PSScreen: Partially Supervised Multiple Retinal Disease Screening

Boyi Zheng
boyi.zheng@oulu.fi
Qing Liu*
qing.liu@oulu.fi

Center for Machine Vision and Signal Analysis (CMVS)
University of Oulu
Oulu, Finland

Abstract

Leveraging multiple partially labeled datasets to train a model for multiple retinal disease screening reduces the reliance on fully annotated datasets, but remains challenging due to significant domain shifts across training datasets from various medical sites, and the label absent issue for partial classes. To solve these challenges, we propose **PSScreen**, a novel **P**artially **S**upervised multiple retinal disease **Screen**ing model. Our PSScreen consists of two streams and one learns deterministic features and the other learns probabilistic features via uncertainty injection. Then, we leverage the textual guidance to decouple two types of features into disease-wise features and align them via feature distillation to boost the domain generalization ability. Meanwhile, we employ pseudo label consistency between two streams to address the label absent issue and introduce a self-distillation to transfer task-relevant semantics about known classes from the deterministic to the probabilistic stream to further enhance the detection performances. Experiments show that our PSScreen significantly enhances the detection performances on six retinal diseases and the normal state averagely and achieves state-of-the-art results on both in-domain and out-of-domain datasets. Codes are available at https://github.com/boyiZheng99/PSScreen.

1 Introduction

Automated detection of retinal disease with fundus images is crucial for efficient and cost-effective large-scale population screening. Fueled by the release of open-access datasets summarized in Fig. 1(a) for specific retinal diseases such as DDR [18] for diabetic retinopathy screening, REFUGE2 [9] for glaucoma screening, PALM [10] for myopia screening etc., numerous works [4, 24, 32, 37] have been developed to train disease-specific screening models on individual training datasets. Although these models are promising in screening for specific diseases on images within specific domains, they are still far from real-world applications where screening for as many retinal diseases as possible on images from various or even unseen domains is desired. Developing a screening model for multiple retinal diseases with strong domain generalization ability is of great significance, yet remains challenging.

The most intuitive way is to train a disease screening model in a fully supervised way with fully labeled training data as illustrated in Fig. 1(b). For example, TrustDetector [12]

[1]* Corresponding author

(a) Open-access datasets for retinal disease screening coming from various domains and following different distributions.

Open-access datasets		Labels for retinal diseases and the normal state							
names	notations and distributions	normal	diabetic retinopathy	glaucoma	cataract	age-related macular degeneration	hypertensive retinopathy	pathologic myopia	
Kaggle-CAT	$D_1 \sim \mathcal{P}_1$	✓	?	✓	✓	?	?	?	partially labelled
DDR	$D_2 \sim \mathcal{P}_2$	?	✓	?	?	?	?	?	
REFUGE2	$D_3 \sim \mathcal{P}_3$	?	?	✓	?	?	?	?	
ADAM	$D_4 \sim \mathcal{P}_4$	?	?	?	?	✓	?	?	
Kaggle-HR	$D_5 \sim \mathcal{P}_5$	?	?	?	?	?	✓	?	
PALM	$D_6 \sim \mathcal{P}_6$	?	?	?	?	?	?	✓	
ODIR	$D_{full} \sim \mathcal{P}_{full}$	✓	✓	✓	✓	✓	✓	✓	fully labelled

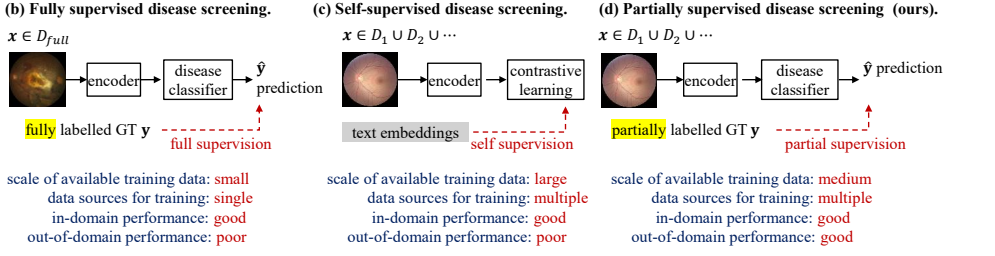


Figure 1: Exampled open-access datasets for retinal disease screening and screening model comparisons under three learning paradigms. (a) lists open-access datasets where "✓" indicates labels for diseases are available while "?" denotes labels are not available. From (b) to (d), we illustrate the pipeline and characteristics of the fully supervised screening model usually trained with a fully labeled dataset, the self-supervised screening model trained with multiple datasets consisting of image-text pairs, and the partially supervised screening model trained with multiple partially labeled datasets.

trains a screening model for three retinal diseases as the dataset collected only provides labels for three diseases while the method in [17] trains the model on ODIR for multiple retinal disease screening. Although they achieve promising performances, the size of training data is limited and they assume that training and test images share the same distribution, which limits their generalization ability to out-of-domain data. The second way is to collect and manually annotate a large-scale dataset e.g. Retina-1M [14], to enable the fully supervised training. However, the annotation is labor-intensive and costly. More recently, with the adaptability of foundation models such as FLAIR [31] and RET-CLIP [6] which are trained with large-scale image-text pairs in a self-supervised learning way as illustrated in Figure 1(c), zero-shot learning for retinal disease screening is emerging. Although large-scale fully annotated training data is not required, the performances for specific disease screening is poor [35].

In this paper, we propose **PSScreen**, a **P**artially **S**upervised multiple retinal disease **Screen**ing model. As illustrated in Fig. 1(d), our PSScreen trains a screening model with multiple partially labeled datasets following different data distributions. Unlike previous partially supervised learning methods for natural images [2, 3, 26, 27] and medical images [34], where both training and testing data come from the same dataset without domain shift, the partially supervised learning in this paper is more challenge due to (1) domain shifts among training datasets collected from various medical sites, (2) the label absent issue for partial classes. To address them, we propose a two-stream network: one learns deterministic features and the other learns probabilistic features via uncertainty injection. Then, under the

guidance of textual information, we decouple two types of features into disease-wise features and align them via feature distillation so that the features learned are robust against the domain shifts. Meanwhile, PSScreen addresses the label absent issue via pseudo label consistency between two streams. Finally, we introduce a self-distillation to transfer task-relevant semantics about known classes from the deterministic to the probabilistic stream to further enhance the detection performances. Experiments show that PSScreen outperforms state-of-the-art methods on in-domain and out-of-domain datasets.

To summarize, our contributions are as follows:

- We propose PSScreen, a multiple retinal disease screening method which trains the disease screening model on a meta-dataset composed of multiple partially labeled datasets following different distributions. To the best of our knowledge, we are the first to train a partially supervised model for screening multiple retinal diseases using multiple datasets from various medical sites.

- We propose a two-stream network that learns deterministic features and uncertainty-injected probabilistic features and aligns them to boost the domain generalization ability at the feature level. Meanwhile, we introduce pseudo label consistency between two streams to address the label absent issue and a self-distillation to transfer the task-relevant semantics about known classes from the deterministic to the probabilistic stream to further enhance the detection performances.

- We validate our PSScreen on the meta-dataset, and demonstrate that our PSScreen achieves state-of-the-art performances. More importantly, validation on six unseen datasets further shows that PSScreen achieves superior domain generalization capability over previous methods.

2 Proposed Method

Problem Formulation. We suppose that (1) there is a meta-dataset $\mathcal{D} = \{D_1, D_2, \ldots, D_K\}$ consisting of K partially labeled datasets which are collected from various medical sites and follow different distributions, (2) each dataset $D_k = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_k}$ with N_k samples where $\mathbf{x}_i$ is the i-th sample and $\mathbf{y}_i \in \{1, 0, -1\}^T$ is the label for T retinal diseases and $y_{i,t} = 0$ indicating that the t-th disease label is unknown, while $y_{i,t} = 1/-1$ indicating the t-th disease label is positive/negative. For simplicity, we use the vector $\boldsymbol{\delta}_i = \mathbb{1}_{\{1,-1\}}(\mathbf{y}_i)$ to indicate whether the label for diseases is known or unknown, where $\mathbb{1}(\cdot)$ is an indicator function. Our goal is to train a multi-label disease screening model on $\mathcal{D}$, which can predict risks for T diseases and well generalize to out-of-domain test data.

Overview. Fig. 2 illustrates the overview of *PSScreen*. PSScreen is a two-stream network composed of six key modules: 1) Deterministic feature learning which outputs deterministic feature maps to preserve task-relevant semantics, 2) Probabilistic feature learning via attaching Domain Shifts with Uncertainty (DSU) block [19] at each stage of the backbone to produce probabilistic feature maps, 3) Text-guided semantic decoupling which decouples global feature maps from the two streams into disease-wise features respectively, 4) Feature distillation which aligns deterministic and probabilistic disease-wise features in the latent space to boost the domain generalization ability, 5) Self-distillation which transfers task-relevant semantics about known classes from the deterministic to the probabilistic stream to

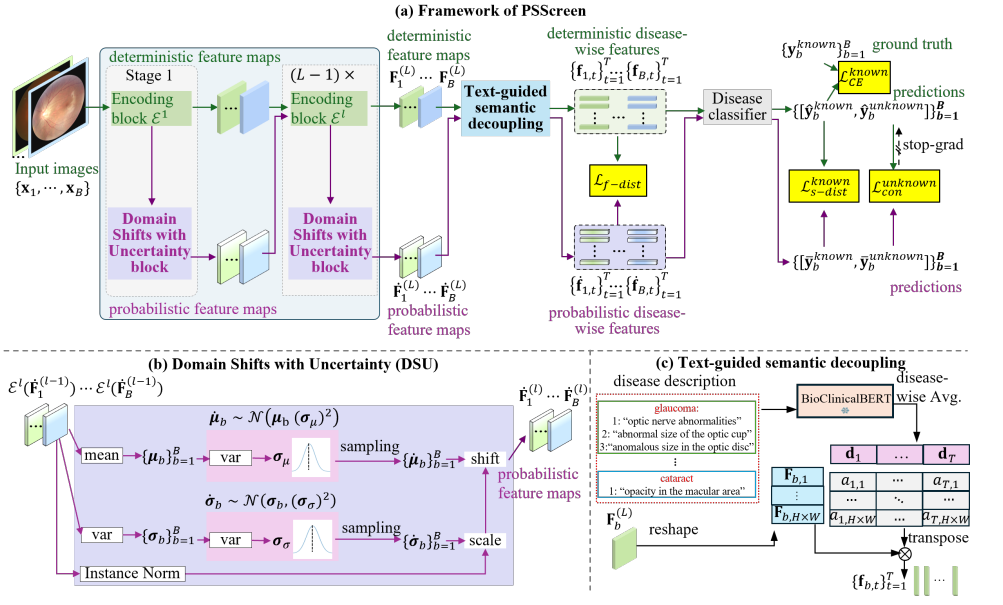


Figure 2: The framework of *PSScreen*. (a) illustrates the training pipeline of PSScreen. With training images, deterministic features and probabilistic features are extracted by the encoding blocks and domain shifts with uncertainty (DSU) blocks, then decoupled by the text-guided semantic decoupling module, finally fed to the disease classifier for multi-disease risk prediction. Feature distillation $\mathcal{L}_{f\text{-}dist}$, self-distillation for known classes $\mathcal{L}_{s\text{-}dist}^{known}$, pseudo label consistency for unknown classes $\mathcal{L}_{con}^{unknown}$, and cross entropy loss for known classes $\mathcal{L}_{CE}^{known}$ are applied for model optimization. (b) and (c) illustrate details for DSU and text-guided semantic decoupling block.

enhance detection performances, and 6) Pseudo label consistency which addresses the label absent issue.

Deterministic Feature Learning. We adopt either a L-stage CNN or vision transformer as the backbone to learn deterministic features from each batch of input images and obtain deterministic feature maps $\{\mathbf{F}_1^{(l)},\cdots,\mathbf{F}_B^{(l)}\}$ where $l\in[1,\cdots,L]$. As deterministic feature maps preserve task-relevant semantics without uncertainty injection, the classification loss for known classes is imposed on predictions derived from them.

Probabilistic Feature Learning via Domain Shifts with Uncertainty (DSU) Block. We estimate the uncertainty via attaching a DSU block [19] into each stage of the backbone and produce probabilistic features to enhance the generalization ability across diverse or even unseen domains. As shown in Fig. 2(b), probabilistic feature maps $\{\dot{\mathbf{F}}_1^{(l-1)},\cdots,\dot{\mathbf{F}}_B^{(l-1)}\}$ from the $(l-1)$-th stage are first fed into the l-th encoding block, yielding $\mathcal{E}^l(\dot{\mathbf{F}}_1^{(l-1)}),\cdots,\mathcal{E}^l(\dot{\mathbf{F}}_B^{(l-1)})$. Then, we calculate channel-wise means $\{\mu_b\}_{b=1}^B$ and variances $\{\sigma_b\}_{b=1}^B$, and model μ_b and σ_b as centers of two independent Gaussian distributions. Their respective scopes, i.e., the uncertainty for the mean σ_μ and uncertainty for the variance σ_σ, are estimated non-parametrically which are variances of $\{\mu_b\}_{b=1}^B$ and $\{\sigma_b\}_{b=1}^B$ respectively. Finally, we randomly sample the mean and variance from these distributions and scale the instance normal-

ized features with the sampled variance and shift them with the sampled mean to obtain the probabilistic feature maps $\{\dot{\mathbf{F}}_1^{(l)}, \cdots, \dot{\mathbf{F}}_B^{(l)}\}$ where $l \in [1, \cdots, L]$.

Text-guided Semantic Decoupling. To focus on disease-specific semantic regions, we introduce a text-guided semantic decoupling module inspired by [1]. As illustrated in Fig 2(c), we use BioClinicalBERT[1] to encode multiple expert knowledge descriptions of each disease from [31] and average them to obtain the disease-wise text embeddings $\{\mathbf{d}_t\}_{t=1}^{T}$. For the visual features $\mathbf{F}_b^{(L)}$ with spatial size $H \times W$, we first reshape them into $\{\mathbf{F}_{b,i}\}_{i=1}^{H\times W}$, then calculate the attention score $\alpha_{t,i}$ between each visual feature $\mathbf{F}_{b,i}$ and text feature $\mathbf{d}_t$ via:

$$\alpha_{t,i} = \frac{\exp\left(\mathbf{v}_{att}^{\top} \tanh\left(\mathbf{W}_{att}^{F}\mathbf{F}_{b,i} \odot \mathbf{W}_{att}^{d}\mathbf{d}_t\right)\right)}{\sum_{j=1}^{H\times W} \exp\left(\left(\mathbf{v}_{att}^{\top} \tanh\left(\mathbf{W}_{att}^{F}\mathbf{F}_{b,j} \odot \mathbf{W}_{att}^{d}\mathbf{d}_t\right)\right)\right)}, \tag{1}$$

where $\odot$ denotes the Hadamard product, and $\mathbf{W}_{att}^{F}$, $\mathbf{W}_{att}^{d}$, and $\mathbf{v}_{att}$ are learnable weights. Finally, we obtain the text-guided disease-wise features via:

$$\mathbf{f}_{b,t} = \sum_{i=1}^{H\times W} \alpha_{t,i} \cdot \mathbf{F}_{b,i}\,. \tag{2}$$

Feature Distillation. To align feature distributions, we minimize the maximum mean discrepancies (MMD) loss [23] which measures the discrepancy between two distributions by comparing their means in a reproducing kernel Hilbert space $\mathcal{H}$. Specifically, we minimize the MMD loss between the text-guided semantic decoupled features from the two-streams:

$$\mathcal{L}_{f\text{-}dist}(\mathbf{f}_{b,t}, \dot{\mathbf{f}}_{b,t}) = \frac{1}{T}\sum_{t=1}^{T} \left\|\phi(\mathbf{f}_{b,t}) - \phi(\dot{\mathbf{f}}_{b,t})\right\|_{\mathcal{H}}^{2}, \tag{3}$$

where $\phi(\cdot)$ denotes the kernel mapping function. In practice, we use a Gaussian kernel to compute the MMD loss. This class-specific alignment ensures that probabilistic disease-wise features closely align with their deterministic counterparts while preserving essential class-specific discriminative information. By aligning the two-stream features, the learned features help the model boost the domain generalization ability.

Classification Loss for Known Classes. We follow [7] and use the partial binary cross entropy loss as the classification loss for known classes:

$$\mathcal{L}_{CE}^{known}(\mathbf{y}_b^{known}, \hat{\mathbf{y}}_b^{known}) = -\frac{1}{\|\delta_b\|_1}\sum_{t=1}^{T}\left(\mathbb{1}_{\{1\}}(y_{b,t})\log(\hat{y}_{b,t}) + \mathbb{1}_{\{-1\}}(y_{b,t})\log(1-\hat{y}_{b,t})\right), \tag{4}$$

where $\hat{y}_{b,t}$ is the prediction corresponding to the deterministic disease-wise feature for t-th disease of the b-th image in current batch, and $y_{b,t}$ is the ground-truth label.

Self-distillation. To supervise the probabilistic feature learning, we introduce self-distillation to transfer the task-relevant semantics from the deterministic to the probabilistic stream via aligning the two-stream network's output distributions of the known classes. To this end, we minimize the KL divergence loss between the final classification predictions of the two streams for known classes:

$$\mathcal{L}_{s\text{-}dist}^{known}\left(\hat{\mathbf{y}}_b^{known}, \bar{\mathbf{y}}_b^{known}\right) = KL\left(\hat{\mathbf{y}}_b^{known} || \bar{\mathbf{y}}_b^{known}\right) = -\frac{1}{||\delta_b||_1}\sum_{t=1}^{T}\delta_{b,t}\cdot\hat{y}_{b,t}\cdot\log\frac{\bar{y}_{b,t}}{\hat{y}_{b,t}}, \tag{5}$$

[1] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

where $\bar{y}_{b,t}$ is the prediction corresponding to the probabilistic disease-wise feature for t-th disease of the b-th image in current batch.

Pseudo Label Consistency. We further enforce the consistency between the pseudo labels derived from the probabilistic features for unknown classes and those derived from deterministic features. In detail, we adopt confidence-based "hard" pseudo labels and samples with predictions greater than a threshold τ are treated as positive and those below $1-\tau$ are treated as negative. With them, the pseudo label consistency loss can be expressed as:

$$\mathcal{L}_{con}^{unknown}(\hat{\mathbf{y}}_b^{unknown}, \bar{\mathbf{y}}_b^{unknown}) = -\frac{1}{T-||\delta_b||_1}\sum_{t=1}^{T}(1-\delta_{b,t})\cdot\left(\mathbb{1}_{(\hat{y}_{b,t}>\tau)}\log(\bar{y}_{b,t}) + \mathbb{1}_{(\hat{y}_{b,t}<1-\tau)}\log(1-\bar{y}_{b,t})\right). \quad (6)$$

To ensure that the generated pseudo labels for unknown classes are sufficiently accurate, τ is set as 0.95 in our experiment.

Total Loss. The total loss function is :

$$\mathcal{L} = \mathcal{L}_{CE}^{known} + \lambda_1\mathcal{L}_{f\text{-}dist} + \lambda_2\mathcal{L}_{s\text{-}dist}^{known} + \lambda_3\mathcal{L}_{con}^{unknown}, \quad (7)$$

where λ_1 and λ_2 are set to 0.05 and 1 respectively, in order to ensure that the different loss components are on a comparable scale. λ_3 is used to control the contribution of the pseudo label loss during training. It is set to 0 for the first 5 epochs and updated to 0.6 starting from the 6-th epoch.

3 Experiment

3.1 Experimental Settings

Datasets. We construct two combined datasets using multiple partially labeled open-access datasets: (1) **meta-dataset**, constituting of six datasets, i.e., DDR [18], ADAM [8], PALM [10], Kaggle-CAT[2], Kaggle-HR[3] and REFUGE2 [9], and (2) **unseen-dataset** constituting of four datasets i.e., APTOS2019 [4], ORIGA^{light} [38], HPMI [13], and RFMiD [25]. The meta-dataset covers all diseases of interest, including diabetic retinopathy (DR), glaucoma, cataract, age-related macular degeneration (AMD), hypertensive retinopathy (HR), pathologic myopia (PM), and the normal state. It is utilized for both training and in-domain validation. The unseen-dataset is used for out-of-domain validation. Besides, the test set of **ODIR dataset** [17] is also used for out-of-domain validation. To further evaluate the model's domain generalization ability, we follow [31] and use **ODIR200×3** to validate the performances under the setting of zero-shot inference. ODIR200×3 is a 600-image subset containing three diseases, i.e., the normal state, cataract, and pathologic myopia and with 200 images in each category. More details about the meta-dataset, unseen-dataset and ODIR200x3 can be found in Supplementary A.

Evaluation Metrics. F-score and quadratic weighted kappa (QWK) are adopted following [17, 37]. For evaluation across multiple tasks and datasets, we compute the average

[2] https://www.kaggle.com/datasets/jr2ngb/cataractdataset

[3] https://www.kaggle.com/datasets/harshwardhanfartale/hypertension-and-hypertensive-retinopathy-dataset

[4] https://www.kaggle.com/competitions/aptos2019-blindness-detection/data

Methods	Source	Meta		Unseen		ODIR		FPS	GFLOPs	#param (M)
		mF	*mQWK*	*mF*	*mQWK*	*mF*	*mQWK*			
Full supervise	–	–	–	–	–	69.8±0.2	52.6±0.7	–	–	–
MultiNets	–	82.7±0.7	73.0±1.4	62.4±1.2	44.8±2.3	56.4±1.4	26.1±2.1	47.7	5240	837.4
MultiHeads	–	83.6±0.7	75.0±1.3	62.4±1.5	44.2±2.0	56.7±0.6	27.7±0.4	382.6	650	105.3
SST [2]	AAAI22	83.4±1.4	74.7±2.8	63.2±0.9	46.7±1.8	55.6±1.7	26.2±2.5	357.9	670	122.9
SARB [26]	AAAI22	84.0±0.9	75.7±1.8	60.5±1.0	40.3±1.2	61.4±0.9	33.8±1.6	361.6	670	112.7
BoostLU [15]	CVPR23	80.3±0.5	69.5±1.1	44.5±0.8	18.0±0.6	50.6±1.0	17.5±2.1	140.5	650	162.2
HST [3]	IJCV24	83.4±0.4	74.7±0.8	62.3±1.7	44.8±3.8	56.0±0.8	27.5±0.8	357.9	670	123.0
CALDNR [27]	TMM24	82.0±1.1	72.8±2.2	47.6±3.1	19.0±6.7	52.1±0.6	19.1±0.8	138.4	660	120.9
PSScreen (ours)	–	**84.2**±0.3	**76.8**±0.8	**65.9**±0.1	**50.9**±0.1	**64.1**±1.0	**39.8**±1.3	373.1	660	116.8

Table 1: Performance comparison of different partially supervised learning methods on three datasets. The best and second-best are highlighted in bold and with an underline. Means and standard deviations are reported over three trails. Results for each disease can be found in Supplementary B.

F-score as:

$$mF = \sum_{t=1}^{T} \sum_{k=1}^{K^{(t)}} \frac{1}{T} \frac{1}{K^{(t)}} F_k^{(t)}, \tag{8}$$

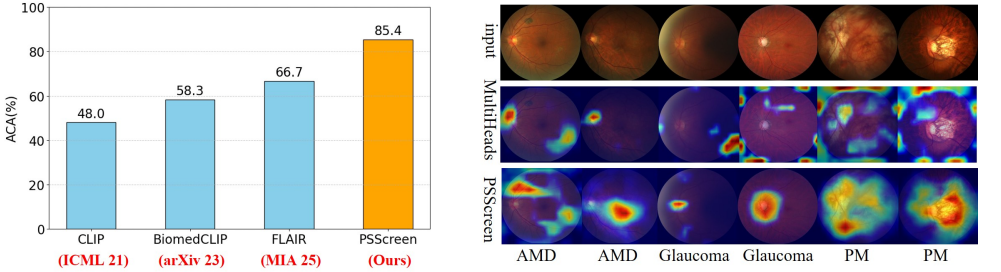
where T is the number of tasks and $K^{(t)}$ is the number of datasets per task. Mean QWK ($mQWK$) is calculated similarly.

Implementation Details. We crop the field of view from each fundus image, then pad the short side with zeros to equal length with the long side and resize it to 512×512. For augmentation, we apply random scaling with a scale factor uniformly sampled from [0.8,1.2] with a probability of 0.5, followed by padding or cropping to maintain the input size. We then apply the augmentation strategies from [29], excluding Cutout. Following [2, 3, 26, 27], we adopt ResNet-101 [11] pretrained on ImageNet [5] as the backbone, with other model parameters initialized randomly. Training uses the ADAM optimizer [16] with a batch size of 16, weight decay of 5×10^{-4}, and an initial learning rate of 1×10^{-5} reduced by a factor of 10 every 10 epochs. PSScreen is trained with 20 epochs in total, and is implemented with PyTorch on one NVIDIA A100 GPU with 40 GB RAM.

3.2 Results

Comparison with Partially Supervised Learning Methods. We compare our PSScreen with two baseline approaches, MultiNets and MultiHeads, as well as five state-of-the-art (SOTA) methods: SST [2], SARB [26], HST [3], BoostLU [15], and CALDNR [27]. MultiNets trains multiple task-specific models for each task and combines all prediction results during testing. MultiHeads consists of a backbone network and a classifier, which outputs prediction probabilities for all classes.

We report performances in Table 1, and observe that on the meta-dataset (1) our PSScreen achieves the best, (2) our PSScreen outperforms the second-best SARB [26] by 0.2% in mF and 1.1% in $mQWK$. Notably, results on two out-of-domain datasets, i.e., unseen and ODIR [17] show that PSScreen gains significant improvements compared to other methods. In detail, on the unseen-dataset, our PSScreen outperforms the second-best SST [2] by 2.7% in mF and 4.2% in $mQWK$. On the ODIR [17], PSScreen outperforms the second-best SARB [26] by 2.7% in mF and 6.0% in $mQWK$. For reference, we also report the performances on the ODIR [17] by the fully supervised model as an upper bound. SOTAs in Table 1 perform poorly on the two out-of-domain datasets as they assume that training and



(a) Comparison with foundation models (b) Visualization of heatmaps

Figure 3: (a) Performance comparison of zero-shot inference with foundation models on the ODIR200x3 dataset. (b) Visualization of heatmaps generated by MultiHeads and PSScreen for three retinal diseases: age-related macular degeneration (AMD), glaucoma, and pathologic myopia (PM).

test images follow the same distribution. Additionally, we compare model efficiency and report Frames Per Second (FPS), Giga Floating Point Operations (GFLOPs), and number of parameters (#param). It can be observed that the FPS of PSScreen is only slightly lower than that of MultiHeads, indicating comparable inference efficiency. Meanwhile, compared to SOTAs, PSScreen significantly improves prediction accuracy without introducing additional computational or parameter overhead.

Comparison of Zero-shot Inference with Foundation Models. We further compare the zero-shot performances of our PSScreen and three latest foundation models, i.e., CLIP [28], BiomedCLIP [36], and FLAIR [31] on ODIR200×3. Following FLAIR, we adopt ACA [39] as the evaluation metric and report ACAs in Fig. 3(a). As shown, foundation models perform suboptimally in zero-shot retinal disease screening. CLIP [28] and BiomedCLIP [36] struggle to detect retinal diseases, while FLAIR [31] fails to generalize well to unseen diseases due to the complex pathological structures of retinal diseases. Our PSScreen achieves the best performance on ODIR200×3, outperforming FLAIR by 18.7%.

Visualization. We use GradCAM [30] to obtain heatmaps from the last convolutional layer of MultiHeads and PSScreen. As shown in Fig. 3(b), MultiHeads often fails to localize lesions, instead attending to domain-specific background. In contrast, our PSScreen consistently highlights lesion areas. For example, as shown in the first column, PSScreen accurately localizes AMD-related hemorrhage; in the second column, it identifies exudates critical for AMD diagnosis. For glaucoma, PSScreen consistently attends to the optic disc and cup, and for PM, PSScreen localizes retinal atrophy regions strongly correlated with the disease.

3.3 Ablation Study

How $\mathcal{L}_{f\text{-}dist}$, $\mathcal{L}_{s\text{-}dist}^{known}$ and $\mathcal{L}_{con}^{unknown}$ Contribute? There are three key loss terms, i.e., $\mathcal{L}_{f\text{-}dist}$ in Eq. 3, $\mathcal{L}_{s\text{-}dist}^{known}$ in Eq. 5 and $\mathcal{L}_{con}^{unknown}$ in Eq. 6. To validate their effectiveness, we conduct ablation experiments and report the performance in Table 2. Removing the three loss items, PSScreen degrades into MultiHeads equiped with the text-guided semantic decoupling, which achieves inferior performance to PSScreen obviously. With $\mathcal{L}_{f\text{-}dist}$ and $\mathcal{L}_{s\text{-}dist}^{known}$

$\mathcal{L}_{f\text{-}dist}$	$\mathcal{L}_{s\text{-}dist}^{known}$	$\mathcal{L}_{con}^{unknown}$	Meta		Unseen		ODIR	
			mF	*mQWK*	*mF*	*mQWK*	*mF*	*mQWK*
			83.5±0.2	75.0±0.4	63.0±0.8	45.4±1.3	57.0±0.6	28.0±1.0
✓			83.4±0.7	75.3±0.2	63.4±0.1	46.7±0.9	58.0±0.4	29.6±0.5
	✓		83.4±0.2	75.8±0.4	64.0±0.3	47.5±0.7	57.5±1.0	28.7±1.5
		✓	83.3±0.7	75.5±1.1	61.8±1.1	44.1±2.1	61.7±0.9	34.8±1.6
	✓	✓	83.8±0.2	76.4±0.6	65.2±0.3	49.5±0.7	63.5±0.8	38.0±1.5
✓	✓	✓	**84.2**±0.3	**76.8**±0.8	**65.9**±0.1	**50.9**±0.1	**64.1**±1.0	**39.8**±1.3

Table 2: The ablation study on the key loss terms of PSScreen.

λ_1	λ_2	λ_3	Meta	Unseen	ODIR
0.1	1.0	0.6	76.1±0.6	50.1±0.2	39.4±0.7
0.05	**1.0**	**0.6**	**76.8**±0.8	**50.9**±0.1	**39.8**±1.3
0.025	1.0	0.6	76.7±0.1	49.4±0.6	39.7±1.0
0.05	0.5	0.6	76.2±0.3	50.0±0.6	38.6±0.9
0.05	**1.0**	**0.6**	**76.8**±0.8	**50.9**±0.1	**39.8**±1.3
0.05	2.0	0.6	75.9±0.4	48.2±1.3	38.7±0.5
0.05	1.0	0.4	76.5±1.3	48.7±1.2	38.9±1.0
0.05	**1.0**	**0.6**	**76.8**±0.8	**50.9**±0.1	**39.8**±1.3
0.05	1.0	0.8	76.6±0.5	50.5±1.4	38.8±0.4

(a) *mQWK* under different loss weight settings.

Methods	Meta	Unseen	ODIR
ResNet-101 [11]	75.0±1.3	44.2±2.0	27.7±0.4
+PSScreen	**76.8**±0.8	**50.9**±0.1	**39.8**±1.3
ConvNeXt-T [22]	75.0±0.7	48.2±1.2	25.4±2.3
+PSScreen	**76.6**±0.2	**49.6**±2.1	**37.0**±1.5
ConvNeXt V2-T [33]	76.3±0.7	48.6±1.2	28.9±1.3
+PSScreen	**77.0**±0.2	**49.1**±0.5	**38.3**±1.0
Swin-T [21]	75.8±0.8	49.5±0.7	25.4±0.3
+PSScreen	**79.3**±0.7	**50.8**±0.6	**35.5**±0.9
VMamba-T [20]	76.3±1.0	47.2±0.5	27.0±0.4
+PSScreen	**78.0**±0.7	**49.8**±1.7	**36.9**±1.8

(b) Performances with different backbones.

Table 3: (a) *mQWK* under different settings of λ_1, λ_2, and λ_3. (b) *mQWK* by MultiHeads and PSScreen with different backbones. Results in terms of *mF* can be found in Supplementary C.1.

separately, in-domain performances and domain generalization are improved. On the contrary, $\mathcal{L}_{con}^{unknown}$ alone leads to the performance degradation which is possibly caused by the loss of task-relevant semantics in probabilistic feature maps. Combining $\mathcal{L}_{s\text{-}dist}^{known}$ and $\mathcal{L}_{con}^{unknown}$ improves the model's performance. Integrating all loss terms achieves the best results, with *mQWK* improvements of 5.5% and 11.8% on the unseen and ODIR datasets, respectively, compared to the degraded variant without the three loss terms, validating the effectiveness of PSScreen.

Influences of Loss Weights λ_1, λ_2, and λ_3. We vary λ_1, λ_2, and λ_3, and report *mQWK* in Table 3(a), which shows that $\lambda_1 = 0.05$, $\lambda_2 = 1.0$, and $\lambda_3 = 0.6$ perform the best performances. The performances on *mF* can be found in Supplementary C.1.

Compatibility with Different Backbones. To further validate the compatibility of our PSScreen, we report *mQWK* of PSScreen with different backbones including ConvNeXt-T [22], ConvNeXt V2-T [33], Swin-T [21], and VMamba-T [20] and compare them with the naive ones, i.e., MultiHeads with various backbones in Table 3(b). The results demonstrate that our PSScreen consistently improves the performances across all three datasets with different backbones. The performances on *mF* can be found in Supplementary C.2.

4 Conclusion and Future Work

In this paper, we propose *PSScreen*, a multiple retinal disease screening model trained on multiple partially labeled datasets following different distributions. PSScreen is a two-stream network and one streams learns deterministic features and the other learns probabilistic features via DSU blocks. The features are then decoupled by the text-guided semantic decoupling modules to facilitate multiple disease detection. To enforce the probabilistic feature learning stream to learn task-relevant semantics, feature distillation and self-distillation

are employed to transfer task-relevant semantics from the deterministic to the probabilistic stream. Meanwhile, pseudo label consistency is imposed to address label absent issue. Extensive experiments on both fully and partially labeled datasets demonstrate that PSScreen achieves state-of-the-art performance across multiple in-domain and out-of-domain datasets.

PSScreen has a lightweight architecture and fast inference speed, making it well-suited for integration into clinical workflows. However, challenges remain in translating screening models from bench to bedside. For example, ophthalmic imaging modalities are diverse, whereas our current focus is limited to fundus images. To enhance clinical applicability, incorporating additional modalities for screening a broader range of retinal diseases is desirable in the future. Even though GradCAM-based visualizations for model's interpretability are provided, a structured evaluation involving clinicians is lacking. Moreover, how to conduct clinician-in-the-loop assessments to verify the model's interpretability remains still an open question and merits further investigation.

5 Acknowledgments

This work was supported in part by Academy Research Fellow project under Grant 355095 and Hunan Provincial Natural Science Foundation of China under Grant 2023JJ30699. As well, the authors wish to acknowledge the CSC–IT Center for Science, Finland, for computational resources.

References

[1] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.

[2] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, and Liang Lin. Structured semantic transfer for multi-label recognition with partial labels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):339–346, June 2022.

[3] Tianshui Chen, Tao Pu, Lingbo Liu, Yukai Shi, Zhijing Yang, and Liang Lin. Heterogeneous semantic transfer for multi-label recognition with partial labels. *International Journal of Computer Vision*, 132(12):6091–6106, July 2024.

[4] Dipankar Das, Deepak Ranjan Nayak, and Ram Bilas Pachori. CA-Net: A novel cascaded attention-based network for multistage glaucoma classification using fundus images. *IEEE Transactions on Instrumentation and Measurement*, 72:1–10, 2023.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Jiawei Du, Jia Guo, Weihang Zhang, Shengzhu Yang, Hanruo Liu, Huiqi Li, and Ningli Wang. RET-CLIP: A retinal image foundation model pre-trained with clinical diagnostic reports. In *Medical Image Computing and Computer Assisted Intervention*, pages 709–719, 2024.

[7] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep ConvNet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.

[8] Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, et al. ADAM Challenge: detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging*, 41(10):2828–2847, 2022.

[9] Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, et al. REFUGE2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening. *arXiv preprint arXiv:2202.08994*, 2022.

[10] Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. Open fundus photograph dataset with pathologic myopia recognition and anatomical structure annotation. *Scientific Data*, 11(1): 99, 2024.

[11] 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*, June 2016.

[12] Yichen Hu, Chao Wang, Weitao Song, Aleksei Tiulpin, and Qing Liu. A scanning laser ophthalmoscopy image database and trustworthy retinal disease detection method. In *Medical Image Computing and Computer Assisted Intervention*, pages 46–56, 2024.

[13] Shenghai Huang, Zhongwen Li, Bing Lin, Shaodan Zhang, Quanyong Yi, and Lei Wang. HPMI: A retinal fundus image dataset for identification of high and pathological myopia based on deep learning. 12 2023.

[14] Lie Ju, Zhen Yu, Lin Wang, Xin Zhao, Xin Wang, Paul Bonnington, and Zongyuan Ge. Hierarchical knowledge guided learning for real-world retinal disease recognition. *IEEE Transactions on Medical Imaging*, 43(1):335–350, 2024.

[15] Youngwook Kim, Jae Myung Kim, Jieun Jeong, Cordelia Schmid, Zeynep Akata, and Jungwoo Lee. Bridging the gap between model explanations in partially annotated multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3408–3417, June 2023.

[16] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, May 2015.

[17] Ning Li, Tao Li, Chunyu Hu, Kai Wang, and Hong Kang. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In *Benchmarking, Measuring, and Optimizing*, pages 177–193, 2021.

[18] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019.

[19] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022.

[20] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision Transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, October 2021.

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, June 2022.

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 97–105, July 2015.

[24] Sandeep Madarapu, Samit Ari, and Kamalakanta Mahapatra. C2x-FNet: Cascaded dense block with twofold cross-feature enhancement module for diabetic retinopathy grading. *IEEE Transactions on Instrumentation and Measurement*, 74:1–10, 2025.

[25] Samiksha Pachade, Prasanna Porwal, Dhanshree Thulkar, et al. Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research. *Data*, 6(2), 2021.

[26] Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. Semantic-aware representation blending for multi-label image recognition with partial labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2091–2098, 2022.

[27] Tao Pu, Qianru Lao, Hefeng Wu, Tianshui Chen, Ling Tian, Jie Liu, and Liang Lin. Category-adaptive label discovery and noise rejection for multi-label recognition with partial positive labels. *IEEE Transactions on Multimedia*, 26:9591–9602, 2024.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 13, pages 8748–8763, 2021.

[29] Manuel Alejandro Rodríguez, Hasan AlMarzouqi, and Panos Liatsis. Multi-label retinal disease classification using Transformers. *IEEE Journal of Biomedical and Health Informatics*, 27(6):2739–2750, 2023.

[30] 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 International Conference on Computer Vision*, Oct 2017.

[31] Julio Silva-Rodríguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A Foundation Language-Image Model of the Retina (FLAIR): encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025.

[32] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware Transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021.

[33] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, June 2023.

[34] Jiayin Xiao, Si Li, Tongxu Lin, Jian Zhu, Xiaochen Yuan, David Dagan Feng, and Bin Sheng. Multi-label Chest X-ray image classification with single positive labels. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024.

[35] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, 91:102996, 2024.

[36] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[37] Yuhan Zhang, Xiao Ma, Kun Huang, Mingchao Li, and Pheng-Ann Heng. Semantic-oriented visual prompt learning for diabetic retinopathy grading on fundus images. *IEEE Transactions on Medical Imaging*, 43(8):2960–2969, 2024.

[38] Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. ORIGA$^{\text{light}}$: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 3065–3068, August 2010.

[39] Ziyuan Zhao, Kerui Zhang, Xuejie Hao, Jing Tian, Matthew Chin Heng Chua, Li Chen, and Xin Xu. BiRA-Net: Bilinear attention net for diabetic retinopathy grading. In *2019 IEEE International Conference on Image Processing*, pages 1385–1389, 2019.