LABEL-ASSEMBLE: LEVERAGING MULTIPLE DATASETS WITH PARTIAL LABELS

Mintong Kang[1], *Bowen Li*[2], *Zengle Zhu*[3], *Yongyi Lu*[2], *Elliot K. Fishman*[4], *Alan Yuille*[2], *Zongwei Zhou*[2,*]

[1]University of Illinois Urbana-Champaign [2]Johns Hopkins University
[3]Tongji University [4]Johns Hopkins University School of Medicine

ABSTRACT

The success of deep learning relies heavily on large labeled datasets, but we often only have access to several small datasets associated with partial labels. To address this problem, we propose a new initiative, "Label-Assemble", that aims to unleash the full potential of partial labels from an assembly of public datasets. We discovered that learning from negative examples facilitates both computer-aided disease diagnosis and detection. This discovery will be particularly crucial in novel disease diagnosis, where positive examples are hard to collect, yet negative examples are relatively easier to assemble. For example, assembling existing labels from NIH ChestX-ray14 (available since 2017) significantly improves the accuracy of COVID-19 diagnosis from 96.3% to 99.3%. In addition to diagnosis, assembling labels can also improve disease detection, e.g., the detection of pancreatic ductal adenocarcinoma (PDAC) can greatly benefit from leveraging the labels of Cysts and PanNets (two other types of pancreatic abnormalities), increasing sensitivity from 52.1% to 84.0% while maintaining a high specificity of 98.0%. Code is available here.

Index Terms— Partial label, diagnosis, detection

1. INTRODUCTION

Recent years have witnessed an increasing number of datasets becoming publicly available thanks to the collective efforts of imaging data archives [1] and international competitions [2, 3]. These datasets are collected, organized, annotated differently, and often come with partial labels. 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 public datasets are often incomparable, heterogeneous, or even conflicting [4, 5]. In this paper, we ponder the question: *Can we integrate and exploit such a great number of publicly available datasets with partial labels to achieve an improved computer-aided diagnosis and detection of specific diseases?*

To address this question, we start by probing a principal hypothesis (see §2.1): *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 propose a new initiative of "Label-Assemble" for leveraging partial labels from an assembly of data on hand. Specifically, we develop a new *class query* to encode different visual tasks, which can dynamically integrate partial labels across different datasets (detailed in §2.2). It is noteworthy that the conventional classification must have a predefined and fixed number of categories, but our class query trained with a question-answer manner can handle arbitrary, varying categories, thus becoming more suitable for multiple datasets with partial labels. Furthermore, pseudo labels and consistency constraints are introduced for the missing part of labels and for mitigating the domain gap across different datasets (see Figure 1B).

We validate the effectiveness of Label-Assemble in both computer-aided disease diagnosis and detection, supported by two clinical applications. **(I)** Assembling existing labels from ChestXray14 (available since 2017) significantly improves the accuracy of COVID-19 diagnosis from 96.3% (previous state of the art [6]) to 99.3%. The experiments show that assembling *pathologically-related* labels can improve the diagnosis accuracy of the interested disease. **(II)** Assembling partial labels can also help disease detection, e.g., the detection of pancreatic ductal adenocarcinoma (PDAC) can greatly benefit from leveraging the labels of Cysts and PanNets (two other types of pancreatic abnormalities), increasing sensitivity from 52.1% (previous state of the art [7]) to 84.0% and maintaining a high specificity of 98.0%. The experiments also verify that assembling *spatially-related* labels can help detect the interested disease more precisely.

In summary, the improved results from Label-Assemble are attributable to our simple yet powerful observation: *learning from the classes of "negative examples" can better delimit the decision boundary of the class of interest*. This observation agrees with the concept of "Near Misses" [8, 9], which proposed to construct negative examples near the decision boundary to facilitate the learning of visual recognizers. These results also suggest that rather than chasing for labels of the interested class, assembling labels of alternative classes can also lead to a substantial performance gain, especially for the minority class, e.g., rare and novel diseases. To our best knowledge, this study is among the first to systematically examine the rationale of assembling multiple datasets and fully exploit the potential of partial labels—the latest attempts [10, 11, 12] built models on the labeled part of the data only.

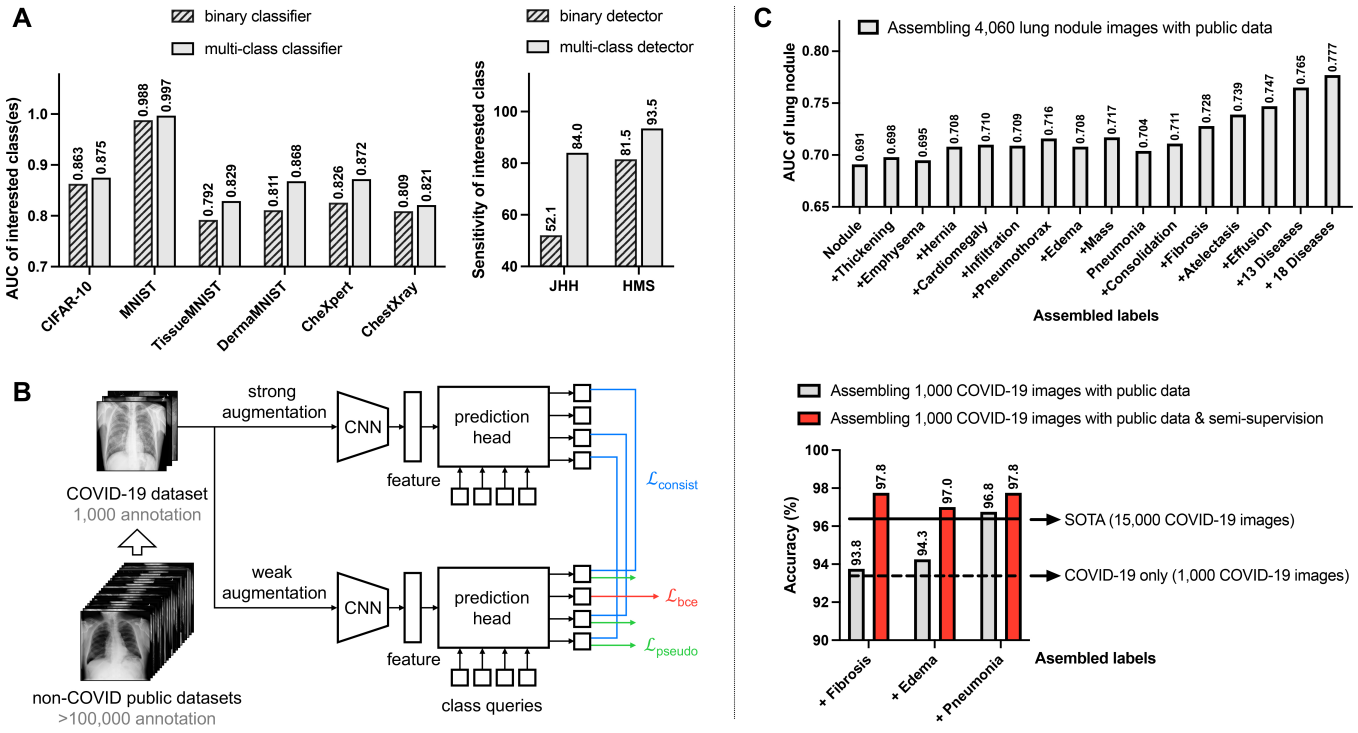

Fig. 1: **Overview.** Our proposed framework is capable of harnessing partially labeled and unlabeled data from heterogeneous sources (e.g., COVID-19 and non-COVID public datasets). **A.** With the same amount of data, learning from classes of "negative examples" benefits the learning of the interested class (see §2.1). This observation is verified by six classification tasks and two detection tasks, serving as the foundation of the Label-Assemble initiative. **B.** Labels in an assembly of public datasets are incomparable and conflicting—negative examples in the COVID-19 dataset can include the positive class in other datasets. A shared CNN extracts image features, and then a prediction head generates the predictions by inner producting features and class queries. A supervised loss (L_{bce}) is used if the label is given; two unsupervised losses (L_{pseudo} & L_{consist}) are used if the label is absent. **C.** Assembling labels of other chest diseases improves lung nodule classification. The performance gain is *positively* correlated to inter-class similarity between nodule and the assembled disease (see §3.1). The Pearson Coefficient is $r = 0.83$; $p = 4.93e\text{-}4$. Assembling 1,000 labeled COVID-19 images with public data (available since 2017), we achieve significantly higher performance than the previous state-of-the-art method, which required over 15,000 labeled COVID-19 images. Again, pathologically similar diseases (i.e., Pneumonia) lead to greater improvement in computer-aided diagnosis of COVID-19.

2. LABEL-ASSEMBLE

2.1. Motivation

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 use six multi-class datasets. For comparison, we train a multi-class classifier and a binary classifier, wherein the interested class is labeled as positive, and the rest classes are negatives. The goal is to classify the interested class. Note that the total numbers of images are the same for training the two classifiers—the only variation is that the makeup of negatives is unknown in the binary classifier, yet it is known in the multi-class classifier. In the DermaMNIST, TissueMNIST, MNIST, and CIFAR10 datasets, "melanoma", "distal convoluted tubule", "zero", and "cat" are the interested classes, respectively. In the ChestXray and CheXpert datasets, five common chest diseases, i.e., "cardiomegaly", "pneumonia", "atelectasis", "edema", "effusion", are the interested classes. Figure 1A shows that in all six datasets, the multi-class classifier consistently outperforms the binary classifier in identifying the interested classes. We attribute the deficient performance of the binary classifier to the lack of fine-grained labels in negative examples. Now, we have reached a conclusion that *learning from the classes of "negative examples" can better delimit the decision boundary of the class of interest*. This conclusion has the potential to accelerate the development circle of computer-aided diagnosis and detection of novel diseases (e.g., COVID-19 in late 2019), whose positive label is hard to collect, yet negative labels are usually available and relatively easier to assemble. Normally, one would not consider using extra labels that seem unrelated to the interested class, but we find that those existing datasets, even if they were not created for the

novel disease, are helpful for improving the performance and reducing annotation efforts (Figure 1C). This has motivated the initiative of Label-Assemble, underlining the necessity of combining multiple datasets with diverse (yet partial) labels.

2.2. Methodology

Dynamic adapter with learnable class queries. The strategy of set prediction was initially proposed for question-and-answer tasks in NLP and has recently demonstrated its power in vision tasks, such as object detection (e.g., DETR [13]), semantic segmentation (e.g., MaskFormer [14]), and medical imaging (e.g., DoDNet [10]. In light of its flexibility and effectiveness, we leverage this training strategy to address the partial label problem for the tasks of disease diagnosis and detection. Specifically, we introduce class queries, which are initialized as one-hot vectors of each class and are *learnable* during the training (differ from [10]). The class query is converted into a tensor with the same dimension as image features using a single *fc* layer. Then a prediction head, technically a linear classification layer, can generate the predictions by integrating features and class queries via inner product operations. As shown in Figure 1B, given class queries (q) and input image (x), our dynamic adapter can compute the output (a) as $a = w(q; \theta_w) * f(x)$, where $*$ is the inner product operation, w is the fully connected layer transforming class queries to classification parameters, and f is the feature extractor (CNN). Subsequently, binary cross entropy loss is used if the label (y) is provided, i.e., $L_{\text{bce}} = -(y \cdot log(a) + (1-y) \cdot log(1-a))$.

Pseudo labels & consistency constraints. To unleash the full potential of unannotated labels, we introduce a sharpening operator to generate pseudo-labels, i.e.,

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

where $\tilde{a}$ is the pseudo-label of the answer, t is the sharpen temperature, and τ is the threshold ($\tau = 0.5$ in our experiments). The prediction beyond (below) the threshold τ can be assigned to a higher (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 [15]). With the sharpening operator, the loss enables the model to operate self-training on unlabeled data, i.e., $L_{\text{pseudo}} = \|a_w - \tilde{a}_w\|_2^2$, where a_w and $\tilde{a}_2$ denote the answer of weakly augmented images and its sharpened pseudo-labels, respectively. To reduce the domain gap across the heterogeneous data sources, we further employ consistency constraints on weakly augmented (a_w) and strongly augmented (a_s) images. The consistency loss can be formulated as $L_{\text{consist}} = \|a_s - \tilde{a}_w\|_2^2$.

Overall loss function. The overall loss function consists of binary cross-entropy regularization for annotated labels as well as pseudo labels & consistency constraints for unlabeled ones, i.e., $L_{\text{total}} = L_{\text{bce}} + L_{\text{pseudo}} + L_{\text{consist}}$. Note that L_{pseudo} and L_{consist} are computed after a few warm-up epochs when the model predictions become fairly stable.

3. EXPERIMENT, RESULT, AND DISCUSSION

Dataset & metric. We evaluate our method on two computer vision datasets (i.e., MNIST, CIFAR10), seven public medical datasets (i.e., COVIDx CXR-2 [6], CheXpert [16], ChestX-ray14 [17], DermaMNIST, TissueMNIST, OrganAMNIST, RetinaMNIST [18]), and two private medical datasets (i.e., JHH and HMS) [7]. Following prior metrics for benchmarking, we evaluate the performance using Area Under the Curve (AUC) for disease diagnosis; sensitivity and specificity for disease detection. All experiments are performed by a statistical analysis based on an independent two-sample *t*-test.

Baseline & implementation. We compare our method with *three* types of baselines: 1) the multi-network strategy [19] (*one-model-one-task*), 2) multi-source learning algorithms [20, 10], and 3) SOTA algorithm [21, 22, 23, 24, 25] on NIH ChestX-ray14 and Standard CheXpert. For a fair comparison, we choose DenseNet121 as the backbone. All experiments run 64 epochs and utilize Adam optimizer 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 τ and the sharpen temperature t are 0.5 and 4.0, respectively.

3.1. Assembling partial labels improves disease *diagnosis*

As shown in §2.1, learning with additional "negative examples" improves the performance of the interested class(es). For example, assembling existing labels from ChestXray14 significantly improves the accuracy of COVID-19 diagnosis from 96.3% to 99.3% and improves the AUC of nodule diagnosis from 0.69 to 0.78. However, how much different classes of "negative examples" contribute to the performance remains unknown. We further delve into this problem and find that *the performance gain is positively related to the pathological similarity between the interested class and the added classes.* Figure 1C illustrates the improvements of classifying "Nodule" by assembling images of 13 different diseases. The Pearson correlation coefficient between the similarity[1] and the performance gain is 0.83, which indicates a significant positive correlation ($p = 4.93e$-4). This means that assembling pathologically similar classes is more beneficial than dissimilar classes for the interested class. A similar observation is obtained in the example of COVID-19 diagnosis (see gray bars in Figure 1C). In practice, it is hard to obtain enough labels for training since novel diseases have limited positive examples. By assembling similar diseases from other publicly available datasets, the model can better identify novel

[1]The similarity is quantified by the Cosine distance between the two learned class queries (see Figure 1B and §2.2).

Table 1: Assembling 75,310 *partial* labels, our method outperforms other methods developed for partial labels, and performs on par with the method learning from 105,434 *full* labels, eliminating the need for additional 40% annotation costs. The performance is measured by AUC. No significant difference ($p > 0.05$) between ours (75K partial labels) and DenseNet (105K full labels).

		CheXpert (val)						ChestX-ray14 (val)					
Method	# labels	Card†	Pneu1†	Atel†	Edema	Effusion	Average	Cons†	Pneu2†	Atel†	Edema	Effusion	Average
DenseNet [19]	37,655	0.646	0.461	0.431	0.791	0.800	0.626	0.693	0.640	0.688	0.737	0.783	0.708
Med3D [20]	75,310	0.751	0.629	0.663	0.839	0.836	0.744	0.700	0.758	0.718	0.732	0.788	0.739
DoDNet [10]	75,310	0.778	0.598	0.646	0.859	0.845	0.745	0.706	0.756	0.721	0.745	0.769	0.740
Ours	75,310	0.832	0.675	**0.702**	**0.867**	**0.886**	0.792	**0.744**	0.805	**0.813**	0.710	0.778	0.770
DenseNet [19]	105,434	**0.835**	**0.683**	0.699	0.864	0.885	**0.793**	0.719	**0.810**	0.740	**0.811**	**0.812**	**0.778**

†Card, Pneu1, Atel, Cons, Pneu2 denote Cardiomegaly, Pneumonia, Atelectasis, Consolidation Pneumothorax, respectively.

diseases, thus relieving the long-tail problem in computer-aided diagnosis. Interestingly, pseudo labels and consistency constraints can largely eliminate the requirement of similar diseases, suggesting that assembling any chest disease (regardless of the specific classes) can achieve equally high performance of COVID-19 diagnosis (see red bars in Figure 1C). These results are encouraging, but more investigation will be needed.

3.2. Assembling partial labels improves disease *detection*

JHH and HMS datasets are used to detect PDAC from CT scans. The detection of PDACs can be influenced by other types of pancreatic abnormalities, e.g., pancreatic cysts and Pancreatic Neuroendocrine Tumors (PanNETs), regarding their appearance, intensity, texture, and so on. With the same number of training cases (1,195) from JHH, we train two models: the first one only segments PDACs from the background, and the second one is trained to segment all three types of tumors from the background. These two models are evaluated on the JHH test set and HMS on the performance of detecting PDACs. As shown in Figure 1A, the performance of PDAC detection in JHH test set increases from 52.1% to 84.0% by exploiting labels of Cysts and PanNETs, while maintaining a high specificity of 98.0%; the performance of PDAC detection in HMS test set increases from 81.5% to 93.5%, while maintaining a high specificity of 90.2%.

3.3. Learning from a mixture of partial labels performs on par with that from full labels

We also compare with the methods [20, 10] developed for partial labels. Med3D [20] adopts the multi-network strategy (one-model-one-task) and DoDNet [10] learns multiple tasks in one network with shared feature extractor. Our method differs from them in *two* perspectives: (1) our adapter with updated encodings enables the model to capture the relations of classes and benefits multi-label learning, and (2) we use pseudo-labeling and consistency loss to exploit unannotated data. To adapt their method to our setting, we utilize 15,062 images from ChestX-ray14 and CheXpert with seven diseases labeled and three out of the seven diseases are shared between the two datasets. The results in Table 1 indicate that **(I)** our method with the adapter and semi-supervised learning framework achieves a better performance of multi-task learning, and **(II)** our method enables learning from partial labels to perform on par with that from full labels while eliminating an additional annotation cost of 40% (75,310 partial labels vs. 105,434 full labels). The obtained results indicate that *it is not necessary to complete the missing labels in an assembly of multiple partially labeled datasets.*

3.4. Exceeding Prior Arts in NIH ChestX-ray14

Table 2: Label-Assemble achieves the best mean performance over all 14 thorax diseases on ChestXray-14 (*official* split).

	Ref. & Year	Architecture	mAUC
Ma et al. [23]	MICCAI 2019	DenseNet (×2)	0.817
Hermoza et al. [22]	MICCAI 2020	DenseNet121	0.821
Kim et al. [21]	CVPR 2021	DenseNet121	0.822
Taslimi et al. [24]	arXiv 2022	SwinT	0.810
Xiao et al. [25]	WACV 2022	ViT-S	0.823
Ours		DenseNet121	**0.832**

Table 2 shows that assembling partial labels from publicly available datasets sets a new state of the art on ChestX-ray14 (mAUC = 0.832), yielding the best performance on 13 out of 14 diseases. Similarly, Label-Assemble is also effective on the CheXpert dataset with a 1.8% improvement over the baseline.

4. CONCLUSION

We propose a new initiative, Label-Assemble, to explore the full potential of an assembly of publicly available datasets with partial labels. The rationale of the initiative is validated on a total of six medical datasets, showing that assembling pathologically-related and spatially-related labels are preferred for disease diagnosis and detection, respectively. This is particularly valuable for novel disease diagnosis, underlining the role of an assembly of existing labels of related diseases, rather than narrowly pursuing expensive labels for the interested class. This work represents the foremost 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.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research. We thank M. R. Hosseinzadeh Taher and F. Haghighi for surveying top solutions on the ChestXray benchmark (Table 2); thank R. Feng for reproducing baseline methods on NIH ChestX-ray14. We also thank J. Liang, Z. Zhu, L. Chen, J. Chen, Y. Bai, G. Li, and X. Li for the constructive suggestions; O. Welsh for improving the writing of this paper.

5. REFERENCES

[1] Y. Yang, X. Mei, P. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, K. Link, T. Yang, C. Cao, et al., "Radimagenet: A large-scale radiologic dataset for enhancing deep learning transfer learning research," 2021.

[2] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, et al., "The medical segmentation decathlon," *arXiv preprint arXiv:2106.05735*, 2021.

[3] U. Baid, S. Ghodasara, M. Bilello, S. Mohan, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al., "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[4] Y. Zhou, Z. Li, S. Bai, C. Wang, X. Chen, M. Han, E. Fishman, and A. L. Yuille, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10672–10681.

[5] K. Yan, J. Cai, Y. Zheng, A. P. Harrison, D. Jin, Y.-b. Tang, Y.-X. Tang, L. Huang, J. Xiao, and L. Lu, "Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct," *IEEE Transactions on Medical Imaging*, 2020.

[6] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[7] Y. Xia, Q. Yu, L. Chu, S. Kawamoto, S. Park, F. Liu, J. Chen, Z. Zhu, B. Li, Z. Zhou, et al., "The felix project: Deep networks to detect pancreatic neoplasms," *medRxiv*, 2022.

[8] P. H. Winston, "The psychology of computer vision.," *Pattern Recognit.*, vol. 8, no. 3, pp. 193, 1976.

[9] N. Gurevich, S. Markovitch, and E. Rivlin, "Active learning with near misses," in *AAAI*, 2006, pp. 362–367.

[10] J. Zhang, Y. Xie, Y. Xia, and C. 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*, 2021, pp. 1195–1204.

[11] G. Shi, L. Xiao, Y. Chen, and S. K. Zhou, "Marginal loss and exclusion loss for partially supervised multi-organ segmentation," *Medical Image Analysis*, vol. 70, pp. 101979, 2021.

[12] X. Fang and P. Yan, "Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3619–3629, 2020.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[14] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[15] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[16] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 590–597.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[18] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, Apr 2021.

[19] Y. Lu, S. Pirk, J. Dlabal, A. Brohan, A. Pasad, Z. Chen, V. Casser, A. Angelova, and A. Gordon, "Taskology: Utilizing task relations at scale," 2021.

[20] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[21] E. Kim, S. Kim, M. Seo, and S. 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)*, June 2021, pp. 15719–15728.

[22] R. Hermoza, G. Maicas, J. C. Nascimento, and G. Carneiro, "Region proposals for saliency map refinement for weakly-supervised disease localisation and classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 539–549.

[23] C. Ma, H. Wang, and S. C. H. Hoi, "Multi-label thoracic disease image classification with cross-attention networks," 2020.

[24] S. Taslimi, S. Taslimi, N. Fathi, M. Salehi, and M. H. Rohban, "Swinchex: Multi-label classification on chest x-ray images with transformers," *arXiv preprint arXiv:2206.04246*, 2022.

[25] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, "Delving into masked autoencoders for multi-label thorax disease classification," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.