IEEE TRANSACTIONS ON
MEDICAL IMAGING

UniChest: Conquer-and-Divide Pre-Training for Multi-Source Chest X-Ray Classification

Tianjie Dai, Ruipeng Zhang, Feng Hong, Jiangchao Yao, *Member, IEEE*,
Ya Zhang, and Yanfeng Wang

Abstract—**Vision-Language Pre-training (VLP) that utilizes the multi-modal information to promote the training efficiency and effectiveness, has achieved great success in vision recognition of natural domains and shown promise in medical imaging diagnosis for the Chest X-Rays (CXRs). However, current works mainly pay attention to the exploration on single dataset of CXRs, which locks the potential of this powerful paradigm on larger hybrid of multi-source CXRs datasets. We identify that although blending samples from the diverse sources offers the advantages to improve the model generalization, it is still challenging to maintain the consistent superiority for the task of each source due to the existing heterogeneity among sources. To handle this dilemma, we design a Conquer-and-Divide pre-training framework, termed as UniChest, aiming to make full use of the collaboration benefit of multiple sources of CXRs while reducing the negative influence of the source heterogeneity. Specially, the *"Conquer"* stage in UniChest encourages the model to sufficiently capture multi-source common patterns, and the *"Divide"* stage helps squeeze personalized patterns into different small experts (query networks). We conduct thorough experiments on many benchmarks, *e.g.*, ChestX-ray14, CheXpert, Vindr-CXR, Shenzhen, Open-I and SIIM-ACR Pneumothorax, verifying the effectiveness of UniChest over a range of baselines, and release our codes and pre-training models at https://github.com/Elfenreigen/UniChest.**

Index Terms—**Chest x-rays, medical imaging diagnosis, conquer and divide, vision-language pre-training.**

I. INTRODUCTION

CHEST X-Ray (CXR) in screening chest diseases is essential to detect and control their fatal infectious impact on human lives in broad countries [1]. To reduce the labor

costs, deep learning techniques have become prevalent for machine-assisted CXR diagnosis, driving medical imaging recognition into the new era [2]. Specially, with the rapid development of pre-training models, extensive studies have been conducted and shown promise in a wide range of tasks and domains [3], drawing increasing attention in medical community.

Recently, Vision-Language Pre-training techniques has significantly improved the performance of machine-aided CXR disease diagnosis [4], [5], [6], [7]. Some studies even have shown the potential of surpassing the experienced radiologists in diagnosing some chest diseases [7]. Besides, in combination with medical domain-specific knowledge, these pretrained models exhibit more reasonable explanations in the lesion grounding [6], [7]. Nevertheless, it is worth noting that these works for CXR VLP only consider pre-training on a single-source dataset *e.g.,* MIMIC-CXR of about 300K samples [8]. Recalling the practice of GPT [9] or CLIP [10] that utilizes billions of multi-source samples, single-source VLP inherently induces drawbacks in the disease coverage and representativeness [8], especially under real-world medical applications.

Motivated by the above limitation, we explore building a more powerful pre-training framework by leveraging multi-source CXR data. Generally, the label space union of multiple sources can help expand the coverage of the disease categories, particularly for rare diseases. Besides, samples from different sources with diverse radiation equipment, collection standards and population distributions, may complement each other [11], which helps enhance the generalization ability of pretrained models. Unfortunately, we argue that it is still very challenging to effectively utilize multi-source CXRs, as the source heterogeneity (shown in Fig. 1) also exacerbates the complexity of the CXR disease data, which could impair the holistic improvement for all source tasks during pre-training.

To address this dilemma, we propose a Conquer-and-Divide pre-training framework, termed as UniChest, which maintains the merits of multi-source data collaboration and simultaneously weakens the negative effect of the source heterogeneity by a proper training isolation. Specifically, at the "Conquer" stage of UniChest, we promote the capture of multi-source common patterns at first in parts of the model, which focuses on enhancing the feature extraction abilities. Then, at the "Divide" stage of UniChest, we introduce a mixture of experts warmed up from previous stage, together
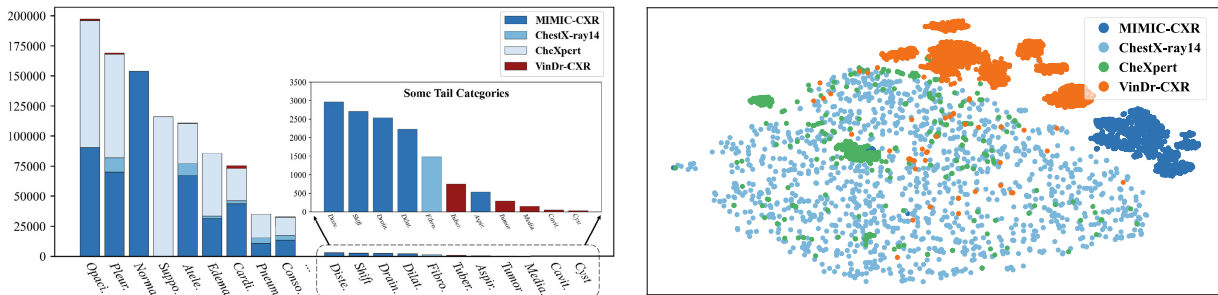
Fig. 1. Left: The class distribution of the multi-source dataset composed of MIMIC-CXR, ChestX-ray14, CheXpert and VinDr-CXR training sets. The initial five characters of each label serve as the abbreviated x-axis annotation. Right: T-SNE visualization *w.r.t.* the visual representations of medical images randomly selected from MIMIC-CXR, ChestX-ray14, CheXpert and VinDr-CXR, which characterizes the source heterogeneity.

with the guidance of source contrastive learning, to squeeze the source-specific patterns into different experts, which reduces the interference among sources during pre-training. With this framework, we can achieve better efficiency and effectiveness for pre-training on the large-scale multi-source CXR data. In a nutshell, the contribution of this work can be summarized as follows:

- We explore pre-training on large-scale multi-source CXR data and propose a Conquer-and-Divide pre-training framework to overcome the dilemma induced by the source heterogeneity in scaling up the data.
- We design a mixture of deep query networks together with a source contrastive learning loss to squeeze source-specific patterns into separate components, promoting a harmonious multi-source collaboration for pre-training.
- We conduct thorough experiments to show the promise of UniChest on multiple datasets, *i.e.*, ChestX-ray14 [12], CheXpert [13], Vindr-CXR [14], Shenzhen [15], Open-I [16] and SIIM-ACR Pneumothorax [17], achieving new state-of-the-art performance for diverse CXR diagnosis.

II. RELATED WORK

A. Deep Learning for Chest X-Ray Disease Diagnosis

Considering the labor and repeatability of human experts in CXR diagnosis, it is possible to find the computer-aided solutions powered by deep learning [18], [19], [20], [21]. Thereby, extensive CNN-based methods for CXRs have been explored in recent years [22]. For instance, Khoiriyah [23] built a network comprising of three convolutional layers and three connected layers, showing remarkable performance in automatic pneumonia detection. The enhancement by transfer learning further reduced the training cost and improved the generalization performance [24], [25]. Specifically, several studies utilized models pre-trained on the natural domain *e.g.,* ImageNet [26] as initialization and finetuned the last layer [27]. In addition, post-hoc techniques can be also very beneficial to enhance the stability and accuracy of the CXR diagnosis. For example, some explorations [28], [29] focus on the model ensemble, *i.e.*, directly summarizing the multiple outputs of a series of models as the final prediction, achieving the remarkable performance.

B. Vision-Language Pre-Training in Medical Domain

Recently, Vision-Language Pre-training (VLP) models have achieved impressive success in natural domain [30], [31], [32], which then drives many extensions in the medical area and improves the ability of machine-aided medical applications. ConVIRT [33] made the first attempt to integrate VLP into medical models, which follows the two-stream paradigm with the bidirectional contrastive learning. GLoRIA [34] explored the fine-grained information contained in the image and report, proposing a framework for learning both global and regional representations of two modalities. MedCLIP [35] proposed one decoupled multimodal contrastive learning framework based on CLIP to scale the usable training data from two distinct sources. BioVIL [4] focused on the representation of radiological reports and proposed a radiology-specific text encoder along with the classical VLP paradigm. CheXzero [5] retrained a pre-trained CLIP model [10] on the CXR data and showed considerable improvement. MedKLIP [6] extracted entities from reports and converted them to the medical-specific knowledge descriptions, which enhanced the model reasoning ability. KAD [7] built up a medical knowledge graph to fine-tune text encoder and performed the image-text contrastive learning with paired chest X-ray pairs, showing state-of-the-art capability on common benchmarks. However, all these works mentioned above are either pre-trained on a single dataset from the identical source or on dual sources like MedCLIP [35], which overlook the non-negligible heterogeneity problem in their corresponding application on the multi-source CXR data.

III. METHODOLOGY

In this section, we first present the problem formulation and the motivation of our study. Then, we will introduce the basic model design that consists of modality-specific backbones and the MoE-QN module. Finally, we provide detailed descriptions and analysis of the Conquer-and-Divide pre-training stages.

A. Problem Formulation

Assuming that we have a training set of N samples collected from the multiple sources, $\mathcal{D} = \{(x_i, Y_i, l_i, t_i)\}_{i=1}^{N}$, where x_i denotes the CXR image, Y_i is the label set indicating what diseases are found for x_i, l_i means the source identity, and t_i denotes the report of x_i. Note that, for those samples that
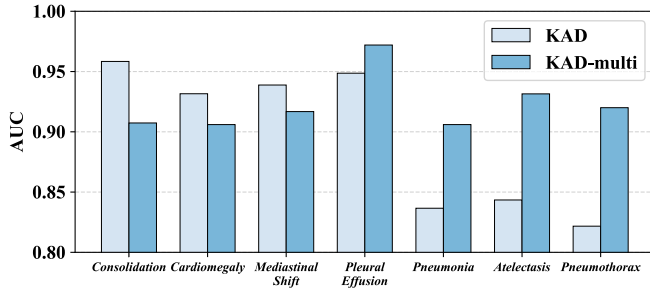
Fig. 2. Inconsistent improvement is achieved when comparing KAD and KAD-Multi in terms of the AUC metric, when scaling up the training data by multiple sources and evaluating on the VinDr-CXR test set.

have no reports, we convert Y_i to t_i correspondingly. Our goal is to train a vision-language pre-training model on the given multi-source data $\mathcal{D}$, which can accurately diagnose the chest diseases. Specifically, in the inference phase, the model can estimate the diseases in the given set for any CXR image. Here, we would like to clarify that our method adheres to the vision-language pre-training paradigm instead of transition supervised learning paradigm, since it inherently extracts representation vectors from images and text data for contrastive pre-training and can perform to open-set and zero-shot evaluations. Conventional supervised learning actually cannot perform such evaluations due to the rigid dimension issue about the prediction as well as the generalization dilemma towards the semantics of new classes.

B. Motivation

Generally, scaling up CXR samples for pre-training should be useful to improve generalization as discussed in the previous sections. However, we should point out that the resulting multi-source samples, on one hand, enjoy the better diversity *w.r.t.* training samples and label space, yet on the other hand, suffer from the non-negligible source heterogeneity issue as shown in Fig. 1. Physically, the locales and time frames of source samples can be quite diverse, as ChestX-ray14 [12] contains the samples captured from 1992 to 2015 in the U.S. while VinDr-CXR [14] consists of CXRs in Vietnam from 2018 to 2020. Even though CXRs might appear indistinguishable to the human eye, the diagnostic models can actually respond very differently in the face of some imperceptive factors like the dosage of X-ray used and the quality of the imaging instrument [36]. From the t-SNE visualization in Fig. 1, we can find that the distributions of various sources exhibit the distinct disparities. To further under the influence of the source heterogeneity, we implement the straightforward pre-training on the hybrid multi-source data and compare its performance with the pre-training on single source. Specifically, we adopt the current state-of-the-art (SOTA) *KAD* [7] to conduct the pre-retraining on MIMIC-CXR dataset, and compare with the pre-training on the multiple sources (termed as *KAD-multi*). As shown in Fig. 2, KAD-multi does not achieve the consistent improvement, and on some certain diseases, KAD-multi even significantly lags behind the vanilla KAD.

With this observation, we re-think the early VLP paradigm that is naively applied in scaling up CXR data and overlooks the source heterogeneity issue. To handle this dilemma, we actually should allow the pre-training to capture the multi-source common patterns and simultaneously can maintain the source-specific patterns, which motivates us to incorporate the philosophy of "Conquer" and "Divide" design for pre-training.

C. Conquer-and-Divide Pre-Training

In this part, we describe the proposed Conquer-and-Divide pre-training framework, which includes the model architecture, the loss design and the training schedules, detailed as follows.

1) The Model Architecture: We follow the prevalent vision-language pre-training paradigm with the proper tailored design to train a diagnosis model, which consists of two modality-specific encoders and one modality-interaction module. Note that, this is different from the classical supervised framework that directly maps the input image space to the label space. The merits are tri-fold: First, we can inject more knowledge into the label space via the textual encoder, which is richer in semantics than the naive one/multi-hot label vector. The second is that we can incorporate the prior knowledge to promote the learning efficiency of vision encoder, when the textual description for medical images is available, e.g., the report information of samples from MIMIC. Finally, VLP allows us to achieve better generalization for open-set categories by means of the powered textual encoder. In the following, we concretely describe the architecture design of each component in our UniChest.

Given a sample (x_i, Y_i, l_i, t_i), we take the ResNet-50 as the visual backbone to encode x_i and adopt the output of the 4-th residual block as the image representation [6], denoting as $\Phi_{\text{image}}(\cdot)$. For the report information t_i, RadGraph [37] is used to extract key entities and filter the irrelevant words. When reports containing more than one sentence, we extract entities sentence by sentence, and use [SEP] token as the separation between different entities.[1] After the entity extraction, we use PubMedBERT [38] pre-trained on Unified Medical Language System (UMLS) data [39] as textual encoder backbone Φ_{text} for generating text representation [7]. For clarity, we summarize the modality-specific encoding process as below,

$$\boldsymbol{I}_i = \Phi_{\text{image}}(x_i) \in \mathbb{R}^{h \times w \times d}, \quad \boldsymbol{T}_i = \Phi_{\text{text}}(t_i) \in \mathbb{R}^d. \quad (1)$$

For the disease prediction, we introduce a MoE-QN Module, namely, mixture of query networks, where each query network consists of a few transformer decoder layers (4-layers in default). The MoE-QN module plays the role of overcoming the source heterogeneity issues by squeezing the source-specific patterns into different query networks, but its training should carefully follow our Conquer-and-Divide schedules, which will be discussed in the subsequent sections. For each query network, we take the fine-grained visual representation $\boldsymbol{I}_i$ as Key and Value, and utilize the textual representation of the disease set $Y_i = \{Y_{i1}, \ldots, Y_{ic}\}$ (encoded

---

[1] For CXR samples without reports, we use their corresponding labels as the input entity and also take [SEP] token to separate the multiple labels.
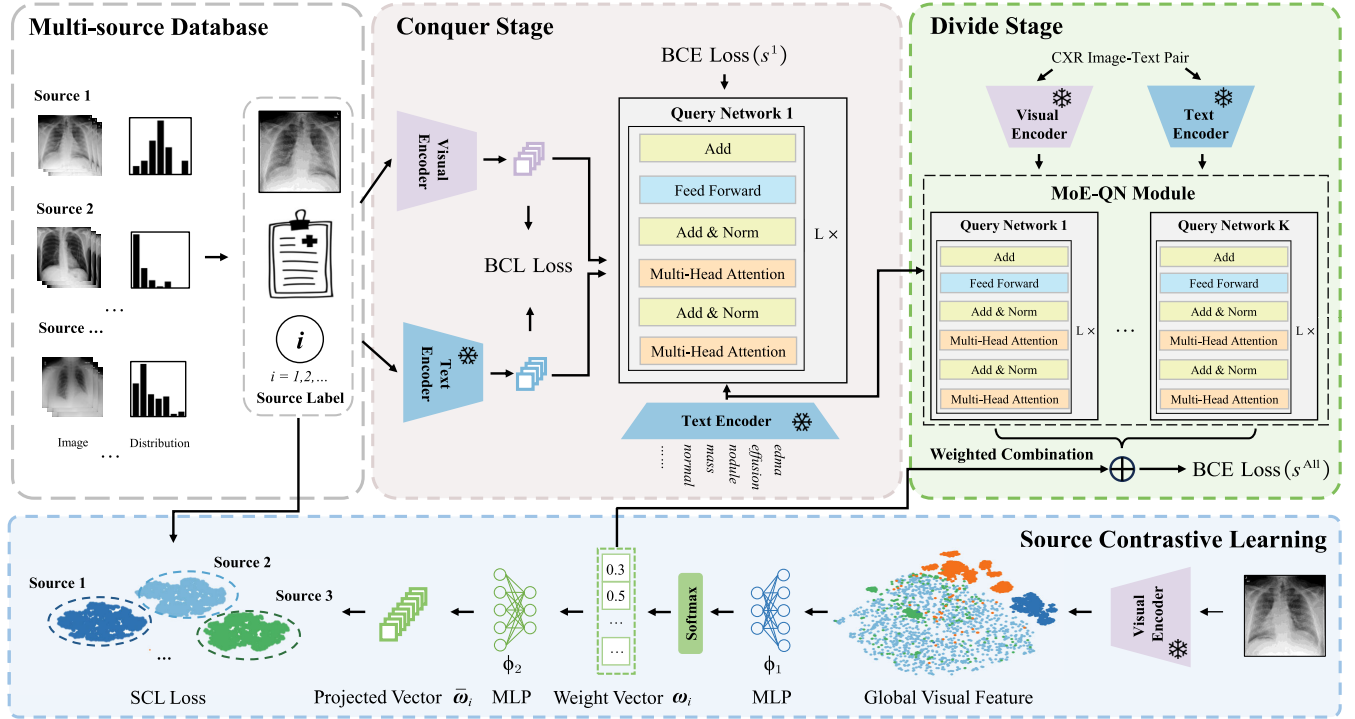
Fig. 3. The framework of UniChest, which consists of two training stages. During the "Conquer" stage, two modality encoders first project visual and textual representations into the common space with alignment, then feed them into the first transformer query networks for prediction. The multi-source common patterns are learnt as much as possible at this stage. During the "Divide" stage, we freeze the modality encoders and squeeze the source-specific patterns via the MoE-QN module with the guidance of the enhanced supervised loss and the source contrastive learning.

by $\Phi_{\text{text}}(\cdot)$ similarly as in Eq. (1)) as Query. The output from the sequential transformer decoder will be fed to one MLP layer to obtain the prediction of each Query Network,

$$s_i^k = \Phi_{\text{QN}}^k(\boldsymbol{I}_i, \Phi_{\text{text}}(Y_i)) \in \mathbb{R}^c, \ k = 1, \ldots, K, \quad (2)$$

where c is the total class number of all sources and K is the number of query networks. Then, we transform $\{s_i^k\}_{k=1}^K$ of all query networks to the total prediction by an automatic linear combination, which will be described at the "Divide" stage.

2) The Training at the "Conquer" Stage: As discussed in the earlier sections, the source heterogeneity issue can be the core bottleneck to deteriorate the performance of pre-training for the consistent improvement for each source task. Therefore, at this "Conquer" stage, we first encourage the model to capture multi-source common patterns as many as possible following the prevalent vision-language pre-training framework like KAD [7]. The training loss of this stage involves two parts, the image-text bidirectional contrastive loss $\mathcal{L}_{\text{BCL}}$ and the specially enhanced multi-label cross-entropy loss $\mathcal{L}_{\text{BCE}}$ (see Section IV-B for its implementation detail) *w.r.t.* the *first* query network prediction. Taking a batch of M samples as an example, $\mathcal{L}_{\text{BCL}}$ emphasizes the alignment between the global visual representation and the textual representation, which could be formulated as follows,

$$\mathcal{L}_{\text{BCL}} = -\log \frac{e^{\bar{\boldsymbol{I}}_i^\top \boldsymbol{T}_i/\tau}}{\sum_{m=1}^M e^{\bar{\boldsymbol{I}}_i^\top \boldsymbol{T}_m/\tau}} - \log \frac{e^{\boldsymbol{T}_i^\top \bar{\boldsymbol{I}}_i/\tau}}{\sum_{m=1}^M e^{\boldsymbol{T}_i^\top \bar{\boldsymbol{I}}_m/\tau}} \quad (3)$$

where $\bar{\boldsymbol{I}}_i$ is the average pooling of $\boldsymbol{I}_i$ along the first two dimensions so that we reduce it to the same dimension of $\boldsymbol{T}_i$, and τ is temperature with the default setting 1.0 following [40]. Regarding $\mathcal{L}_{\text{BCE}}$, we only enforce the supervision on the *first*

query network and compute the enhanced multi-label cross-entropy loss at this stage. Note that, samples in some domains may have smaller label space than the union label space of all sources. In this case, we neglect the computation between prediction and missing classes of these samples, even if it may belong to them (but are unobserved). Finally, we sum up $\mathcal{L}_{\text{BCL}}$ and $\mathcal{L}_{\text{BCE}}$ (only involves the first query network) as the overall loss for each mini-batch at the "Conquer" stage as follows

$$\mathcal{L}_{\text{Conquer}}(s^1) = \mathcal{L}_{\text{BCL}} + \mathcal{L}_{\text{BCE}}(p, Y)\Big|_{p=s^1}. \quad (4)$$

3) The Training at the "Divide" Stage: Previous stage follows the classical pre-training spirit, where we treat multiple sources equally, and expect that the model can sufficiently learn the multi-source common patterns. At the "Divide" phase, we mainly explore to mediate the negative impact induced by the source heterogeneity. Concretely, on the basis of the warming-up during the "Conquer" stage, we train all the query networks (MoE-QN Module) with *freezing* all the other parts of the model. In the MoE-QN Module, the new query networks introduced at the this stage are initialized from randomness. Here, the multi-label cross-entropy loss will be implemented on the ensemble prediction s^{All} from $\{s^k\}_{k=1}^K$, unlike using s^1 in Eq. (4). The final prediction is characterized by the following equation

$$s_i^{\text{All}} = \lambda s_i^1 + (1 - \lambda) \sum_{k=2}^K s_i^k \omega_i^k, \quad (5)$$

where λ is the hyperparameter to balance the training of the first query network and the remaining query networks, and ω_i^k is the learnable weight to summarize the contribution of the source-specific modules for the diagnosis prediction. Note

TABLE I

COMPARISON BETWEEN UNICHEST AND SINGLE-SOURCE PRE-TRAINING WITH FINE-TUNING BASELINES ON CHESTX-RAY14.
AUC SCORES OF 14 CLASSES AND THE aAUC SCORE ARE LISTED. EACH LABEL IS ABBREVIATED, WHICH
IS ALIGNED WITH LABELS IN THE LEFT PANEL OF FIG. 4

Method	Mean	Ate.	Car.	Eff.	Inf.	Mas.	Nod.	Pne.	Pne.	Con.	Ede.	Emp.	Fib.	Thi.	Her.
ConVIRT	0.808	0.771	0.867	0.825	0.703	0.818	0.761	0.722	0.857	0.747	0.854	0.901	0.809	0.771	0.909
GLoRIA	0.800	0.760	0.855	0.818	0.700	0.814	0.749	0.715	0.828	0.739	0.832	0.887	0.813	0.767	0.921
BioViL	0.800	0.765	0.871	0.824	0.697	0.819	0.752	0.710	0.845	0.742	0.842	0.871	0.821	0.759	0.888
MedKLIP	0.801	0.764	0.849	0.823	0.697	0.820	0.747	0.712	0.839	0.751	0.848	0.879	0.817	0.777	0.892
KAD	0.825	0.785	0.897	0.840	0.713	0.836	0.771	0.740	0.874	0.753	0.860	0.916	0.829	0.778	0.961
UniChest	**0.858**	**0.823**	**0.922**	**0.861**	**0.739**	**0.850**	**0.778**	**0.777**	**0.933**	**0.795**	**0.893**	**0.958**	**0.877**	**0.837**	**0.977**

TABLE II

COMPARISON WITH OTHER STATE-OF-THE-ART METHODS ON FINE-TUNING CLASSIFICATION TASK WITH DIFFERENT TRAINING DATA PORTIONS
ON SIIM-ACR PNEUMOTHORAX. FT DENOTES FINE-TUNING PRE-TRAINED MODEL, ZS DENOTES ZERO-SHOT CLASSIFICATION

Data Portion	1%				10%				100%			
Method	aAUC↑	aF1↑	aACC↑	mAP↑	aAUC↑	aF1↑	aACC↑	mAP↑	aAUC↑	aF1↑	aACC↑	mAP↑
ConVIRT (FT)	0.7134	0.4651	0.6375	0.4224	0.7826	0.5346	0.7190	0.4640	0.8989	0.6858	0.8351	0.7039
GLoRIA (FT)	0.7439	0.4948	0.7030	0.4595	0.8538	0.6091	0.7874	0.6250	0.9018	0.6898	0.8450	0.7269
BioVIL (FT)	0.6947	0.4685	0.6244	0.3816	0.7775	0.5393	0.7372	0.5183	0.8688	0.6364	0.8004	0.6624
MedKLIP (FT)	0.8527	0.5952	0.7925	0.5858	0.9071	0.6845	0.8389	0.6823	0.9194	0.7246	0.8689	0.7510
KAD (FT)	0.8541	0.5810	0.8097	0.4822	0.9023	0.6498	0.8490	0.6506	0.9203	0.6788	0.8655	0.7079
UniChest (ZS)	0.9260	0.6928	**0.8807**	0.6929	0.9260	0.6928	0.8807	0.6929	0.9260	0.6928	0.8807	0.6929
UniChest (FT)	**0.9375**	**0.7182**	0.8801	**0.7326**	**0.9418**	**0.7314**	**0.8881**	**0.7508**	**0.9437**	**0.7254**	**0.8885**	**0.7696**

that, the intuition behind Eq. (5) is to inherit the training gains at the "Conquer" stage by λ, and simultaneously squeeze the source-specific patterns to the other query networks by ω_i^k.

Source Contrastive Learning. In Eq. (5), we introduce the learnable weights to incorporate the source-specific patterns into the newly introduced query networks during the "Divide" Stage. However, without any guidance, it is challenging to achieve this goal by optimization. Here, we introduce a source contrastive learning to promote the desire. Specifically, we transform the global visual representation $\bar{I}_i$ into a $K-1$ simplex via one-layer MLP $\Phi_1(\cdot)$ and Softmax layer as the weight vector, namely, $\omega_i = \Phi_1(\bar{I}_i) \in \mathbb{R}^{K-1}$. Then, ω_i is projected into a higher dimension space by another one-layer MLP $\Phi_2(\cdot)$, denoted as $\bar{\omega}_i = \Phi_2(\omega_i)$, which is the high-dim vector to perform a source contrastive learning with the guidance of source id l in the following

$$\mathcal{L}_{\text{SCL}} = -\log \frac{\sum_{l_j=l_i}^{j \neq i} e^{\bar{\omega}_i^\top \bar{\omega}_j / \tau}}{\sum_{m \neq i}^{M} e^{\bar{\omega}_i^\top \bar{\omega}_m / \tau}}. \qquad (6)$$

In Eq. (6), the numerator $\sum_{l_j=l_i}^{j \neq i} e^{\bar{\omega}_i^\top \bar{\omega}_j / \tau}$ computes the sum of exponential similarities of the high-dim weight vectors $\bar{\omega}$ between sample i and all other samples j from the same source l, encouraging to pull the samples with the same source identify closer together. The denominator $\sum_{m \neq i}^{M} e^{\bar{\omega}_i^\top \bar{\omega}_m / \tau}$ calculates the sum of exponential similarities of the high-dim weight vectors between sample i and all other samples in the batch, providing a normalization factor to ensure stability of the loss. Generally, above contrastive learning helps us to learn the similar weight vector for the samples from the same source, which makes the source-specific patterns learned by the similar query networks. Under this mechanism, we naturally learn an automatic optimal assignment for the remaining query networks to overcome the source heterogeneity issue. Then, with Eq. (5) and Eq. (6), the overall loss at the "Divide" stage

can be formulated as

$$\mathcal{L}_{\text{Divide}} = \mathcal{L}_{\text{SCL}} + \mathcal{L}_{\text{Conquer}}(s^{\text{All}}) \qquad (7)$$

4) Difference From the Pre-Training With Fine-Tuning: The proposed Conquer-and-Divide pre-training is intrinsically different from the ordinary fine-tuning after pre-training, although both of them are a two-stage process. First, the two-stage training of UniChest is towards all multi-source data, while the fine-tuning after pre-training is narrowed down to a single-source scenario. When the target dataset's corresponding training set is seen during multi-source pre-training, this leads to the distinction in the generalization ability of two frameworks, where UniChest is significantly stronger than pre-training with fine-tuning as supported by Table I. Second, the "Divide" stage is to properly mediate the heterogeneity patterns in a broad sense, which does not mean multi-source pre-training contradicts with the conventional pre-training and fine-tuning diagram. Conversely, the downstream task can still apply the ordinary fine-tuning ways to achieve fast adaptation for unseen sources. For instance, we fine-tune the pre-trained UniChest on SIIM-ACR Pneumothorax [17] with different training data ratios following MedKLIP [6]. Overall, UniChest's zero-shot capability is already comparable to the 100% fine-tuned performance of other models and the diagnostic performance is further enhanced after further fine-tuning as shown in Table II. Finally, when it comes to the subsequent experimental comparison, we mainly focus on the zero-shot performance of pre-training, instead of the performance by the fine-tuning, which follows the prevalent pre-training spirit that tries to avoid the expensive tuning cost for downstream tasks as much as possible.

IV. EXPERIMENTS

In this section, we will introduce the datasets for pre-training and downstream tasks, evaluation metrics and

TABLE III
STATISTICS OF DATASETS FOR PRE-TRAINING AND DOWNSTREAM

Purpose	Dataset	# Samples	# Diseases	Region	Collection Institution	Time Scope
Pre-training	MIMIC-CXR [8]	348900	41	Northeast USA	MIT	2011-2016
	ChestX-ray14 [12]	98637	14	Northeast USA	NIH	1992-2015
	CheXpert [13]	223414	14	Western USA	Stanford Hospital	2002-2017
	VinDr-CXR [14]	15000	28	Vietnam	VinBrain	2018-2020
	Multi-CXR	685951	60	USA & Vietnam	MIT, *etc.*	1992-2020
Downstream	Shenzhen [15]	662	1	Southeast China	Guangdong Medical College	Sep. 2012
	Open-I [16], [41]	3547	14	USA	NLM	/
	SIIM-ACR [6], [17]	2135	1	Northeast USA	Kaggle	/
	PadChest [16]	39053	193	Spain	University of Alicante	2009-2017
	ChestX-Det10 [42]	542	/	Northeast USA	Deepwise	1992-2015

implementation details will also be detailedly described. At last, we will present the experimental results of our proposed method compared with other baselines.

A. Datasets

We combine some common CXR datasets containing **MIMIC-CXR** [8], **ChestX-ray14** [12], **CheXpert** [13] and **VinDr-CXR** [14] for pre-training, namely **Multi-CXR**, which yields a grand total of 685,951 Chest X-Ray images.

For evaluating the generalization performance of the pre-trained model, we adopt **Shenzhen** [15], **Open-I** [16], **SIIM-ACR Pneumothorax** [17] and **PadChest** [16] for zero-shot classification and **ChestX-Det10** [42] which is the subset of ChestX-ray14 for intuitive lesion grounding. More detailed statistics of pre-training and downstream datasets can be found in Table III. We can observe that significant diversity exists in the collection time, population distribution, disease category coverage and sample scales among different datasets.

B. Implementation Details

The pre-training process of UniChest is conducted on a single NVIDIA A100 GPU for 30 epochs in the first "Conquer" stage and 20 epochs in the second "Divide" stage. The starting checkpoint for the second stage is the one that performed the best on the ChestX-ray14 validation set. The number of transformer decoder layers of each query network in MoE-QN Module is set to 4. The hyperparameter λ which weights the first query network during the "Divide" stage is set to be 0.5 and the total number of query network K is 4 in default. The dimension of $\bar{\omega}$ which is the projection of the MLP layer $\Phi_2(\cdot)$ is set to be 32 during source contrastive learning. The temperature parameter τ in Eq. (6) is set to 1.0. We set the training batch size as 32 and resize input images as 512×512. We adopt the AdamW optimizer in conjunction with the cosine annealing scheduler for managing the learning rate, where the initial learning rate is 1×10^{-5}. To perform data augmentation, we additionally utilize the Fourier amplitude mixup method, which has been demonstrated to be effective for medical imaging data [43], [44]. Besides, we utilize the ASL loss [45] as $\mathcal{L}_{BCE}$ to promote balanced training. These techniques have demonstrated promising outcomes in the analysis of CXR data. For MedKLIP-multi and KAD-multi, we strictly follow the training hyper-parameters and

implementation details described in their original papers and official codes. Their pre-training dataset is also Multi-CXR that is identical to UniChest. In MedKLIP-multi, for samples without location annotations (excluding MIMIC-CXR), we mask the calculation of location contrastive loss directly.

C. Evaluation Metrics

Some common metrics for multi-label classification are adopted to evaluate the model performance, *i.e.*, area under curve (AUC), F1 score, accuracy (ACC) and average precision (AP) for each category and their average value, namely, average AUC (aAUC), average F1 (aF1), average accuracy (aACC) and mean average precision (mAP), for comprehensive comparison. Following the strategy of Med-KLIP [6], the final binary threshold of each category prediction is the value when the maximum F1 score is achieved and ACC metric also adopts this threshold.

D. Baselines

We consider a wide range of baselines for CXR pre-training, including **ConVIRT** [33], **GLoRIA** [34], **MedCLIP** [35], **BioVIL** [4], **CheXzero** [5], **MedKLIP** [6], and **KAD** [7].

ConVIRT trains two modality-specific encoders by bidirectional contrastive loss to learn visual representations. **GLoRIA** utilizes both global and fine-grained features for medical VLP. **MedCLIP** [35] trains one CLIP-based framework on two CXR datasets including MIMIC-CXR and CheXpert. **BioVIL** proposes a radiology-specific text encoder for the subsequent classical pipeline of VLP. **CheXzero** retrains one CLIP model with a corpus of the medical domain. **Med-KLIP** designs one novel entity extraction and transition module to inject domain-specific knowledge into the process of VLP. **KAD** incorporates a medical knowledge graph to further improve the capability of current VLP models in Chest X-Ray, showing SOTA performance on some common public benchmarks.

E. Results

In this part, we provide the results of in-domain evaluation on pre-training datasets and zero-shot evaluation on

Comparison of UniChest With Baselines on in-Domain Classification. Four Metrics Including aAUC, aF1, aACC and mAP Scores Are Reported. For All Datasets, the Metrics All Refer to the Macro Average on All Diseases

| Dataset | ChestX-ray14 | | | | CheXpert | | | | VinDr-CXR | | | |
Method	aAUC↑	aF1↑	aACC↑	mAP↑	aAUC↑	aF1↑	aACC↑	mAP↑	aAUC↑	aF1↑	aACC↑	mAP↑
ConVIRT	0.5804	0.1623	0.5385	0.1008	0.6640	0.3609	0.7232	0.2990	0.6803	0.1748	0.8233	0.1129
GLoRIA	0.6099	0.1803	0.5616	0.1220	0.6889	0.4094	0.7472	0.3580	0.6775	0.1953	0.7796	0.1380
MedCLIP	0.6696	0.2061	0.7001	0.1424	0.7228	0.4149	0.7719	0.3613	0.6958	0.2079	0.8163	0.1503
BioViL	0.5766	0.1561	0.5842	0.0948	0.5769	0.3080	0.6446	0.2157	0.5856	0.1507	0.7726	0.1048
CheXzero	0.6872	0.2205	0.7698	0.1597	0.7537	0.4488	0.8136	0.4069	0.7407	0.2354	0.8562	0.1713
MedKLIP	0.7233	0.2541	0.7946	0.1860	0.8389	0.5385	0.8672	0.5130	0.7770	0.2433	0.8517	0.1897
KAD	0.7933	0.3363	0.8639	0.2746	0.8165	0.5467	0.7835	0.5200	0.7599	0.2960	0.8724	0.2315
KAD-CXR14	0.8380	0.4006	0.8913	0.3401	0.7455	0.4296	0.7658	0.3901	0.7711	0.2399	0.8410	0.1740
KAD-CXP	0.7014	0.2444	0.7744	0.1767	0.8804	0.5964	0.8846	0.5606	0.7382	0.2256	0.8101	0.1619
KAD-VC	0.6372	0.1881	0.6776	0.1275	0.7228	0.4161	0.7811	0.3285	0.8621	0.3839	0.9418	0.3255
MedKLIP-multi	0.7915	0.2994	0.8691	0.2313	0.7010	0.4244	0.7395	0.3410	0.7879	0.2802	0.9167	0.2196
KAD-multi	0.8431	0.4077	0.8905	0.3457	0.8819	0.5912	0.8713	0.5626	0.8716	0.3698	0.9390	0.3202
UniChest	**0.8584**	**0.4293**	**0.8999**	**0.3797**	**0.9005**	**0.6446**	**0.8912**	**0.6328**	**0.8807**	**0.4028**	**0.9501**	**0.3520**
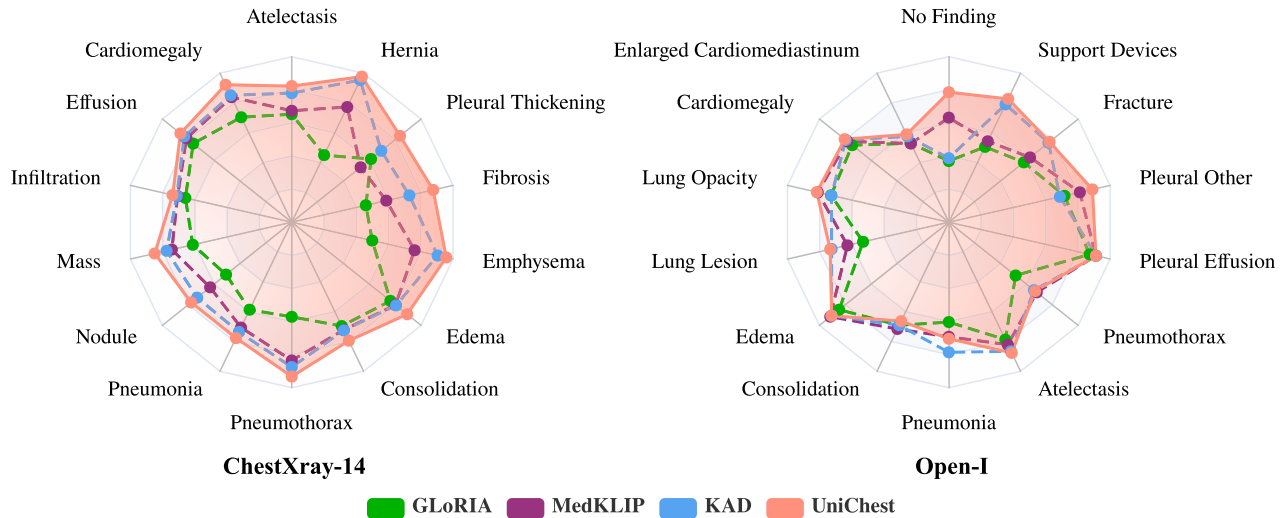
Fig. 4. Per-category performance of different methods on ChestX-ray14 (left) and Open-I (right). AUC scores of each category are displayed. 1 and 0 are adopted as the maximum and minimal values for each category in the radar chart.

downstream datasets. Additionally, we showcase the lesion grounding capability through intuitive examples.

1) In-Domain Evaluation on Pre-Training Datasets: We compare UniChest with other baselines on the test set of pre-training datasets in Table IV. UniChest demonstrates a notable improvement in various evaluation metrics compared to the best baseline models. Specifically, for ChestX-ray14, UniChest shows a significant improvement of 6.51%, 9.30%, 3.60% and 10.51% respectively as to aAUC, aACC, aF1 and mAP than the best baseline. In terms of fine-grained classification for each category, UniChest demonstrates persistent gains as shown in Fig. 4 (left). For CheXpert, UniChest outperforms 6.16%, 9.79%, 2.40% and 11.28%. In the case of VinDr-CXR, UniChest surpasses the best baselines by 12.08%, 10.68%, 7.77% and 12.05%. By introducing multi-source CXR datasets and the Conquer-and-Divide pre-training framework, the diagnosis ability of VLP model has been significantly enhanced.

To mitigate the impact of inconsistent training data, we also conduct variants of the baselines that utilize the same training data with UniChest. These variants include models trained on single-source data (*i.e.,* KAD-CXR14, KAD-CXP, and KAD-VC in Table IV) as well as models trained on multi-source data (*i.e.,* MedKLIP-multi and KAD-multi). KAD-CXR14 outperforms KAD on ChestX-ray14 with improvements of 4.47% in aAUC, 6.43% in aACC, 2.74% in aF1, and 6.55% in mAP. However, it falls short compared to UniChest, which achieves better results by 2.04%, 2.87%, 0.86%, and 3.96% respectively. KAD-CXP and KAD-VC also demonstrate similar performance. Moreover, it is worth noting that while the single dataset-based KAD models may perform reasonably well on their corresponding datasets, their generalization capability to other datasets is limited. This is evident in the case of KAD-VC's performance on ChestX-ray14, where it may not generalize effectively. For the baseline variants that utilize

TABLE V

COMPARISON OF UNICHEST WITH BASELINES ON ZERO-SHOT CLASSIFICATION. FOUR METRICS INCLUDING aAUC, aF1, aACC, AND mAP SCORES ARE REPORTED. FOR SINGLE-LABELED DATASET, WE REPORT AUC, F1, ACC, AND AP SCORES

| Dataset Method	Shenzhen				Open-I				SIIM-ACR Pneumothorax			
	AUC↑	F1↑	ACC↑	AP↑	aAUC↑	aF1↑	aACC↑	mAP↑	AUC↑	F1↑	ACC↑	AP↑
ConVIRT	0.7166	0.6908	0.6375	0.7619	0.6265	0.1855	0.7504	0.1115	0.6356	0.4322	0.4778	0.2948
GLoRIA	0.5831	0.6733	0.5060	0.6001	0.6449	0.2218	0.8134	0.1535	0.5342	0.3770	0.4047	0.2084
MedCLIP	0.6545	0.6933	0.5160	0.6321	0.6609	0.2422	0.8406	0.1593	0.6822	0.4611	0.5564	0.3159
BioViL	0.6227	0.6872	0.5544	0.6174	0.5669	0.1515	0.6966	0.0941	0.4678	0.3109	0.1836	0.1761
CheXzero	0.7801	0.7302	0.6692	0.7969	0.6887	0.2429	0.8689	0.1850	0.6879	0.4722	0.5588	0.3240
MedKLIP	0.7909	0.7741	0.6906	0.7986	0.7219	0.2897	0.9048	0.2119	0.8924	0.6833	0.8428	0.6869
KAD	0.8856	0.8474	0.7893	0.8141	0.7346	0.3248	0.8808	0.2603	0.8967	0.6480	0.8385	0.6093
KAD-CXR14	0.8764	0.8037	0.8000	0.8947	0.6986	0.2676	0.8617	0.1856	0.9162	0.6561	0.8348	0.6872
KAD-CXP	0.7204	0.6793	0.6031	0.7611	0.7701	0.3314	0.9011	0.2631	0.9132	0.6554	0.8456	0.6771
KAD-VC	0.8832	0.8448	0.8578	0.9186	0.7012	0.2112	0.8468	0.1503	0.8002	0.4995	0.7661	0.4012
MedKLIP-multi	0.8431	0.7780	0.7855	0.8817	0.6440	0.2270	0.7943	0.1469	0.8958	0.6888	0.8412	0.6853
KAD-multi	0.8794	0.8419	0.8547	0.9180	0.7736	0.3409	0.9065	0.2749	0.9016	0.6515	0.8395	0.5847
UniChest	**0.9499**	**0.8906**	**0.8913**	**0.9598**	**0.7830**	**0.3520**	**0.9104**	**0.2877**	**0.9260**	**0.6928**	**0.8807**	**0.6929**
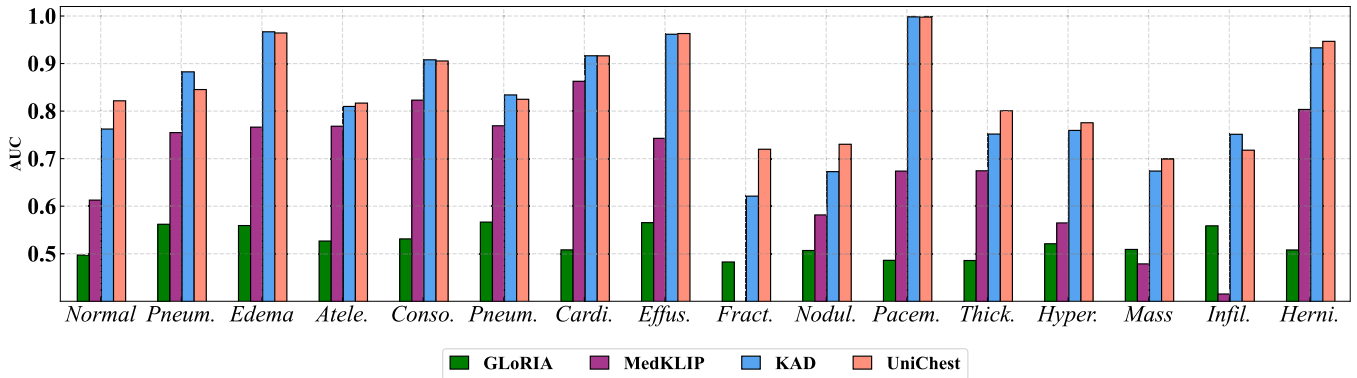
Fig. 5. Per-category performance of 16 *seen* categories during pre-training in PadChest.

multi-source data, MedKLIP-multi and KAD-multi, their overall performance exceeds that of their vanilla methods and achieves comparable or even better results than using only the corresponding single-source data. However, these two methods of direct data replenishment lag behind our proposed UniChest with a substantial margin. Generally speaking, the comparison with these variants validates the necessity and importance of developing a multi-source CXR foundational model and demonstrates the effectiveness of our framework design.

2) Zero-Shot Classification Evaluation: We conduct the zero-shot evaluation to assess the generalization ability of CXR pre-training models. The results in Table V demonstrate the superiority of UniChest across various datasets. In the case of the Shenzhen dataset, UniChest surpasses the best baseline KAD by 6.43%, 4.32%, 10.20% and 14.57% respectively as to AUC, ACC, F1 and AP. For the SIIM-ACR Pneumothorax dataset, UniChest outperforms 2.93%, 0.95%, 3.79% and 0.60% respectively compared with the best baseline. As to Open-I dataset, UniChest achieves significant improvements by 4.84%, 2.72%, 0.56% and 2.74%. As indicated by the per-category evaluation in Fig. 4 (right), UniChest achieves performance improvements across the majority of the categories. For instance, the classification ability is sharply

consolidated of *No Finding*, *Pleural Other* and *Support Devices* with a margin of over 3%. For PadChest, UniChest obtains an average AUC of 0.8403 for 16 *seen* categories during pre-training, surpassing KAD by 1.51%. Specifically, the diagnosis capability of 11 classes achieves SOTA as shown in Fig. 5, among which the improvements of *fracture*, *nodule* and *pleural thickening* are around or over 5%. For unseen pathologies in Fig. 6, UniChest also showcases superior performance in the given 16 diseases, demonstrating its value in rare disease diagnosis. Besides, it is worth noting that the multi-source data baseline variants (*e.g.*, KAD-multi) exhibit relatively modest and inconsistent improvements compared to their vanilla methods and single-source data baseline variants (*e.g.*, KAD-CXR14), which emphasizes the necessity of our method design in enhancing the model's generalization ability. In summary, the results in Table V validate the significance of the Conquer-and-Divide pre-training framework for multi-source CXR samples, which indicates its potential in assisting clinical human diagnosis and highlights its unignorable value.

3) Qualitative Grounding Visualization: In Fig. 7, we present several examples of lesion grounding on ChestX-Det10 by UniChest. To provide an intuitive visualization, we generate spectrum heatmaps on the original CXR images based on
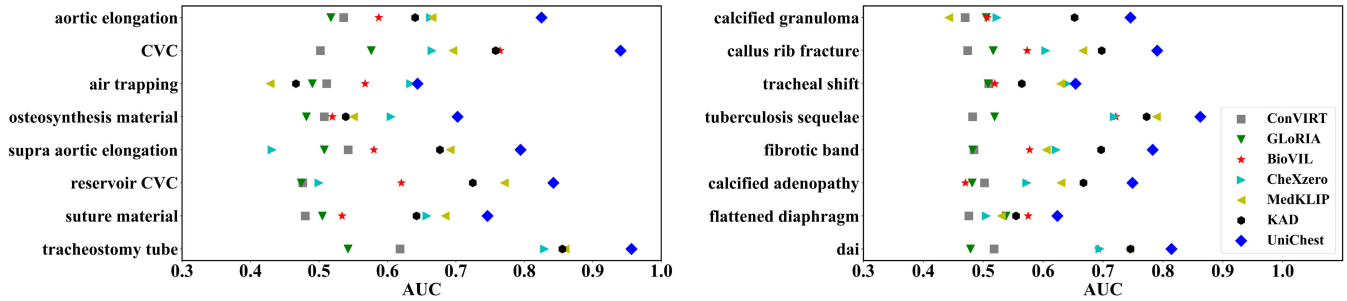
Fig. 6. Per-category performance of 16 randomly selected *unseen* categories with more than 50 positive samples in PadChest.
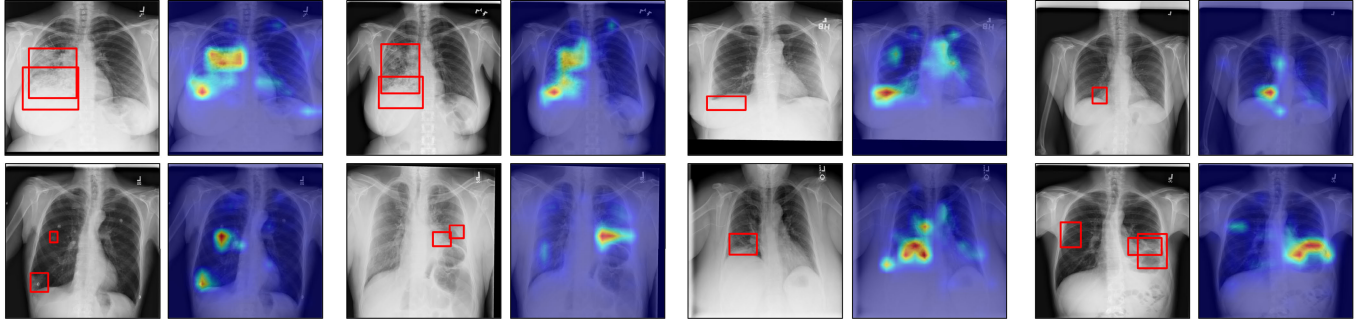

Fig. 7. Some visualization examples of lesion grounding performance of UniChest on ChestX-Det10. In the left CXR images, bounding boxes are abnormal areas manually annotated by the consensus of some board-certified radiologists. In the right heatmaps generated by UniChest, the color temperature in the heatmaps indicates the attention focus of the model, with higher temperatures indicating greater attention and suggesting a higher likelihood of abnormality in the corresponding region.

TABLE VI
ABLATION STUDY ON DIFFERENT STAGES OF UNICHEST. FOUR METRICS INCLUDING aAUC, aF1, aACC AND mAP SCORES ARE REPORTED.
FOR SINGLE-LABELED DATASET, WE REPORT AUC, F1, ACC AND AP SCORES

Dataset Method	Shenzhen				SIIM-ACR Pneumothorax				ChestX-ray14			
	AUC↑	F1↑	ACC↑	AP↑	AUC↑	F1↑	ACC↑	AP↑	aAUC↑	aF1↑	aACC↑	mAP↑
KAD *(single-source SOTA)*	0.8856	0.8474	0.7893	0.8141	0.8967	0.6833	0.8428	0.6869	0.7933	0.3363	0.8639	0.2746
UniChest *(only Conquer stage)*	0.9013	0.8558	0.8594	0.9313	0.9019	0.6450	0.8494	0.6601	0.8468	0.4143	0.8957	0.3543
UniChest *(with equal weights)*	0.9463	0.8851	0.8838	0.9508	0.8687	0.5867	0.8234	0.5507	0.8528	0.4250	0.8971	0.3722
UniChest	**0.9499**	**0.8906**	**0.8913**	**0.9598**	**0.9260**	**0.6928**	**0.8807**	**0.6929**	**0.8584**	**0.4293**	**0.8999**	**0.3797**

the MoE-QN Module's regional cross-attention maps in transformer decoder layers. By comparing the model-detected lesions with the bounding boxes annotated by expert clinicians, we observe a strong alignment between model findings and the diagnoses made by experts, demonstrating the reasoning ability and interpretability of UniChest.

V. ABLATION STUDY

A. The Effectiveness of Conquer-and-Divide Framework

1) On "Conquer" Stage: Data replenishment from multiple sources is an intractable issue mentioned in Section III. We compare the impact of incorporating multi-source training data in the "Conquer" stage of our UniChest on the diagnostic performance, as displayed in the second row of Table VI. It turns out that simply augmenting the training data with multiple sources does not consistently yield significant performance gains compared to using a single source, as shown in the first row of the table. According to Table VI, explicit

data replenishment improves the performance on ChestX-ray14 by 6.08%, aligning with the training data distribution. However, the performance of the zero-shot evaluation varies between the two inference sets. For the Shenzhen dataset, the influence of the "Conquer" stage has a significantly positive effect, as all four metrics experience considerable improvement. In the case of SIIM-ACR Pneumothorax, the AUC and ACC values are comparable to previous state-of-the-art approaches, but the other two metrics decrease by approximately 3.26%.

2) On "Divide" Stage: A comparison of the second and fourth rows of Table VI reveals that UniChest significantly outperforms the "Conquer" stage in terms of zero-shot generalization by 4.86% (AUC), 3.48% (F1), 3.19% (ACC), and 2.85% (AP) for Shenzhen, and 2.41% (AUC), 4.78% (F1), 3.13% (ACC), and 3.28% (AP) for SIIM-ACR Pneumothorax. To further present the effectiveness of the soft-gating mechanism and $\mathcal{L}_{SCL}$ in the MoE-QN structure during the "Divide" stage, we present the results of removing the weight
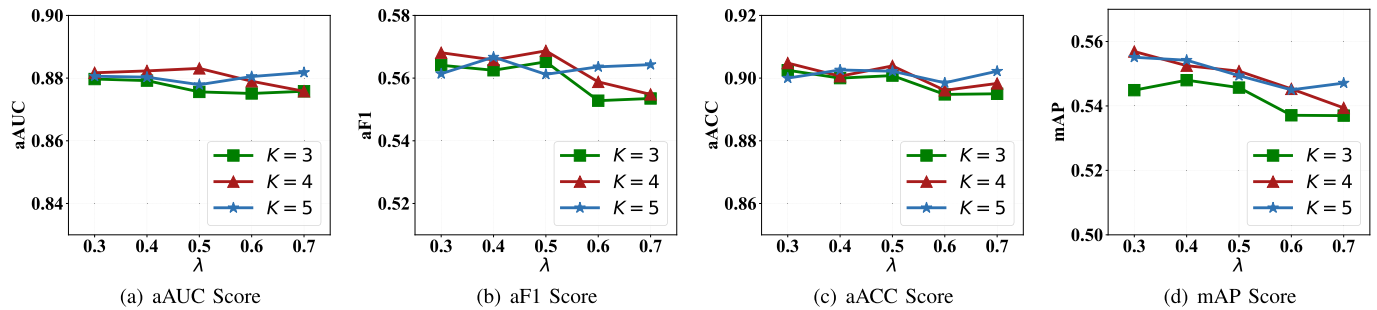
(a) aAUC Score   (b) aF1 Score   (c) aACC Score   (d) mAP Score

Fig. 8. Ablation study on different hyper-parameter combinations, where λ is the weight of the first query network during the "Divide" stage, and K is the total number of query networks.

TABLE VII

ABLATION STUDY ON THE ROLE OF THE REPORT. REPORT IN THE FIRST ROW MEANS ADOPTING REPORT ENTITIES AS THE INPUT OF THE TEXT ENCODER IF THE SAMPLE IS WITH REPORT. "LABEL CONCAT." DENOTES WE CONCATENATE CORRESPONDING POSITIVE LABELS AS THE INPUT. "PARTIAL" MEANS ADOPTING "LABEL CONCAT." ONLY FOR SAMPLES WITHOUT REPORTS WHILE "ALL" MEANS ADOPTING "LABEL CONCAT." REGARDLESS OF WHETHER ACCOMPANIED BY REPORTS OR NOT

| Report | Label Concat. | | aAUC | aF1 | aACC | mAP |
	Partial	All				
			0.8716	0.5450	0.8939	0.5175
✓			0.8742	0.5488	0.8926	0.5217
		✓	0.8733	0.5624	0.8995	0.5458
✓	✓		**0.8831**	**0.5687**	**0.9039**	**0.5508**

TABLE VIII

ABLATION STUDY ON DIFFERENT MODALITY ENCODERS

Method	aAUC	aF1	aACC	mAP
w/ DenseNet-121	0.8794	0.5531	**0.9078**	0.5410
w/ ClinicalBERT	**0.8849**	0.5643	0.8997	0.5427
UniChest	0.8831	**0.5687**	0.9039	**0.5508**

generation process and replacing it with equal weights in the third row of Table VI. As shown on SIIM-ACR Pneumothorax, our mechanism significantly enhances zero-shot generalization. In addition to the performance improvement, it's also important to note that the introduction of the MoE-QN structure at the "Divide" stage does not significantly increase the computational cost compared to previous single-stage pre-training frameworks. Concretely, the increase in computational complexity is marginal, growing only from 44.3 GFLOPs in KAD to 45.5 GFLOPs in UniChest, which constitutes an increase of approximately 2.7%.

B. The Robustness Under Different Hyper-Parameters

The main hyper-parameters in our UniChest framework are the weight λ of the first query network during the "Divide" stage, and the total number of query networks K. In the previous sections, we set the default values of λ as 0.5 and K as 4. In Fig. 8, we present results for various combinations of $\lambda \in [0.3, 0.7]$ and $K \in [3, 5]$, showcasing the average numerical outcomes across six different datasets, including ChestX-ray14, CheXpert, VinDr-CXR, Shenzhen, SIIM-ACR Pneumothorax, and Open-I. For instance, the default setting achieves best numerical results as to aAUC and aF1, while the range of variation for the maximum and minimum values of aAUC and aACC are below 1%, which validates the robustness of our method under different hyper-parameter combinations.

C. The Role of Report in Pre-Training

As stated in Section III, following the paradigm of vision-language pre-training, we use entities extracted from the report by NER tool as the input of the text encoder Φ_{text} if the corresponding report is available, while we concatenate their positive labels as input content for those without reports. In our multi-source pre-training dataset Multi-CXR, there are a total of 685,951 samples, of which 348,900 samples from MIMIC-CXR have reports, meaning that half of the samples come with reports while the rest samples have no corresponding reports. It's worth noting that numerous datasets, which currently present only labels and images, such as CheXpert and ChestX-ray14, were originally accompanied by reports. The labels from these datasets were derived using NLP tools akin to our entity extraction method. Nevertheless, due to certain practical restrictions, including privacy concerns, these reports are not publicly available. Consequently, opting for label concatenation as a substitute is a reasonable manner. To further study the role of the report, we consider three different processing methods. First, we discard the text input directly. Second, we utilize only reports and cancel the input of positive labels for samples without reports. Third, we adopt positive label concatenation as the input for the text encoder for all samples whether they have a report or not. The overall result shown in Table VII drops a little when we underutilize the text input, emphasizing the importance of semantically rich reports in the development of the CXR diagnosis model.

D. Performance Under Different Backbones

We also explore the influence of modality backbones in the whole architecture. Firstly, we substitute ResNet-50 with DenseNet-121 [46] for the visual backbone and then replace the default fine-tuned PubMedBERT with ClinicalBERT [47] as textual encoder. As shown in Table VIII, we notice the

average numerical outcomes across six different datasets of the three settings are comparable, demonstrating the insensitivity of backbone selection.

VI. Discussion and Conclusion

In this paper, we propose a novel Conquer-and-Divide pre-training framework for multi-source Chest X-Rays, namely UniChest, which is among the first attempts to fuse and utilize CXR samples from various origins harmonically. Our method effectively balances the benefits of multi-source data while minimizing the negative impacts the inter-source heterogeneity brings. In the "Conquer" stage, UniChest enhances feature extraction by sharing model components across different sources. In the "Divide" stage, it employs a mixture of deep query modules and utilizes a novel source-contrastive learning loss to isolate source-specific patterns, further reducing cross-source interference. Through extensive experiments on a range of benchmark datasets, we show the robustness and effectiveness of UniChest under diverse settings in CXR diagnosis.

Despite its powerful performance, UniChest still has some limitations in its design and application. Firstly, similar to MedKLIP and KAD, UniChest is limited to generating coarse-grained grounded heatmaps by utilizing cross-attention maps, which falls short of meeting the requirements for precise pixel-level segmentation. Therefore, the development of a comprehensive universal CXR model that combines both classification and fine-grained lesion grounding is a promising avenue for benefiting the medical community, which would not only offer efficiency but also ensure trustworthiness in daily applications. Secondly, although the Conquer-and-Divide framework is precise and meaningful and we have proposed one implementation paradigm within this framework. There is room for further exploration of other concrete frameworks that align with this spirit. We hope that our UniChest will inspire the exploration and utilization of multi-source CXRs.

References

[1] J. Kim and K. H. Kim, "Role of chest radiographs in early lung cancer detection," *Transl. Lung Cancer Res.*, vol. 9, no. 3, pp. 522–531, Jun. 2020.

[2] H. J. Shin, K. Han, L. Ryu, and E.-K. Kim, "The impact of artificial intelligence on the reading times of radiologists for chest radiographs," *NPJ Digit. Med.*, vol. 6, no. 1, p. 82, Apr. 2023.

[3] W. Wang et al., "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14408–14419.

[4] B. Boecking et al., "Making the most of text semantics to improve biomedical vision–language processing," in *Proc. ECCV*, 2022, pp. 1–21.

[5] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," *Nature Biomed. Eng.*, vol. 6, no. 12, pp. 1399–1406, Sep. 2022.

[6] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "MedKLIP: Medical knowledge enhanced language-image pre-training for X-ray diagnosis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21372–21383.

[7] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nature Commun.*, vol. 14, no. 1, p. 4542, Jul. 2023.

[8] A. E. W. Johnson et al., "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*.

[9] R. Dale, "GPT-3: What's it good for?" *Natural Lang. Eng.*, vol. 27, no. 1, pp. 113–118, 2021.

[10] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[11] Y. Yang, H. Wang, and D. Katabi, "On multi-domain long-tailed recognition, imbalanced domain generalization and beyond," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 57–75.

[12] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[13] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 590–597.

[14] H. Q. Nguyen et al., "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Sci. Data*, vol. 9, no. 1, p. 429, Jul. 2022.

[15] S. Jaeger, S. Candemir, S. Antani, Y. X. Wang, P. X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant. Imag. Med. Surg.*, vol. 4, p. 475, Dec. 2014.

[16] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016.

[17] (2019). *Society for Imaging Informatics in Medicine: Siim-ACR Pneumothorax Segmentation*. [Online]. Available: https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation

[18] B. Van Ginneken, B. M. Ter Haar Romeny, and M. A. Viergever, "Computer-aided diagnosis in chest radiography: A survey," *IEEE Trans. Med. Imag.*, vol. 20, no. 12, pp. 1228–1241, Dec. 2001.

[19] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "Artificial intelligence in radiology," *Nat. Rev. Cancer*, vol. 18, no. 8, pp. 500–510, 2018.

[20] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," *Med. Phys.*, vol. 47, no. 5, pp. e218–e227, 2020.

[21] S. T. H. Kieu, A. Bade, M. H. A. Hijazi, and H. Kolivand, "A survey of deep learning for lung disease detection on medical images: State-of-the-art, taxonomy, issues and future directions," *J. Imag.*, vol. 6, no. 12, p. 131, Dec. 2020.

[22] D. Meedeniya, H. Kumarasinghe, S. Kolonne, C. Fernando, I. D. L. T. Díez, and G. Marques, "Chest X-ray analysis empowered with deep learning: A systematic review," *Appl. Soft Comput.*, vol. 126, Sep. 2022, Art. no. 109319.

[23] S. A. Khoiriyah, A. Basofi, and A. Fariza, "Convolutional neural network for automatic pneumonia detection in chest radiography," in *Proc. Int. Electron. Symp. (IES)*, Sep. 2020, pp. 476–480.

[24] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan, *Transfer Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[25] S. Won Lee et al., "Darwin's neural network: AI-based strategies for rapid and scalable cell and coronavirus screening," 2020, *arXiv:2007.11653*.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[27] R. Choudhuri and A. Paul, "Multi class image classification for detection of diseases using chest x ray images," in *Proc. 8th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2021, pp. 769–773.

[28] M. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem, "Efficient pneumonia detection in chest Xray images using deep transfer learning," *Diagnostics*, vol. 10, no. 6, p. 417, Jun. 2020.

[29] V. Chouhan et al., "A novel transfer learning based approach for pneumonia detection in chest X-ray images," *Appl. Sci.*, vol. 10, no. 2, p. 559, Jan. 2020.

[30] F. Bianchi, G. Attanasio, R. Pisoni, S. Terragni, G. Sarti, and S. Lakshmi, "Contrastive language-image pre-training for the Italian language," 2021, *arXiv:2108.08688*.

[31] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. ICML*, 2021, pp. 4904–4916.

[32] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 9694–9705.

[33] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Proc. 7th Mach. Learn. Healthcare Conf.*, 2022, pp. 2–25.

[34] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3922–3931.

[35] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," 2022, *arXiv:2210.10163*.

[36] R. S. R. Silva and P. Fernando, "Effective utilization of multiple convolutional neural networks for chest X-ray classification," *Social Netw. Comput. Sci.*, vol. 3, no. 6, p. 492, Sep. 2022.

[37] S. Jain et al., "RadGraph: Extracting clinical entities and relations from radiology reports," in *Proc. NeurIPS Datasets Benchmarks Track*, 2021, pp. 1–12.

[38] Y. Gu et al., "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, pp. 1–23, Oct. 2021.

[39] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. 90001, pp. 267D–270, Jan. 2004.

[40] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[41] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 12, pp. 6070–6080, Dec. 2022.

[42] J. Liu, J. Lian, and Y. Yu, "ChestX-det10: Chest X-ray dataset on detection of thoracic abnormalities," 2020, *arXiv:2006.10550*.

[43] R. Zhang, Q. Xu, C. Huang, Y. Zhang, and Y. Wang, "Semi-supervised domain generalization for medical image analysis," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[44] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1013–1023.

[45] T. Ridnik et al., "Asymmetric loss for multi-label classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 82–91.

[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[47] E. Alsentzer et al., "Publicly available clinical BERT embeddings," in *Proc. 2nd Clin. Natural Lang. Process. Workshop*. Minneapolis, MI, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78.