropped from the center slices of the 3D bounding boxes in axial/coronal/sagittal views (planes). The images are resized into \(1 \times 28 \times 28\) to perform multi-class classification of 11 body organs. 115 and 16 CT scans from the source training set are used as training and validation set, respectively. The 70 CT scans from the source test set are treated as the test set.

\(^4\)Our baseline refers to PhoenixMark0/ChestX-ray14-Classification

\(^5\)Our baseline refers to LalehSeyyed/CheXclusion

\(^6\)The proportion is the same as ChestXray official split
Figure 5. A model can better classify “positive” images if it identifies classes in the “negative”. For binary classification, the model is only provided with a single labeled disease, while for the multi-class scenario, the model can access annotations of all diseases. Note that the number of images is the same in the two cases. The results in DermaMNIST and TissueMNIST demonstrate that the model in the binary setting has poor performance. It is due to the lack of fine-grained labels in negative examples, which potentially causes confusion between similar diseases.

Figure 6. Our dynamic adapter is more robust to novel class attacks. The novel class occurs in the test set only, not in the training or validation set. The neural network tends to be over-confident in the training set, thus suffering from the misidentification of the novel class in the test. DermaMNIST and TissueMNIST demonstrate that training with our dynamic adapter can mitigate this problem and enable progressively incorporating more novel classes.

B.4. Results on MedMNIST

To thoroughly validate our hypothesis that a model can better segment “foreground” pixels if it recognizes classes in the “background”, we have conducted extensive experiments on DermaMNIST and TissueMNIST (shown in Figure 5). Additional studies on the novel class attack have been presented in Figure 6, which further consolidates that our dynamic adapter is more robust to novel class attack.

C. Natural Imaging Benchmark

C.1. Cityscapes

The Cityscapes dataset [15] is tasked for urban scene understanding. There are totally 30 classes and only 19 classes are used for parsing evaluation. The dataset contains 5K high-quality pixel-level finely annotated images and 20K coarsely annotated images. The finely annotated 5K images are divided into 2,975/500/1,525 images for training, validation, and testing. We used it in Table 1 to demonstrate the initiative of “Data-Assemble”. The reason “bus” is chosen as our interest class is that it’s the object highly related to other classes. Concretely, we calculate the total number of co-occurrences with other objects for each label and normalize the times by the number of the occurrence of the label itself. The calculation result is shown in Figure 7.

C.2. MNIST

MNIST [33] is comprised of 10-class 28×28 pixel grayscale images of handwritten digits. It contains 60,000 images for training and 10,000 images for testing.

C.3. MNIST-novel

To demonstrate that our dynamic adapter is more robust to novel class attacks, we construct a new dataset named
Figure 7. The co-occurrence rate of each category in Cityscapes. We first count the inter-class co-occurrence, resulting in a 2D matrix where rows and columns are the numbers of classes. Each cell in the matrix is the frequency of the two classes occurring in the same image based on the ground truth. We then compute the total co-occurrence \((N_c)\) and occurrence \((D_c)\) for each class \(c\) in the Cityscapes dataset. This figure plots the normalized co-occurrence frequency \((N_c/D_c)\). “Bus” is chosen as the class of interest in Table 1 since it has the highest co-occurrence with other classes; therefore, the “bus” class is expected to benefit more if the model extensively learns from the labels of other classes (“non-bus”).

“MNIST-novel” (exampled in Figure 10) with handwritten digits in MNIST [33]. For the experiments in Table 7, we took out the images of number “0” from the training set of MNIST, while the test set remains the same. As a result, the number “0” is a novel class that the model would not see in the training phase. Note that we only trained the model using 1% of the entire training set for Table 7 to avoid the performance being too high. The AUC score is used for evaluating the binary classification: positive if the image contains the novel class, otherwise, negative. Generally speaking, any number could be considered as a novel class and, therefore, one can construct numerous different versions of MNIST-novel for a thorough examination.

C.4. MNIST-multi

To demonstrate that our dynamic adapter benefits multi-label classification, we construct a new dataset named “MNIST-multi” with handwritten digits in MNIST [33]. MNIST-multi is a more challenging dataset for multi-label classification because the images in the training set only contain one digit, while the images in the test set contain two digits. Concretely, we resize the \(28 \times 28\) images in MNIST to the size of \(28 \times 14\), and concatenate two images to form one image with \(28 \times 28\) resolution. Figure 11 illustrates examples in “MNIST-multi”.

Figure 8. The model can segment “bus” from the image more precisely if it is exposed to all 19 classes in Cityscapes instead of that exposed to only five partial classes. The reason is that the model trained with partial classes cannot identify the boundary well. In this example, without learning from “car” in Cityscapes-five, the model predicts a course boundary between the bus and the car.

Figure 9. Learning from classes of “negative examples” can better delimit the decision boundary of the class of interest. In this example, we discover that fine-grained labels for the class of “non-three” positively affect the classification of “three”. Now, imagine “three” as a rare disease whose labels are difficult to collect; this paper suggests assembling existing labels of related, common diseases to assist with diagnosing the rare disease (“three”).
Table 12. The results on MNIST reveal that (1) learning from a mixture of 3,000 partial labels performs on par with that from 4,200 full labels, therefore eliminating the need of additional 40% annotation costs; (2) learning from alternative classes boosts the performance of classifying the classes of interest, e.g., learning from Num 1 and Num 2 results in a noticeable improvement of Num 6 (95.2%→96.9%) classification; similarly, learning from Num 6 and Num 7 results in an improvement of Num 1 (95.6%→98.0%) classification. In this experimental setting, $D_0 = \{1, 2, 3, 4, 5\}$ and $D_1 = \{3, 4, 5, 6, 7\}$. Note that $D_0$ and $D_1$ are taken from the original MNIST dataset; both of them include 300 images. The gaps in color are statistically significant ($p < 0.05$).

Figure 10. To demonstrate that our dynamic adapter is more robust to novel class attacks, we construct MNIST-novel, which derives from MNIST, to evaluate the property of identifying out-of-distribution data in the test set. The novel class occurs in the test set only, not in the training or validation set. In this example, the number “0” is the novel class, which only occurs in the test set. Similar methods are utilized for experiments in DermaMNIST and TissueMNIST. Our experiments in Table 7 and Figure 6 demonstrate that our dynamic adapter is more robust to novel class attacks than conventional training, not only on MNIST-novel, but also on the tasks of Dermatoscope and Kidney Cortex Microscope image analysis.

Figure 11. We illustrate the effectiveness of dynamic adapter for multi-label classification by utilizing a novel dataset, named MNIST-multi. The training set is similar to the original MNIST dataset, which only presents one digit in each image; on the other hand, the test set presents two-digit in each image. MNIST-multi is developed to examine the robustness of the trained model when encountering multi-label scenarios in the test phase. Our experiments in Table 8 demonstrate that our dynamic adapter significantly improves multi-label classification not only on MNIST-multi, but also on two multi-label thorax X-ray datasets, in comparison with conventional multi-label training (appending “sigmoid” in the last layer).

C.5. CIFAR10

CIFAR10 [30] is a benchmark dataset for low resolution image classification. CIFAR10 has 50k training images and
Figure 12. Visualization results on NIH ChestXray. By learning from alternative classes of “negative examples”, the performance gain is positively correlated to inter-class similarity. It suggests that rather than narrowly chasing extensive labels for the class of interest (“nodule” in this example), assembling existing labels of related diseases is a more effective and efficient choice—a unique annotation scheme for computer-aided diagnosis of rare diseases and emerging pandemics, wherein “positive examples” are hard to collect, yet “negative examples” are relatively easier to assemble. The tabular results are presented in Table 5.

| Similarity* | AUC of “Card” (%) |
|-------------|-------------------|
| Cardiomegaly | 85.9              |
| + Pneumonia  | 86.3 (↑ 0.4)      |
| + Effusion   | 85.2 (↓ 0.7)      |
| + Edema      | 85.5 (↓ 0.4)      |
| + Consolidation | 86.3 (↑ 0.4)   |
| + Atelectasis| 86.9 (↑ 1.0)      |
| + Pneumothorax| 87.0 (↑ 1.3)     |
| + 11 Diseases† | 87.5 (↑ 1.6)     |

*The cosine distance between features of Cardiomegaly and the added disease.
†11 diseases from CheXpert.

Table 13. By learning from alternative classes in the “negative examples”, the performance gain is positively correlated to inter-class similarity ($r = 0.768; p = 7.43e-2$). It also suggests that rather than narrowly pursuing extensive labels for the class of interest (i.e. “cardiomegaly” in this example), assembling existing labels of related diseases is a more effective and efficient choice—a unique annotation scheme for computer-aided diagnosis of rare diseases and emerging pandemics, wherein “positive examples” are hard to collect, yet “negative examples” are relatively easier to assemble.

10k test images from 10 classes. Each image is in the resolution of 32×32.

C.6. CIFAR10-novel

To demonstrate that our dynamic adapter is more robust to the novel class attack, we construct a new dataset named “CIFAR10-novel” with natural images in CIFAR10 [30]. Concretely, in Table 7, we took out the images of “cat” from the training set of CIFAR10, while the test set remains the same. As a result, the class “cat” is a novel class that the model would not see in the training phase. The AUC score is used for evaluating the binary classification: positive if the image contains the novel class, otherwise, negative. Generally speaking, any class could be considered as a novel class and, therefore, one can construct numerous different versions of CIFAR10-novel for a more thorough examination.

D. Additional Results

D.1. Learning from a Mixture of Partial Labels Performs on Par with That from Full Labels

To further explore the effectiveness of our framework on partially labeled datasets, we have also conducted experiments on 600 images from MNIST with seven digits, focusing on the effectiveness of pseudo-labeling, consistency constraints, and our dynamic adapter. In Table 12, compared with the model trained only on $D_1$ (with digits 3,4,5,6,7), including more data from $D_0$ (with digits 1,2,3,4,5) in the training set leads to 3.9% improvement (90.6% → 94.5%) of mAUC. Noticeably, the number of la-
labels for digit 6 and digit 7 stays the same before and after assembling the two datasets, yet by learning from more data in \( D_0 \), they remarkably achieve an mAUC improvement of 0.9% and 1.1%, respectively. Moreover, pseudo-labeling and consistency constraints enable the model to leverage those unlabeled data (e.g. digit 1 and digit 2 are not annotated in \( D_1 \)), and the model shows an mAUC improvement of 1.0 points (94.4%→95.4%). Furthermore, our dynamic adapter produces an additional 0.5% improvement, leading to the final mAUC of 95.9%—on par with the performance (96.1%) of the upper bound model trained with full labels. As a result, an additional label cost of 40% is saved (i.e. 3,000 labels vs. 4,200 labels). These results indicate that our framework, aggregating partially labeled data from other datasets, can achieve similar results to full label training.

D.2. Annotating “Negative Examples” Facilitates Long Tail and Rare Disease Classification

In this section, we further explore how different “negative examples” influence the positive ones. We found that, following our assembling strategy, the performance gain of a positive class is related to the similarity between it and the negative classes. Table 13 shows the improvements of classifying Cardiomegaly by learning other diseases. The Pearson correlation coefficient between similarity and AUC improvement is 0.768, which indicates a relatively strong positive correlation. This means that learning with similar classes is more beneficial than learning with dissimilar classes for classifying the class of interest (i.e. Cardiomegaly in this setting). However, it is hard to obtain enough ground truth for training since rare diseases have limited positive examples. By incorporating similar diseases from other datasets, the model can better identify rare diseases, thus mitigating the long-tail problem in computer-aided diagnosis.

E. Visualization Result

To understand why learning from negative classes can facilitate the identification of the class of interest, we visualize the prediction of the model trained in Cityscapes-five and Cityscapes in Figure 8. The model can segment “bus” from the image more precisely if it is exposed to all 19 classes in Cityscapes in contrast with that exposed to five partial classes only. The reason is that the model trained with partial classes can not identify the boundary well. In this example, without learning from “car” in Cityscapes-five, the model predicts a course boundary between the bus and the car.

We also provide more visualization results of the inter-class relationships established by our dynamic adapter7.

---

7The chord diagram is drawn at app.flourish.studio.