Probabilistic Integration of Object Level Annotations in Chest X-ray Classification

Tom van Sonsbeek[1], Xiantong Zhen[1,2], Dwarikanath Mahapatra[2], and Marcel Worring[1]

[1]University of Amsterdam, Amsterdam, the Netherlands
[2]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

Abstract

Medical image datasets and their annotations are not growing as fast as their equivalents in the general domain. This makes translation from the newest, more data-intensive methods that have made a large impact on the vision field increasingly more difficult and less efficient. In this paper, we propose a new probabilistic latent variable model for disease classification in chest X-ray images. Specifically we consider chest X-ray datasets that contain global disease labels, and for a smaller subset contain object level expert annotations in the form of eye gaze patterns and disease bounding boxes. We propose a two-stage optimization algorithm which is able to handle these different label granularities through a single training pipeline in a two-stage manner. In our pipeline global dataset features are learned in the lower level layers of the model. The specific details and nuances in the fine-grained expert object-level annotations are learned in the final layers of the model using a knowledge distillation method inspired by conditional variational inference. Subsequently, model weights are frozen to guide this learning process and prevent overfitting on the smaller richly annotated data subsets. The proposed method yields consistent classification improvement across different backbones on the common benchmark datasets Chest X-ray14 and MIMIC-CXR. This shows how two-stage learning of labels from coarse to fine-grained, in particular with object level annotations, is an effective method for more optimal annotation usage.

1. Introduction

The recent big advances in vision can be attributed to two main factors: algorithmic innovation and large amounts of data. Especially the availability of well-annotated datasets is showing to be a decisive factor [42, 6, 50]. This is a noteworthy development when looking at the role the general
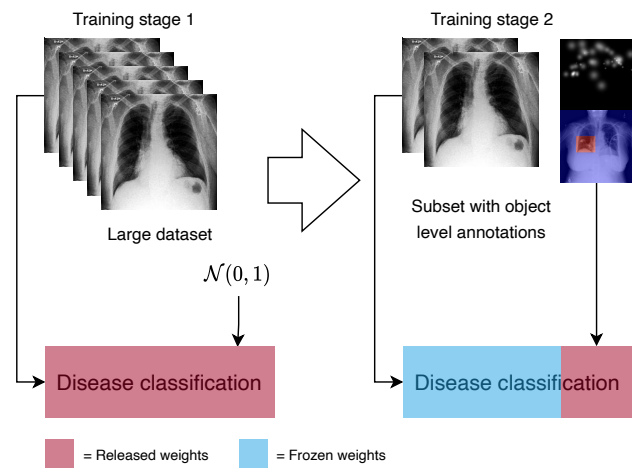

Figure 1: Model architecture overview for probabilistic integration of object level annotations.

vision domain has had on the medical imaging domain in recent years.

Applying the early deep learning based generic computer vision solutions in the medical domain showed to work well, concretely: 1) The introduction of CNNs led to the first deep learning methods which outperformed radiologists[34]. 2) Although there exists an obvious gap between generic and medical data [33, 1] fine-tuning pretrained models from the general vision domain as a basis of new models on medical data gives good results [30, 4].

A remaining challenge is whether the data-driven breakthroughs in the general vision domain can be translated to the medical domain in a similar way. In the general domain it is possible to increase annotated dataset sizes through crowdsourcing, web scraping, and label prediction through sophisticated vision-language methods. This allows for powerful models like the Vision Transformer [6] and many methods building and innovating based on that [29, 3, 49].

Adaptations of these methods are leading to competitive results in the medical domain [41, 10]. However compared to previous methods their effect is not as large ($\sim 1\%$) as seen in the general domain ($\sim 3\%$). This can possibly be attributed to domain shift, scarcity of (annotated) data, and annotation quality.

Getting large scale annotations is more complicated in the medical domain, where reliable annotations originate from trained medical experts [52]. In practise this means that current public medical datasets are small, and fully annotated, or large, and only partly annotated. Automatic sourcing of annotations through metadata or other information sources is an option when it is not possible to obtain expert level annotations on an entire dataset, even though this leads to a lower annotation quality [53, 32].

This shortage of expert level annotations is especially relevant for chest X-ray scans. The chest X-ray is one of the most common medical imaging modalities. The low cost, non-invasive nature and low patient impact explain their use as a primary diagnostic tool. The high volume of scans makes applying automated deep learning methods an interesting avenue that has been explored over the last years. As a consequence, there are multiple publicly available datasets exceeding 100k scans. Annotations on these datasets are often limited to global disease labels extracted from their corresponding radiology reports using Natural Language Processing (NLP) [44, 17]. Usage of radiology reports offers annotations for the entire dataset, but is also prone to unintended mislabeling and biases [39]. Access to higher quality expert-level annotations can contribute to improved accuracies.

Next to global disease labels, many x-ray datasets contain more precise object level annotations, made by clinicians for a subset of the data. At this stage these are: 1) bounding box annotations [44, 24], describing the region of interest (ROI) within the X-ray in which signs of the disease are located. 2) eye gaze information [18, 24], the extracted gaze pattern of clinicians, tracked while they analyze and report on a chest X-ray scan using specialised software. These eye gaze maps contain valuable insights regarding the ROI and the analysis process of clinicians, since they show the exact locations that contribute to the decision making and reporting of the expert annotator. So far these object level annotations have not been widely used for improvement of classification performance. Bounding box information has been used to verify the location-awareness of classification methods. The recently introduced public eye gaze datasets have limited exploration so far.

To improve disease classification, we see an opportunity for a scenario where we use a large dataset with global disease labels together with a smaller subset which contains more rich object level annotations. By doing so we encounter two challenges. First, the method should be able to learn from a dataset containing different granularities of labels, namely global disease labels and smaller subsets of eye gaze information maps from clinicians and disease bounding box annotations. Secondly, when training a deep neural network with small amounts of data, there is a risk of overfitting and loss of generalization to large datasets. We propose a two-stage optimization algorithm to incorporate object level annotations into representation learning of chest X-ray images. This unified training strategy can integrate different types of image labels. In this paper we make the following contributions:

- We propose a new probabilistic latent variable model for disease classification which is able to learn image representations by leveraging annotation of different granularities.

- We propose a two-stage optimization strategy enabling the model to learn low level features with large base datasets and more relevant features to diseases by integrating object level annotations.

- We conduct extensive experiments which show that by combining the variational model and the two-stage training strategy it is possible to consistently improve disease classification performance, across two chest X-ray classification benchmark datasets, by over 3%.

2. Related works

Chest X-ray classification   Strides have been made on the problem of disease classification from chest X-rays over the last years. Since the introduction of ChexNet [34], many new solutions for this classification problem have been provided, pushing the field forward. Methods range from supervised [48, 41, 32, 38, 20], to semi-supervised learning [28, 26]. A number of these methods use object level annotations to check the location-awareness of disease classification [27, 45, 35, 47, 7]. Li *et al.* [25] show that disease localization on large X-ray datasets can be aided by a subset of bounding box annotations. To the best of our knowledge, no existing methods are using object level annotations with the objective of enhancing classification performance on large scale X-ray datasets.

Incorporation of eye gaze information   This year, two datasets containing eye gaze information of clinicians analysing chest X-ray scans were released. Huang *et al.* [15] showed that even a small scale eye gaze dataset has the ability to improve disease classification performance. Zhu *et al.* [54] had similar findings and furthermore demonstrated that gaze information can also be used to generate more useful attention/saliency maps. These findings are promising and show that even with small amounts of data,
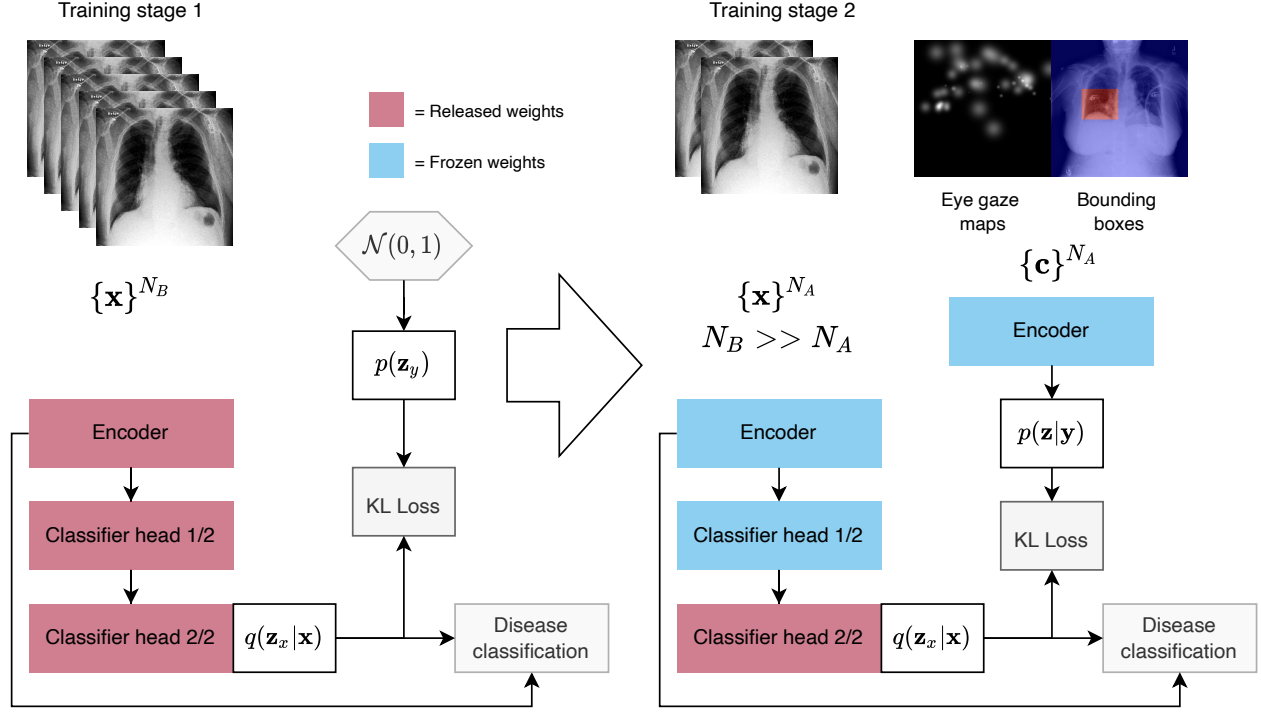
Figure 2: Architecture of probabilistic integration of object level annotations in chest X-ray classification. The architecture is composed of two learning stages: At first the entire model is trained on a large scale base dataset. Secondly fine-grained object level annotations on a data subset are infused through the variational prior, enabling better disease classification performance.

eye gaze maps contain valuable information that improve disease classification results. It is still an open challenge for these methods to generalise to larger datasets, since these models are trained and tested on $< 1\%$ of available chest X-ray images of the dataset they originate from.

Probabilistic latent variable models A major successful application area of probabilistic latent variable models is in the multi-modal domain, due to the inherent versatility of using latent variables across domains. Examples of recent successful applications are in cross-modal retrieval [5] and multi-modal pose generation [23]. The use cases of probabilistic models also extend to the medical domain. For example in multi-modal segmentation of brain MRI scans [8] and abdominal CT scans [51]. Furthermore, multi-modal methods regarding chest X-rays, like report generation [31] and image-text disease classification [43] are based on probabilistic modelling methods. When we consider object level annotations as an additional data modality, the use of probabilistic latent variable models could be suitable.

3. Methodology

Given a chest X-ray image, we would like to classify it into different categories of diseases. We frame disease clas-

sification based on chest X-ray images as a conditional variational inference problem by defining a probabilistic latent variable model. The latent variable is defined as the feature representation of the chest X-ray image. We will first give the preliminaries on variational auto-encoders, based on which we introduce the conditional inference model by designing a conditional prior. After that we will describe our two stage optimization method as shown in Fig. (2).

3.1. Preliminaries

The variational auto-Encoder (VAE) [22, 36] is a probabilistic generative model which has been successful in many applications. The VAE has shown to be effective in learning low-dimensional representations of images.

Specifically, given an input image $\mathbf{x}$ from a data distribution $p(\mathbf{x})$, we would like to learn its n-dimensional vector representation $\mathbf{z}$ in the latent space to infer the posterior $p(\mathbf{z}|\mathbf{x})$ over the latent variable $\mathbf{z}$. However, it is intractable to directly infer the posterior. Instead, we introduce a variational posterior $q(\mathbf{z}|\mathbf{x})$ to approximate $p(\mathbf{z}|\mathbf{x})$ by minimising the KL divergence between them:

$$D_{\text{KL}}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})] \tag{1}$$

By applying the Bayes' rule to Eq. (1), we obtain the

well-known evidence lower bound (ELBO):

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}[\log p(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})], \quad (2)$$

in which the posterior $q(\mathbf{z}|\mathbf{x})$ is dependent on $\mathbf{x}$ and the prior $p(\mathbf{z})$ generally is assumed to be an isotropic Gaussian distribution $\mathcal{N}(0, I)$ over $\mathbf{z}$.

The VAE is primarily used for generative modeling, while in this work, we would like to conduct supervised learning for disease classification. To do so, we design a new objective based on Eq. (2) by replacing the data likelihood with a conditional likelihood $p(\mathbf{y}|\mathbf{z}, \mathbf{x})$. This gives rise to the objective function for supervised learning as follows:

$$\mathcal{L} = \mathbb{E}[\log p(\mathbf{y}|\mathbf{z}, \mathbf{x})] - \beta D_{\text{KL}}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]. \quad (3)$$

where $\mathbf{y}$ is the disease label corresponding to the input image $\mathbf{x}$ and β is the hyperparameter to control the weight of the KL term which is regarded as a regularizer. Intuitively, the objective in Eq. (3) encourages the model to learn a compact latent representation of the input image $\mathbf{x}$ to maximally predict its disease label.

3.2. Conditioning on Object Level Annotations

The prior in Eq. (3) is non-informative which serves to remove the redundancy in the latent representation. However, we would like the prior to be able to incorporate prior knowledge in order to achieve more informative representations. To this end, we propose to design a conditional distribution for the prior by leveraging extra label information provided by object level annotations, in particular bounding boxes and eye gaze maps.

To be more specific, for a given image $\mathbf{x}$ in a data subset, we also have an object level annotation $\mathbf{c}$ associated with it, which has identical dimensions as $\mathbf{x}$. By specifying the prior as $p(\mathbf{z}|\mathbf{c})$ that is conditioned on the gaze map, we obtain a new objective function as follows:

$$\mathcal{L}_{\text{SUB}} = \mathbb{E}[\log p(\mathbf{y}|\mathbf{z}, \mathbf{x})] - \beta D_{\text{KL}}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{c})], \quad (4)$$

The knowledge contained in the gaze map is used through the data-dependent prior by minimising the KL term, which guides the model to extract the most relevant features from the input image. This objective provides a principled formalism to incorporate prior knowledge into representation learning.

3.3. Two-stage optimization

In general, we could directly use the objective in Eq. (4) to learn feature representations. In practice the extra annotations are not available for large scale datasets that usually only contain the global label annotations. The object level annotation, e.g., the gaze maps, are only provided for small scale datasets. In this case, we would not be able train a deep model from scratch on those small datasets.

In order to leverage both large scale datasets and also the annotations in small subsets of them, we propose a two-stage optimization. Our objective functions of variational inference provides the flexibility to do so. We consider a set of images $\{\mathbf{x}\}_B^{N_B}$, with global labels $\mathbf{c}_B$ as a large base dataset which contains only the disease labels. This dataset will be used for pre-training in stage 1. Furthermore we define a subset $\{\mathbf{x}\}_S^{N_S} \in \{\mathbf{x}\}_B^{N_B}$ with $N_B >> N_S$. It contains both additional object level annotations $\mathbf{y}_S$ and also global labels $\mathbf{c}_S$. This dataset will be used for fine-tuning in Stage 2. The learning stages are designed to optimally utilise the existing label space.

Stage 1: Pre-training   In this stage, we train the model on the large scale base dataset by using the optimization objective in Eq. (3). We define the posterior as $q(\mathbf{z}_B|\mathbf{x}_B)$ conditioned on the input image $\mathbf{x}_B$ from the base datasets and the prior $p(\mathbf{z})$ as a normal Gaussian distribution. Leveraging a large dataset during this stage leads to a comprehensive extraction of relevant image features.

Stage 2: Fine-tuning   In this stage, we further fine-tune the model on the small dataset with object level annotations. We train the model with the optimization objective in Eq. (4). In this case, the prior is defined as $p(\mathbf{z}_S|\mathbf{c})$. This final stage allows for an optimal contextualization of image features already learnt in the earlier layers of the model. To combat overfitting and preserve generalization to the base dataset, model weights are frozen in earlier model layers.

3.4. Qualitative evaluation

The reasoning behind this method is that the inclusion of object-level annotations from clinicians can contribute to the model processing the image spatially in a similar way as clinicians do. The last step in which pixel level clinician annotations are infused is meant to accentuate certain features and locations in the image that are important according to the eye gaze and bounding box maps. This should be visible by comparing class activation maps (CAMs) of images from the base dataset before and after the last training step is applied. We define the following metric measuring this, with $S(\cdot)$ being a similarity metric like Mean Squared Error (MSE) or Dice:

$$\Delta S = \frac{S(\mathbf{c}, CAM(\mathbf{x})_{BASE})}{S(\mathbf{c}, CAM(\mathbf{x})_{SUB})} * 100 \quad (5)$$

3.5. Implementation

The implementation of this method (Fig. (2)) is done with deep neural networks by adopting the amortization technique [22]. Both the variational posterior $q(\mathbf{z}|\mathbf{x})$ and the priors $p(\mathbf{z}|\mathbf{c}), p(\mathbf{z})$ are parameterised as fully factorised Gaussian distributions. The reparameterization trick [22]

enables sampling from these distributions: $\mathbf{z}$: $\mathbf{z}^{(\ell)} = f(\mathbf{x}, \epsilon^{(\ell)})$ with $\epsilon^{(\ell)} \sim \mathcal{N}(0, I)$, and $f(\cdot)$ as a deterministic differentiable function.

In the posterior $q(\mathbf{z}|\mathbf{x})$, $\mathbf{x}$ is considered to be the CNN representation of an X-ray image. In a similar fashion, $\mathbf{c}$ in the prior $p(\mathbf{z}|\mathbf{c})$ is a CNN representation of an object level annotation, which could either be a gaze map or bounding boxes. Both prior and posterior are inferred by a multi-layer perceptron (MLP).

4. Data and experimental settings

4.1. Datasets

The proposed method is evaluated on two large public chest X-ray datasets which also contain smaller subsets of object level annotations by clinicians.

4.1.1 Chest X-ray14

This dataset consists of $113,120$ frontal chest X-ray scans [44] of size 1024×1024 with 14 disease labels sourced from the accompanying radiology reports (which are not public) through a rule-based extraction method. A subset of bounding box annotations is available for this dataset. These are 1600 disease bounding boxes distributed over 983 X-ray scans. However, all of these annotations are contained in the test set. Using these annotations means that results can not be reported on the official test set. Instead, five-fold cross validation is applied, similarly adopted by Li *et al.* [25] to solve this issue. We acknowledge that this approach makes comparability with previous methods that evaluate on the entire test set less reliable, since there is a mismatch in the assignment of train and test datasets. However, since this dataset serves as the major benchmark for disease classification of chest X-rays in recent years it is still included.

4.1.2 MIMIC-CXR

This currently largest chest X-ray dataset contains $377,110$ images (sized 2500×3000) which are a combination of frontal and sagittal views for $227,827$ studies in total. Labels are extracted through a process similar to Chest X-ray14, albeit with a slightly different label space [16].

MIMIC-CXR has three different subsets that can produce fine-grained pixel level expert annotations. They originate from REFLACX [24] and EGD-CXR [18]. REFLACX contains eye gaze information for 2616 X-ray scans. Additionally, for the same subset, disease bounding boxes (BB) are also provided. EGD-CXR contains eye gaze maps for 1083 X-ray scans. Since these annotations are spread across train and test sets results can be reported on the official test set.

4.2. Experimental settings

X-ray images are standardised by normalisation and rescaling to size 224×224 with center-cropping. This is in adherence with standards in this field, despite causing a slight performance drop [9]. To measure consistency of our method we evaluate our experiments on commonly used CNN backbones, which are pre-trained on ImageNet [11]. The CNN backbones are: VGG16 [40], ResNet50 [12] and DenseNet121 [14]. Recent works showed the latter works best on Chest X-ray images [34, 46]. Finetuning is applied on the CNN backbone [19, 33].

Posterior $q(\mathbf{z}|\mathbf{x})$ is inferred by two sequential two-layer MLPs with hidden dimension 512. Note that the weights of the first MLP will be kept frozen during the conditioning on object level annotations in training stage 2. Prior $p(\mathbf{z}|\mathbf{c})$ is similarly generated through a two-layer MLP with 512 hidden dimensions.

The object level annotations (bounding boxes, eye gaze maps) are incorporated as pixel level annotations. They are represented as images with values ranging between 0 and 1. These should represent a pseudo-segmentation map of the crucial regions of the image. These maps will be passed through an ImageNet pre-trained CNN encoder of the same type as the X-ray image encoder.

Eye gaze datasets contain fixations points of the radiologist on specific coordinates within the image, combined with how long these fixations lasted. Each fixation is characterised by a Gaussian with radius depending on the fixation's length in seconds (with empirical multiplier $a_{\sigma GAZE} = 10$). Bounding boxes are similarly mapped to the shape of the original X-ray images. To prevent edge issues with the CNN encoder, a Gaussian smoothing of the edges is applied with $\sigma_{BB} = 5$.

Grad-CAM [37] is used to compute CAMs. Training was done with one Ryzen 2990WX CPU and one NVIDIA RTX 2080ti GPU, using Adam [21] optimization and early stopping with a tolerance of 1%.

5. Results and discussion

5.1. Improving disease classification

The performance of our method across different CNN backbones is shown in Table 1. We see that the addition of object level annotations improves classification results compared to base model results. This consistent improvement compared to the baseline model is the main feature of our method. Additionally, it shows to be competitive with prior works, performing better or within $\sim 1\%$ in AUC score for Chest X-ray14, and reaching best scores on MIMIC-CXR. In this comparison, however, the difference in train/test split on Chest X-ray14 should be taken in consideration.

The robustness and consistency of the method are reflected in the effectiveness of our method over multiple CNN backbones and two different datasets. Results on MIMIC-CXR indicate that eye-gaze information is a more valuable object level annotation for our method than bound-

	Backbone	Setting	AUC	F1
Chest X-ray14				
Semi-supervised				
Aviles *et al.* [2]	Graph	-	0.789	-
Liu *et al.* [28]	DenseNet169	-	0.792	-
Liu *et al.* [26]	DenseNet169	-	0.811	-
Supervised				
Wang *et al.* [44]	ResNet50	-	0.745	-
Yao *et al.* [48]	DenseNet121	-	0.761	-
Guendel *et al.* [7]	DenseNet121	-	0.807	-
Kim *et al.* [20]	DenseNet121	-	**0.820**	-
ViT [41]	Transformer	-	0.779	-
Taslimi *et al.* [41]	Transformer	-	0.810	-
Li *et al.* [25][1]	ResNet50	Base model	0.746	-
		+Bounding boxes	0.797	-
Ours[1]	VGG16	Base model	0.754	0.24
		+Bounding boxes	0.786	0.25
	ResNet50	Base model	0.763	0.24
		+Bounding boxes	0.793	**0.26**
	DenseNet121	Base model	0.772	0.24
		+Bounding boxes	0.809	0.25
MIMIC-CXR				
Pooch *et al.* [32]	DenseNet121	-	0.828	-
Seyyed *et al.* [38]	DenseNet121	-	0.834	-
Ours	VGG16	Base model	0.806	0.24
		+BB - REFLACX	0.814	0.25
		+Gaze - REFLACX	0.831	0.26
		+Gaze - EGD-CXR	0.827	0.26
		+Gaze - EGD-CXR & BB - REFLACX	0.829	0.26
		+Gaze - EGD-CXR & Gaze - REFLACX	0.811	0.24
	ResNet50	Base model	0.804	0.24
		+BB - REFLACX	0.813	0.25
		+Gaze - REFLACX	0.831	0.25
		+Gaze - EGD-CXR	0.834	0.26
		+Gaze - EGD-CXR & BB - REFLACX	0.809	0.25
		+Gaze - EGD-CXR & Gaze - REFLACX	0.832	0.26
	DenseNet121	Base model	0.807	0.25
		+BB - REFLACX	0.821	0.26
		+Gaze - REFLACX	0.827	**0.27**
		+Gaze - EGD-CXR	**0.836**	**0.27**
		+Gaze - EGD-CXR & BB - REFLACX	0.815	0.25
		+Gaze - EGD-CXR & Gaze - REFLACX	0.835	**0.27**

Table 1: Disease classification performance of proposed method in AUC and F1 scores. Base model scores indicate the performance after the training stage I with a large base dataset. Other scores indicate the performance on the base dataset test set after the integration of object level annotations subsets in training stage 2.

---

[1]Five-fold cross validation

	Setting	ΔMSE (%)	ΔDice (%)
Chest X-ray14			
VGG16	Base model	-	-
	+Bounding boxes	+14	+5
ResNet50	Base model	-	-
	+Bounding boxes	+15	+8
DenseNet121	Base model	-	-
	+Bounding boxes	+6	+7
MIMIC-CXR			
VGG16	Base model	-	-
	+Bounding boxes	+7	+4
	+Gaze - REFLACX	+12	+10
	+Gaze - EGD-CXR	+13	+9
	+Gaze - EGD-CXR & BB - REFLACX	+3	+4
	+Gaze - EGD-CXR & Gaze - REFLACX	+13	+9
ResNet50	Base model	-	-
	+Bounding boxes	+8	+6
	+Gaze - REFLACX	+15	+9
	+Gaze - EGD-CXR	+16	+8
	+Gaze - EGD-CXR & BB - REFLACX	+7	+6
	+Gaze - EGD-CXR & Gaze - REFLACX	+16	+8
DenseNet121	Base model	-	-
	+Bounding boxes	+8	+5
	+Gaze - REFLACX	+17	+8
	+Gaze - EGD-CXR	+15	**+11**
	+Gaze - EGD-CXR & BB - REFLACX	**+19**	+5
	+Gaze - EGD-CXR & Gaze - REFLACX	+14	+7

Table 2: Difference in MSE and Dice similarity score between GRADCAM activation map and object level clinician annotation before and after fine-tuning with this subset of object level annotations.

ing boxes. A peculiar finding in Table 1 is the performance difference between integration of REFLACX and EGD-CXR eye gaze maps. The sizes of these datasets are ∼3k and ∼1k respectively. It would be expected that the larger dataset would be more effective in leveraging higher classification scores, while the contrary is shown in the results. It can not be said with certainty to which dataset property this can be attributed to, but we can conclude that data quality in these object level annotations is an important factor. Simultaneous infusion of object annotation of both sources (EGD-CXR and REFLACX) was tested on MIMIC-CXR. The results reveal that that simultaneous integration of two eye gaze subsets yields better results than integration of a bounding box and eye gaze subset. Therefore we conclude that a certain consistency within the object level annotations is beneficial to reach optimal performance. Results split by disease class are listed in supplementary material A.

	MIMIC-CXR subset			Chest X-ray14 subset	VGG16		ResNet50		Densenet121	
	REFLACX gaze	REFLACX BB	EGD Gaze	BB	AUC	F1	AUC	F1	AUC	F1
Chest X-ray 14	✗	✗	✗	✗	0.754	0.24	0.763	0.25	0.772	0.24
	✓	✓	✓	✓	0.769	0.24	0.757	0.25	0.789	0.25
	✓	✓	✓	✗	0.737	0.24	0.731	0.22	0.766	0.24
	✗	✗	✗	✓	**0.786**	**0.25**	**0.793**	**0.26**	**0.809**	**0.26**
MIMIC-CXR	✗	✗	✗	✗	0.806	0.24	0.804	0.24	0.807	0.25
	✓	✓	✓	✓	0.813	0.24	0.818	**0.26**	0.826	0.25
	✗	✗	✗	✓	0.794	0.23	0.788	0.24	0.798	0.24
	✓	✓	✓	✗	**0.827**	**0.26**	**0.834**	**0.26**	**0.836**	**0.27**

Table 3: Cross-domain performance, for which annotation subsets (eye gaze maps and bounding boxes (BB)) are exchanged between datasets during the second training step. In this setting the base model is trained only on the source dataset (left column).

				VGG16		ResNet50		Densenet121	
	Encoder	CH1	CH2	AUC	F1	AUC	F1	AUC	F1
Chest X-ray14	✗	✓	✗	0.751	0.25	0.771	0.25	0.759	0.24
	✗	✓	✓	0.767	0.25	0.764	**0.26**	0.768	**0.26**
	✓	✓	✓	0.723	0.25	0.722	0.24	0.723	0.23
	✗	✗	✓	**0.786**	**0.26**	**0.793**	**0.26**	**0.809**	**0.26**
MIMIC-CXR	✗	✓	✗	0.779	0.19	0.765	0.17	0.782	0.16
	✗	✓	✓	0.783	0.20	0.788	0.17	0.792	0.18
	✓	✓	✓	0.690	0.16	0.703	0.16	0.710	0.17
	✗	✗	✓	**0.827**	**0.26**	**0.834**	**0.26**	**0.836**	**0.27**

Table 4: Ablation study over second training step. The checkmark indicates whether the model components' weights are released during the second training step. The model components (Fig. 2) are: encoder, classifier head (CH) 1 and 2.

5.2. Grad-CAM similarity to object level annotations

By using our method on integration of expert object level annotations, we guide our classification method to consider an X-ray image in a spatially similar way as a clinician. Table 2 shows the difference in similarity between the CAM of an X-ray and their object level annotation after (1) the base dataset training and (2) after the subset training with object level annotations. A positive value means that the similarity to the object level annotations increased after training with the subset containing object level annotations. As the results in Table 2 show an increase in these similarity scores, we confirm that the reasoning behind our methodology is valid. The increase in similarity scores is higher for eye gaze maps than for bounding boxes. This can be an indication that eye gaze maps are a more informative object level annotation than bounding boxes.

In Fig. (3) several examples are shown to illustrate how infusion of object level annotations can benefit classification capability. Fig. (3A) shows how the eye gaze pattern of the clinician has a bottom-left focus, which seems to be adopted through training stage 2. After this stage the model no longer incorrectly detects the 'Effusion' label. In Fig. (3B-C) similar GradCAM shift pattern can be observed. With an unexpressive eye gaze map as in Fig. (3D) we see limited effects of conditioning on the gaze map. In

this instance we can see that the missed 'Infiltration' label after stage 1 training is also not detected after stage 2 training.

5.3. Cross-dataset learning

Another interesting but challenging setting to evaluate our method on is when the object-level annotations used in the second training step originate from another similar dataset. The transferability and similarity within chest X-ray datasets has been studied in earlier works and indicated that there is a non-negligible domain shift between current public chest X-ray datasets [33, 19, 32, 13]. Our results in Table 3 confirm this domain shift. Using the annotation subset of a different base dataset consistently does not lead to classification improvements. Interestingly though, when including more out-of-dataset subsets, the performance compared to base dataset training can improve, while still being lower compared to in-dataset training.

5.4. Effect of layer freezing in two-stage training

We show different settings with frozen or released weights over the final training step where we condition the model on object-level annotations in Table 4. These experiments confirm the benefit of only updating the model weights in the last model layers when fine-tuning with a small data subset. Unfreezing the weights of the earlier layers leads to a steep decline in performance. Furthermore
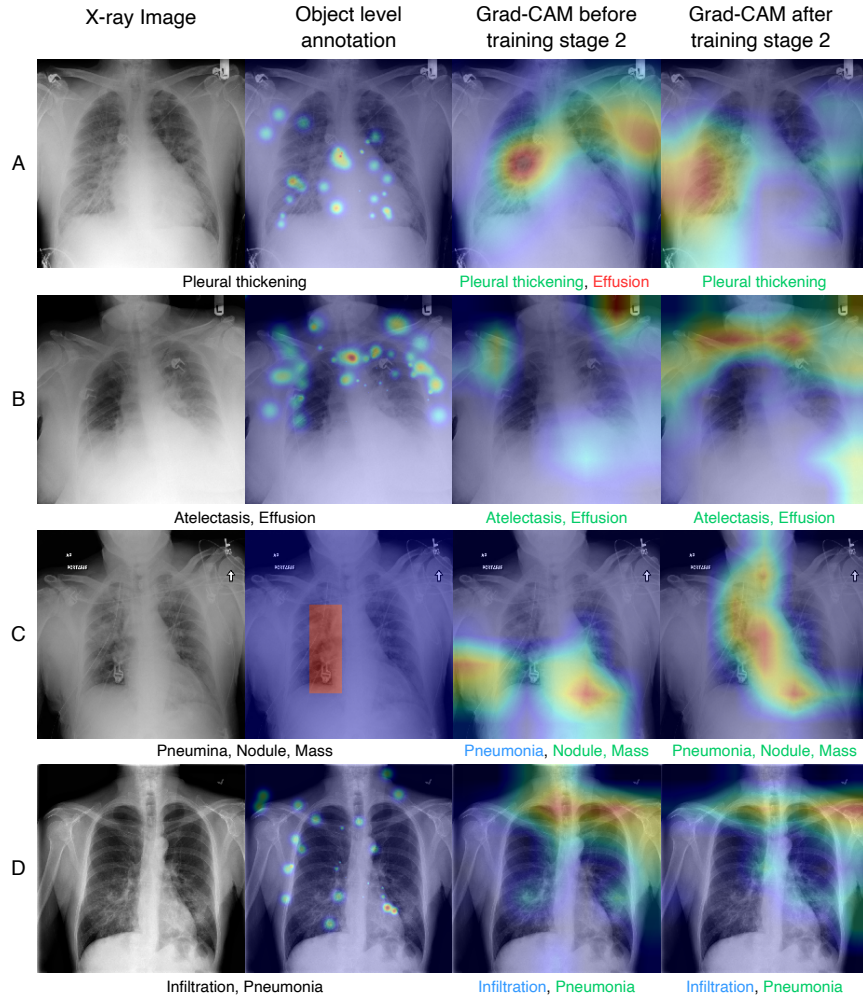
Figure 3: Visualisation of change in Grad-CAM before and after the second training stage with conditioning on object level annotations. Green, blue and red labels stand for correct, missed and wrong prediction respectively. (A) EGD-CXR eye gaze (B) REFLACX eye gaze (C) Chest X-ray14 bounding box (D) REFLACX eye gaze.

they show that fine-tuning of the last layers is needed to obtain the best classification results. An additional downside of unfreezing the earlier model layers is the faster occurrence of overfitting which is also detrimental for the generalization on the classification on the base dataset.

6. Conclusion

In this paper we introduce a probabilistic latent variable model for classification of chest X-rays. It tackles the problem of label scarcity in the medical imaging domain. This model is able to process different types of label granularities, resulting in an efficient usage of all available labels. To achieve this, a two-stage method is introduced. In its first stage global image features are learnt through disease classification on a large base dataset. In the second stage, model weights in earlier layers are frozen. This enables condi-

tioning on a small data subset of object level annotations in the form of eye gaze maps and bounding boxes for better contextualization of features learnt in the first training stage. This simple, yet effective, approach proves it effectiveness by consistently improving performance on multiple datasets. It provides an interesting outlook on how to manage data and how to utilise smaller subsets of rich annotated data in an effective manner.

Acknowledgements

References

[1] Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J Humaidi, Omran Al-Shamma, Mohammed A Fadhel, Jinglan Zhang, J Santamaría, and Ye Duan. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13(7):1590, 2021.

[2] Angelica I Aviles-Rivero, Nicolas Papadakis, Ruoteng Li, Philip Sellars, Qingnan Fan, Robby T Tan, and Carola-Bibiane Schönlieb. Graphxnet - chest x-ray classification under extreme minimal supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 504–512. Springer, 2019.

[3] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[4] Veronika Cheplygina. Cats or cat scans: Transfer learning from natural or medical image source data sets? *Current Opinion in Biomedical Engineering*, 9:21–27, 2019.

[5] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8415–8424, June 2021.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[7] 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.

[8] Mohammad Hamghalam, Alejandro F Frangi, Baiying Lei, and Amber L Simpson. Modality completion via gaussian process prior variational autoencoders for multi-modal glioma segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 442–452. Springer, 2021.

[9] Md Inzamam Ul Haque, Abhishek K Dubey, and Jacob D Hinkle. The effect of image resolution on automated classification of chest x-rays. *medRxiv*, 2021.

[10] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. Transformers in medical image analysis: A review. *arXiv preprint arXiv:2202.12165*, 2022.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, June 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 3–13. Springer, 2021.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE CVPR*, pages 4700–4708, 2017.

[15] Yifei Huang, Xiaoxiao Li, Lijin Yang, Lin Gu, Yingying Zhu, Hirofumi Seo, Qiuming Meng, Tatsuya Harada, and Yoichi Sato. Leveraging human selective attention for medical image analysis with limited training data. *arXiv preprint arXiv:2112.01034*, 2021.

[16] 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.

[17] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

[18] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific data*, 8(1):1–18, 2021.

[19] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y Ng, and Pranav Rajpurkar. Chextransfer: performance and parameter efficiency of imagenet models for chest x-ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 116–124, 2021.

[20] 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*, pages 15719–15728, 2021.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International conference on learning representations*, 2014.

[23] Jogendra Nath Kundu, Rahul M V, Jay Patravali, and Venkatesh Babu RADHAKRISHNAN. Unsupervised cross-dataset adaptation via probabilistic amodal 3d human pose completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[24] Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. Reflacx, a dataset of reports and eye-tracking data for local-

ization of abnormalities in chest x-rays. *Scientific data*, 9, 2022.

[25] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8290–8299, 2018.

[26] Fengbei Liu, Yu Tian, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Self-supervised mean teacher for semi-supervised chest x-ray classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 426–436. Springer, 2021.

[27] Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[28] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE transactions on medical imaging*, 39(11):3429–3440, 2020.

[29] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022.

[30] Mohammad Amin Morid, Alireza Borjali, and Guilherme Del Fiol. A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine*, 128:104115, 2021.

[31] Ivona Najdenkoska, Xiantong Zhen, Marcel Worring, and Ling Shao. Variational topic inference for chest x-ray report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 625–635. Springer, 2021.

[32] Eduardo HP Pooch, Pedro Ballester, and Rodrigo C Barros. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In *International Workshop on Thoracic Image Analysis*, pages 74–83. Springer, 2020.

[33] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

[34] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:1711.05225*, 2017.

[35] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Phil Chen, Amirhossein Kiani, Jeremy Irvin, Andrew Y Ng, and Matthew P Lungren. Chexpedition: investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting. *arXiv preprint arXiv:2002.11379*, 2020.

[36] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv:1401.4082*, 2014.

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE ICCV*, pages 618–626, 2017.

[38] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[39] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] Sina Taslimi, Soroush Taslimi, Nima Fathi, Mohammadreza Salehi, and Mohammad Hossein Rohban. Swinchex: Multi-label classification on chest x-ray images with transformers. *arXiv preprint arXiv:2206.04246*, 2022.

[42] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

[43] Tom van Sonsbeek, Xiantong Zhen, Marcel Worring, and Ling Shao. Variational knowledge distillation for disease classification in chest x-rays. In *International Conference on Information Processing in Medical Imaging*, pages 334–345. Springer, 2021.

[44] 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 CVPR*, pages 2097–2106, 2017.

[45] Joy Wu, Yaniv Gur, Alexandros Karargyris, Ali Bin Syed, Orest Boyko, Mehdi Moradi, and Tanveer Syeda-Mahmood. Automatic bounding box annotation of chest x-ray data for localization of abnormalities. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 799–803. IEEE, 2020.

[46] Yuan Xue and Xiaolei Huang. Improved disease classification in chest x-rays with transferred features from report generation. In *International Conference on Information Processing in Medical Imaging*, pages 125–138. Springer, 2019.

[47] Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang. Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 103–110, 2018.

[48] Li Yao, Jordan Prosky, Eric Poblenz, Ben Covington, and Kevin Lyman. Weakly supervised medical diagnosis

and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*, 2018.

[49] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021.

[50] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022.

[51] Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 589–599. Springer, 2021.

[52] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[53] Tongxue Zhou, Su Ruan, and Stéphane Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, page 100004, 2019.

[54] Hongzhi Zhu, Robert Rohling, and Septimiu Salcudean. Multi-task unet: Jointly boosting saliency prediction and disease classification on chest x-ray images. *arXiv preprint arXiv:2202.07118*, 2022.