PSScreen：部分监督的多种视网膜疾病筛查

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

利用多个部分标注的数据集来训练一个用于多种视网膜疾病筛查的模型可以减少对完全标注数据集的依赖，但由于来自不同医疗机构的训练数据集之间的显著领域转移和部分类别的标签缺失问题，这仍然具有挑战性。为了解决这些挑战，我们提出了 PSScreen，一种新颖的部分监督多种视网膜疾病筛查模型。我们的 PSScreen 由两条流组成，一条学习确定性特征，另一条通过不确定性注入学习概率特征。然后，我们利用文本引导将两种特征解耦为疾病特异性特征，并通过特征蒸馏对齐它们，以提升领域泛化能力。同时，我们在两个流之间引入伪标签一致性以解决标签缺失问题，并引入自蒸馏以从确定性流向概率流传输有关已知类别的任务相关语义，从而进一步增强检测性能。实验表明，我们的 PSScreen 显著提升了对六种视网膜疾病和正常状态的检测性能，并在域内和域外数据集上均取得了最先进的结果。代码可在 https://github.com/boyiZheng99/PSScreen 获取。

1 引言

利用眼底图像进行视网膜疾病的自动检测对于高效且具有成本效益的大规模人群筛查至关重要。通过开源数据集的发布（如图 1 (a) 所示），这些数据集针对特定的视网膜疾病，如用于糖尿病视网膜病变筛查的 DDR [18]，用于青光眼筛查的 REFUGE2 [9]，以及用于近视筛查的 PALM [10] 等，驱动了许多工作 [4, 24, 32, 37] 的发展，这些工作通过单独的训练数据集来训练特定疾病的筛查模型。尽管这些模型在特定领域的图像上进行特定疾病的筛查方面表现出色，但在实际应用中，人们期望能够在来自各种甚至未见过的领域的图像上筛查尽可能多的视网膜疾病，这方面这些模型仍有不足。开发一种具有强大域泛化能力的多种视网膜疾病筛查模型具有重要意义，但仍然充满挑战。

最直观的方法是用完全标记的训练数据以完全监督的方式训练疾病筛查模型，如图 1 (b) 所示。例如，TrustDetector [12] 为三种视网膜疾病训练筛查模型，因为收集的数据集只提供了三种疾病的标签，而 [17] 中的方法是在 ODIR 上训练用于多种视网膜疾病筛查的模型。尽管它们取得了令人满意的表现，但训练数据的规模有限，并且假设训练和测试图像共享相同的分布，这限制了其对域外数据的泛化能力。第二种方法是收集并人工注释一个大规模数据集，例如 Retina-1M [14]，以实现完全监督的训练。然而，注释过程劳心费力且成本高昂。最近，随着基础模型如 FLAIR [31] 和 RET-CLIP [6] 的适应性增强，这些模型通过大规模图像-文本对的自监督学

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

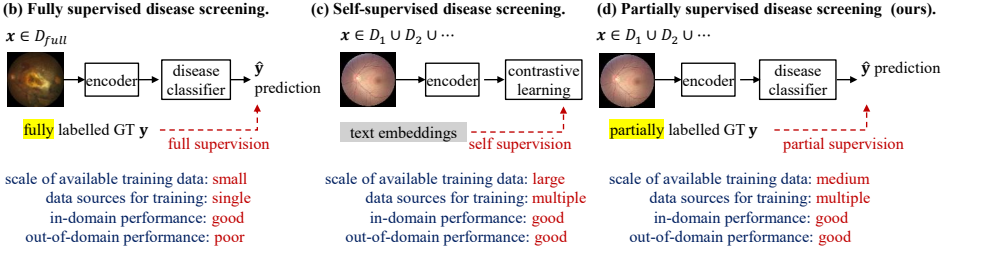

Figure 1: 用于视网膜疾病筛查的开源数据集示例以及在三种学习范式下的筛查模型比较。(a) 列出了开源数据集，其中"✓"表示疾病的标签是可用的，而"?"表示标签不可用。从 (b) 到 (d)，我们展示了完全监督筛查模型的流程和特征，该模型通常使用全标记数据集进行训练，使用包含图像-文本对的多个数据集进行训练的自监督筛查模型，以及使用多个部分标记数据集进行训练的部分监督筛查模型。

习方式进行训练，如图 1 (c) 所示，用于视网膜疾病筛查的零样本学习正在兴起。虽然不再需要大规模的完全注释训练数据，但对于特定疾病筛选的表现较差 [35] 。

在本文中，我们提出了 PSScreen，一个部分监督的多种视网膜疾病筛查模型。如图 1 （d）所示，我们的 PSScreen 在多个部分标记的数据集上训练筛查模型，而这些数据集遵循不同的数据分布。与以前针对自然图像 [2, 3, 26, 27] 和医学图像 [34] 的部分监督学习方法不同，这些方法的训练和测试数据都来自没有领域转移的同一数据集，本论文中的部分监督学习具有更大的挑战性，因为（1）来自不同医疗场所的训练数据集中存在领域转移，（2）部分类别的标签缺失问题。为了解决这些问题，我们提出了一个双流网络：一个学习确定性特征，另一个通过不确定性注入学习概率特征。然后，在文本信息的指导下，我们将两种类型的特征解耦为按疾病划分的特征，并通过特征蒸馏进行对齐，以确保所学特征对领域转移具有鲁棒性。同时，PSScreen 通过两条路径之间的伪标签一致性解决了标签缺失问题。最后，我们引入自我蒸馏，将关于已知类别的任务相关语义从确定性流转移到概率流，以进一步增强检测性能。实验表明，PSScreen 在领域内和领域外数据集上均优于最新的方法。

总结起来，我们的贡献如下：

- 我们提出了 PSScreen，一种多种视网膜疾病筛查方法，该方法在由多个不同分布的部分标注数据集组成的元数据集上训练疾病筛查模型。据我们所知，我们是首次使用来自不同医疗机构的多个数据集训练部分监督模型以筛查多种视网膜疾病。

- 我们提出了一种双流网络，该网络学习确定性特征和注入不确定性的概率特征，并对它们进行对齐，以提高特征层面的域泛化能力。同时，我们在两个流之间引入伪标签一致性，以解决标签缺失问题，并通过自蒸馏将关于已知类别的任

Figure 2: PSScreen 的框架。(a) 展示了 PSScreen 的训练流程。通过训练图像，确定性特征和概率性特征由编码块和具有不确定性的域转移 (DSU) 块提取，然后由文本引导的语义解耦模块解耦，最后输入疾病分类器进行多疾病风险预测。为了优化模型，应用了特征蒸馏 $\mathcal{L}_{f\text{-}dist}$ 、已知类别的自蒸馏 $\mathcal{L}_{s\text{-}dist}^{known}$ 、未知类别的伪标签一致性 $\mathcal{L}_{con}^{unknown}$ ，以及已知类别的交叉熵损失 $\mathcal{L}_{CE}^{known}$ 。(b) 和 (c) 展示了 DSU 和文本引导语义解耦块的细节。

务相关语义从确定性流传递到概率流，从而进一步增强检测性能。

- 我们在元数据集上验证了我们的 PSScreen，并证明我们的 PSScreen 达到了最先进的性能。更重要的是，在六个未见数据集上的验证进一步表明，PSScreen 比以前的方法更具优越的领域泛化能力。

2 提出的方法

问题表述。我们假设：(1) 存在一个元数据集 $\mathcal{D}=\{D_1,D_2,\ldots,D_K\}$ ，由 K 个部分标记的数据集组成，这些数据集是从各个医疗站点收集的，并遵循不同的分布；(2) 每个数据集 $D_k=\{(x_i,y_i)\}_{i=1}^{N_k}$ 有 N_k 个样本，其中 x_i 是第 i 个样本，$y_i\in\{1,0,-1\}^T$ 是表示 T 种眼底疾病的标签，而 $y_{i,t}=0$ 表示第 t 种疾病的标签未知，而 $y_{i,t}=1/-1$ 表示第 t 种疾病标签为阳性/阴性。为简化起见，我们使用向量 $\delta_i=\mathbb{1}_{\{1,-1\}}(y_i)$ 来表示疾病的标签是已知或未知，其中 $\mathbb{1}(\cdot)$ 是一个指示函数。我们的目标是在 $\mathcal{D}$ 上训练一个多标签疾病筛查模型，该模型能够预测 T 种疾病的风险，并且能够很好地泛化到域外测试数据。

概述。图 2 展示了 PSScreen 的概览。PSScreen 是一个由六个关键模块组成的双流网络：1）确定性特征学习，输出确定性特征图以保持任务相关语义，2）通过在主干的每个阶段附加不确定性域转移（DSU）块 [19] 的概率特征学习，生成概率特征图，3）文本引导的语义解耦，将来自两个流的全局特征图分别解耦为疾病特征，4）特征蒸馏，在潜在空间中对齐确定性和概率性疾病特征以提高域泛化能力，5）自蒸馏，将关于已知类别的任务相关语义从确定性流传输到概率性流以增强检测性能，6）伪标签一致性，解决标签缺失问题。

确定性特征学习。我们采用 L 阶段的 CNN 或视觉转换器作为骨干，从每批输入图像中学习确定性特征，并获得确定性特征图 $\{F_1^{(l)},\cdots,F_B^{(l)}\}$ ，其中 $l\in[1,\cdots,L]$ 。由于确定性特征图保留了与任务相关的语义而不引入不确定性，因此对已知类别的分类损失施加在从中导出的预测上。

通过不确定性域转移 (DSU) 模块进行概率特征学习。我们通过在每个骨干网阶段附加一个 DSU 模块 [19] 来估计不确定性，并生成概率特征以增强在不同甚至未见过的领域中的泛化能力。如图 2 (b) 所示，从第 $(l-1)$ 阶段的概率特征图 $\{\dot{F}_1^{(l-1)},\cdots,\dot{F}_B^{(l-1)}\}$ 首先被输入到第 l 编码模块中，生成 $\boldsymbol{\varepsilon}^l(\dot{F}_1^{(l-1)}),\cdots,\boldsymbol{\varepsilon}^l(\dot{F}_B^{(l-1)})$ 。然后，我们计算通道均值 $\{\mu_b\}_{b=1}^B$ 和方差 $\{\sigma_b\}_{b=1}^B$ ，并将 μ_b 和 σ_b 建模为两个独立的高斯分布的中心。它们各自的范围，即均值的不确定性 σ_μ 和方差的不确定性 σ_σ ，分别无参数估计为 $\{\mu_b\}_{b=1}^B$ 和 $\{\sigma_b\}_{b=1}^B$ 的方差。最后，我们从这些分布中随机采样均值和方差，并使用采样得到的方差缩放实例归一化特征，再使用采样得到的均值对其进行平移，以获得概率特征图 $\{\dot{F}_1^{(l)},\cdots,\dot{F}_B^{(l)}\}$ ，其中 $l\in[1,\cdots,L]$ 。

文本引导的语义解耦。为了专注于疾病特定的语义区域，我们引入了一种受到 [1]

启发的文本引导语义解耦模块。如图 2 (c) 所示，我们使用 BioClinicalBERT [1] 对每种疾病的多个专家知识描述从 [31] 进行编码，并将其平均以获得疾病级的文本嵌入 $\{d_t\}_{t=1}^{T}$ 。对于空间大小为 $H \times W$ 的视觉特征 $F_b^{(L)}$ ，我们首先将其重塑为 $\{F_{b,i}\}_{i=1}^{H\times W}$ ，然后通过如下方式计算每个视觉特征 $F_{b,i}$ 和文本特征 d_t 之间的注意力得分 $\alpha_{t,i}$ ：

$$\alpha_{t,i} = \frac{\exp\left(v_{att}^{\top} \tanh\left(W_{att}^{F} F_{b,i} \odot W_{att}^{d} d_t\right)\right)}{\sum_{j=1}^{H\times W} \exp\left(\left(v_{att}^{\top} \tanh\left(W_{att}^{F} F_{b,j} \odot W_{att}^{d} d_t\right)\right)\right)}, \tag{1}$$

，其中 $\odot$ 表示哈达玛积，W_{att}^{F} 、W_{att}^{d} 和 v_{att} 是可学习的权重。最后，我们通过以下方式获得文本引导的疾病级特征：

$$f_{b,t} = \sum_{i=1}^{H\times W} \alpha_{t,i} \cdot F_{b,i} \,. \tag{2}$$

特征蒸馏。为了对齐特征分布，我们最小化最大均值差异（MMD）损失 [23] ，该损失通过比较再生核希尔伯特空间 $\mathcal{H}$ 中的均值来衡量两个分布之间的差异。具体来说，我们最小化来自双流的文本引导语义解耦特征之间的 MMD 损失：

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

，其中 $\phi(\cdot)$ 表示核映射函数。在实践中，我们使用高斯核来计算 MMD 损失。这种特定类别的对齐方式确保了概率性疾病特征与其确定性对应物紧密对齐，同时保留了基本的类别特异性判别信息。通过对齐双流特征，学习到的特征可帮助模型提升领域泛化能力。已知类别的分类损失。我们遵循 [7] ，并使用部分二元交叉熵损失作为已知类别的分类损失：

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

，其中 $\hat{y}_{b,t}$ 是对应于当前批次中第 b 个图像的第 t 个疾病的确定性疾病特征的预测，$y_{b,t}$ 是真实标签。

自蒸馏。为了监督概率特征的学习，我们引入自蒸馏，通过对已知类别的双流网络输出分布进行对齐，将任务相关的语义从确定性流转移到概率性流。为此，我们最小化已知类别的两个流最终分类预测之间的 KL 散度损失：

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

这里 $\bar{y}_{b,t}$ 是当前批次中第 b 张图像的第 t 种疾病对应的概率疾病特征的预测。

伪标签一致性。我们进一步加强了从未知类别的概率特征和确定性特征中导出的伪标签之间的一致性。具体来说，我们采用基于置信度的“硬”伪标签，并将预测值大于阈值 τ 的样本视为正样本，而低于 $1-\tau$ 的样本视为负样本。通过它们，伪标签一致性损失可以表示为：

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

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

为了确保为未知类别生成的伪标签足够准确，我们在实验中将 τ 设置为 0.95。

总损失。总损失函数为：

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

，其中 λ_1 和 λ_2 分别设为 0.05 和 1，以确保不同损失组件处于可比规模。λ_3 用于控制训练期间伪标签损失的贡献。在前 5 个周期中设为 0，从第 6 个周期开始更新为 0.6。

3 实验

3.1 实验设置

数据集。我们使用多个部分标注的开放访问数据集构建两个组合数据集：(1) 元数据集，由六个数据集组成，即 DDR [18] 、ADAM [8] 、PALM [10] 、Kaggle-CAT [2] 、Kaggle-HR [3] 和 REFUGE2 [9] ；(2) 未见数据集，由四个数据集组成，即 APTOS2019 [4] 、ORIGA ^{light} [38] 、HPMI [13] 和 RFMiD [25] 。元数据集涵盖所有关注的疾病，包括糖尿病视网膜病变（DR）、青光眼、白内障、年龄相关性黄斑变性（AMD）、高血压视网膜病变（HR）、病理性近视（PM）和正常状态。它用于训练和域内验证。未见数据集用于域外验证。此外，ODIR 数据集 [17] 的测试集也用于域外验证。为了进一步评估模型的领域泛化能力，我们遵循 [31] 并使用 ODIR200 × 3 在零样本推理的设置下验证性能。ODIR200 × 3 是一个包含三个疾病的 600 张图像子集，即正常状态、白内障和病理性近视，每个类别包含 200 张图像。关于元数据集、未见数据集和 ODIR200x3 的更多详细信息可在补充材料 A 中找到。

评估指标。根据 [17, 37] ，采用 F-score 和加权二次卡帕（QWK）。为了在多个任务和数据集上进行评估，我们计算平均 F-score 如下：

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

，其中 T 是任务数量，$K^{(t)}$ 是每个任务的数据集数量。平均 QWK（$mQWK$ ）也是类似计算的。

实现细节。我们从每张眼底图像中裁剪视野，然后用零填充短边以使其长度等于长边，并将其调整为 512×512 的大小。对于增强，我们以 0.5 的概率应用从 [0.8,1.2] 中均匀采样的缩放因子进行随机缩放，然后通过填充或裁剪来保持输入大小。然后，我们应用来自 [29] 的增强策略，排除 Cutout。遵循 [2, 3, 26, 27] ，我们采用在 ImageNet [5] 上预训练的 ResNet-101 [11] 作为骨干，其它模型参数随机初始化。训练使用 ADAM 优化器 [16] ，批次大小为 16，权重衰减为 5×10^{-4} ，初始学习率为 1×10^{-5} ，每 10 个 epoch 减少 10 倍。PSScreen 总共训练 20 个 epoch，并使用 PyTorch 在一台具有 40 GB RAM 的 NVIDIA A100 GPU 上实现。

3.2 结果

与部分监督学习方法的比较。我们将我们的 PSScreen 与两种基线方法 MultiNets 和 MultiHeads 以及五种最新的（SOTA）方法进行比较：SST [2] 、SARB [26] 、HST

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

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

[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: 三个数据集上不同局部监督学习方法的性能比较。最佳和第二佳结果以粗体和下划线标出。均值和标准差是基于三次试验报告的。每种疾病的结果可以在补充材料 B 中找到。

Figure 3: （a）基础模型在 ODIR200x3 数据集上进行零次推理的性能比较。（b）MultiHeads 和 PSScreen 为三种视网膜疾病生成的热图可视化：年龄相关性黄斑变性（AMD）、青光眼和病理性近视（PM）。

[3] 、BoostLU [15] 和 CALDNR [27] 。MultiNets 为每个任务训练多个特定任务的模型，并在测试时结合所有预测结果。MultiHeads 则由一个骨干网络和一个分类器组成，输出所有类别的预测概率。

我们在表格 1 中报告了性能，并观察到在元数据集上 (1) 我们的 PSScreen 达到了最佳表现，(2) 我们的 PSScreen 在 *mF* 中超过了第二好的 SARB [26] 0.2 %，并在 *mQWK* 中超过了 1.1 % 。值得注意的是，在两个域外的数据集上，即未见过的和 ODIR [17] ，结果显示与其他方法相比，PSScreen 取得了显著的提升。具体来说，在未见过的数据集上，PSScreen 在 *mF* 中超过了第二好的 SST [2] 2.7 %，并在 *mQWK* 中超过了 4.2 % 。在 ODIR [17] 中，PSScreen 在 *mF* 中超过了第二好的 SARB [26] 2.7 %，在 *mQWK* 中超过了 6.0 % 。为了参考，我们还报告了 ODIR [17] 上完全监督模型的性能作为上界。表格 1 中的 SOTAs 在这两个域外数据集上表现不佳，因为它们假设训练和测试图像遵循相同的分布。此外，我们比较了模型效率并报告了每秒帧数 (FPS)、千兆浮点运算量 (GFLOPs) 和参数数量 (# param)。可以观察到 PSScreen 的 FPS 仅比 MultiHeads 略低，表明推理效率相当。同时，与 SOTAs 相比，PSScreen 在不增加额外计算或参数开销的情况下显著提高了预测准确性。

零样本推理与基础模型的比较。我们进一步比较了我们的 PSScreen 和最新的三个基础模型，即 CLIP [28] 、BiomedCLIP [36] 和 FLAIR [31] 在 ODIR200 × 3 上的零样本性能。遵循 FLAIR，我们采用 ACA [39] 作为评估指标，并在图 3 (a) 中报告了 ACA 结果。如图所示，基础模型在零样本视网膜疾病筛查中表现不佳。CLIP [28] 和 BiomedCLIP [36] 难以检测视网膜疾病，而 FLAIR [31] 由于视网膜疾病的复杂病理结构难以很好地推广到未见疾病。我们的 PSScreen 在 ODIR200 × 3 上实现了最佳性能，比 FLAIR 高出 18.7 %。

可视化。我们使用 GradCAM [30] 从 MultiHeads 和 PSScreen 的最后一个卷积层获得热图。如图 3 (b) 所示，MultiHeads 往往不能准确定位病变，而是关注特定领域的背景。相反，我们的 PSScreen 始终突出显示病变区域。例如，如第一列所示，PSScreen 准确定位与 AMD 相关的出血；在第二列中，它识别出对于 AMD 诊断至关重要的渗出物。对于青光眼，PSScreen 始终关注视盘和视杯，而对于 PM，PSScreen 强烈定位与疾病高度相关的视网膜萎缩区域。

$\mathcal{L}_{f\text{-}dist}$	$\mathcal{L}_{s\text{-}dist}^{known}$	$\mathcal{L}_{con}^{unknown}$	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: PSScreen 关键损失项的消融研究。

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) 在不同的 λ_1 、λ_2 和 λ_3 设置下的 *mQWK* 。(b) 使用不同骨干的 MultiHeads 和 PSScreen 得到的 *mQWK* 。关于 *mF* 的结果可以在附录 C .1 中找到。

3.3 消融研究

$\mathcal{L}_{f\text{-}dist}$ 、$\mathcal{L}_{s\text{-}dist}^{known}$ 和 $\mathcal{L}_{con}^{unknown}$ 的贡献如何？有三个关键的损失项，即方程 3 中的 $\mathcal{L}_{f\text{-}dist}$ 、方程 5 中的 $\mathcal{L}_{s\text{-}dist}^{known}$ 和方程 6 中的 $\mathcal{L}_{con}^{unknown}$ 。为了验证它们的有效性，我们进行了消融实验，并在表格 2 中报告了性能。去掉这三个损失项后，PSScreen 退化为配备文本引导的语义解耦的多头机制，其性能明显不如 PSScreen。在单独添加 $\mathcal{L}_{f\text{-}dist}$ 和 $\mathcal{L}_{s\text{-}dist}^{known}$ 后，域内性能和域泛化性能均得到提升。相反，单独使用 $\mathcal{L}_{con}^{unknown}$ 导致性能下降，这可能是由于在概率特征图中与任务相关的语义丢失所致。结合 $\mathcal{L}_{s\text{-}dist}^{known}$ 和 $\mathcal{L}_{con}^{unknown}$ 提高了模型的性能。集成所有损失项达到了最佳效果，与没有这三个损失项的退化版本相比，在未见过的数据集和 ODIR 数据集上的改进分别为 5.5 % 和 11.8 %，验证了 PSScreen 的有效性。

损失权重 λ_1 、λ_2 和 λ_3 的影响。我们改变 λ_1 、λ_2 和 λ_3 ，并在表格 3 (a) 中报告 *mQWK* ，结果显示 $\lambda_1 = 0.05$ 、$\lambda_2 = 1.0$ 和 $\lambda_3 = 0.6$ 达到了最佳表现。关于 *mF* 的表现可以在补充材料 C .1 中找到。

与不同主干网络的兼容性。为了进一步验证我们 PSScreen 的兼容性，我们报告了 PSScreen 与不同主干网络（包括 ConvNeXt-T [22] 、ConvNeXt V2-T [33] 、Swin-T [21] 和 VMamba-T [20] ）的 *mQWK* ，并将它们与简单版本进行比较，也就是表格 3 (b) 中列出的具有各种主干网络的多头机制。结果表明，我们的 PSScreen 在所有三种数据集和不同的主干网络下都能稳定地提高性能。关于 *mF* 的性能可以在补充材料 ?? 中找到。

4 结论与未来工作

在本文中，我们提出了 PSScreen，这是一个多疾病视网膜筛查模型，它是在多种不同分布的部分标记数据集上训练的。PSScreen 是一个双流网络，其中一个流学习确定性特征，另一个流通过 DSU 模块学习概率特征。然后，这些特征通过文本指导的语义解耦模块进行解耦，以促进多疾病检测。为了强迫概率特征学习流学习与任务

相关的语义，特征蒸馏和自蒸馏被用来将与任务相关的语义从确定性流转移到概率流。同时，针对缺失标签问题施加伪标签一致性。在充分和部分标记数据集的广泛实验中表明，PSScreen 在多个同域和跨域数据集上达到了最先进的性能。

PSScreen 具有轻量级架构和快速推断速度，非常适合集成到临床工作流程中。然而，将筛查模型从实验室应用到临床仍面临挑战。例如，眼科成像模式多种多样，而我们目前的重点仅限于眼底图像。为了增强临床适用性，未来可考虑增加额外的成像模式以筛查更广泛的视网膜疾病。尽管提供了基于 GradCAM 的模型解释性可视化，但缺乏涉及临床医生的结构化评估。此外，如何进行临床医师参与的评估以验证模型的可解释性仍然是一个未解决的问题，值得进一步研究。

5 致谢

本工作部分得到了芬兰科学院研究员项目的资助，项目编号为 355095，以及中国湖南省自然科学基金项目的资助，项目编号为 2023JJ30699。此外，作者还要感谢芬兰 CSC–IT 科学中心提供的计算资源。

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.

6

PSScreen 补充材料：部分监督的多种视网膜疾病筛查

A 开放获取数据集的详细信息。

表 4 展示了我们研究中使用的所有开放获取数据集的详细信息。为了建立实验评估协议，我们对其中的一部分进行了处理。RFMiD [25] 包含 46 种视网膜疾病或结构异常类别，其中仅使用正常状态和 AMD 类别用于我们的研究。对于 ODIR [17]，“其他疾病”类别被排除，因为它与我们的研究无关，并且根据 ODIR 官方补充资料 [1] 排除有问题的图像（如受镜头灰尘影响或质量不佳的图像），最终得到包含 6,961 张训练图像和 988 张测试图像的数据集。

Group	Dataset	Resolution	# images	Original Splitting			Labels							train/test
				Train	Valid	Test	N	D	G	C	A	H	P	
Meta	Kaggle-CAT[2]	2592 × 1728 or 2464×1632	600	360	120	120	✓		✓	✓				✓/✓
	DDR [18]	max: 5184×3456 min: 512×512	12522	6261	2504	3757		✓						✓/✓
	REFUGE2 [9]	max: 2124×2056 min: 1634×1634	2000	1200	400	400			✓					✓/✓
	ADAM [8]	2124 × 2056 or 1444×1444	1200	400	400	400					✓			✓/✓
	Kaggle-HR[3]	800×800	712	427	142	143						✓		✓/✓
	PALM [10]	2124 × 2056 or 1444×1444	1200	400	400	400							✓	✓/✓
Unseen	RFMiD [25]	max: 4288×2848 min: 2048×1536	3200	1920	640	640	✓				✓			✗/✓
	APTOS2019[4]	max: 4288×2848 min: 474×358	3662	—	—	—		✓						✗/✓
	ORIGA light [38]	3072 × 2048	650	—	—	—			✓					✗/✓
	HPMI [13]	512 × 512	4011	—	—	—							✓	✗/✓
ODIR	ODIR [17]	max: 5184×3456 min: 160x120	7949	6961	—	988	✓	✓	✓	✓	✓	✓	✓	✗/✓
ODIR 200×3	ODIR 200x3 [31]	max: 5184×3456 min: 868x793	600	—	—	—	✓			✓			✓	✗/✓

Table 4: 本研究中使用的开放获取数据集的详细信息。其中，“N”表示正常状态，“D”表示糖尿病视网膜病变（DR），“G”表示青光眼，“C”表示白内障，“A”表示年龄相关性黄斑变性（AMD），“H”表示高血压视网膜病变（HR），和“P”表示病理性近视（PM）。

B 每种疾病的实验结果

B .1 各疾病在元数据集上的表现

我们分别报告了 PSScreen、基线和现有 SOTA 在每种疾病的元数据集上的 *mF* 和 *mQWK*，其结果分别在表 5 和表 6 中可以找到。可以观察到，我们的 PSScreen 在 *mF* 中超越了排名第二的 SARB [26] 0.2 %，在 *mQWK* 中超越了 1.1 %。更具体地说，PSScreen 在检测青光眼、AMD、HR 和 PM 时取得了最佳或第二佳的成绩。尽管在 REFUGE2 的训练和测试集之间存在显著的领域转变，我们的 PSScreen 在 QWK 指标上比排名第二的 MultiHeads 在 REFUGE2 上高 0.9 %，进一步证明了我们模型的良好泛化能力。

[1] https://odir2019.grand-challenge.org/Download/
[2] https://www.kaggle.com/datasets/jr2ngb/cataractdataset
[3] https://www.kaggle.com/datasets/harshwardhanfartale/hypertension-and-hypertensive-retinop
[4] https://www.kaggle.com/competitions/aptos2019-blindness-detection/data

Methods	T1: Normal	T2: DR	T3: Glaucoma			T4: Cataract	T5: AMD	T6: HR	T7: PM	*mF*
	Kaggle-CAT	DDR	REFUGE2	Kaggle-CAT	Average	Kaggle-CAT	ADAM	Kaggle-HR	PALM	
MultiNets	78.1 ±1.2	67.8 ±0.9	75.7 ±1.3	89.5 ±1.0	82.6 ±1.0	88.0 ±2.0	85.3 ±1.7	80.4 ±3.7	96.7 ±0.4	82.7 ±0.7
MultiHeads	86.8 ±0.8	66.8 ±1.0	77.2 ±1.4	85.4 ±0.4	81.3 ±0.6	88.8 ±3.2	84.4 ±1.2	81.1 ±1.7	96.3 ±0.1	83.6 ±0.7
SST *AAAI2022* [2]	87.6 ±0.8	67.1 ±1.1	73.6 ±1.1	87.3 ±1.5	80.5 ±0.6	92.3 ±2.8	83.7 ±2.2	77.6 ±3.4	95.2 ±0.7	83.4 ±1.4
SARB *AAAI2022* [26]	86.2 ±3.3	66.7 ±1.6	77.0 ±2.4	86.9 ±0.7	81.9 ±1.5	90.0 ±0.9	85.9 ±0.9	81.4 ±2.1	96.0 ±1.0	84.0 ±0.9
BoostLU *CVPR2023* [15]	80.2 ±4.2	62.0 ±1.3	72.6 ±1.8	85.8 ±3.1	78.6 ±2.6	90.0 ±2.0	76.8 ±2.3	78.8 ±1.5	95.9 ±0.4	80.3 ±0.5
HST *IJCV2024* [3]	87.0 ±0.5	67.0 ±0.3	72.3 ±2.5	89.0 ±3.1	80.7 ±0.8	92.6 ±1.7	81.2 ±2.1	79.6 ±1.0	95.6 ±1.0	83.4 ±0.4
CALDNR *TMM2024* [27]	82.5 ±3.1	63.8 ±1.3	74.6 ±4.0	88.6 ±1.6	81.6 ±1.4	93.1 ±1.5	75.6 ±7.3	81.8 ±3.8	95.7 ±0.2	82.0 ±1.1
PSScreen *Ours*	86.4 ±1.7	64.7 ±1.7	77.7 ±1.5	89.0 ±3.1	83.3 ±2.0	89.1 ±2.4	84.5 ±1.4	85.0 ±1.5	96.5 ±0.4	84.2 ±0.3

Table 5: 在元数据集上比较每种疾病的 F-score。最优和次优的结果分别以粗体和下划线标出。均值和标准差基于三次试验报告。

Methods	T1: Normal	T2: DR	T3: Glaucoma			T4: Cataract	T5: AMD	T6: HR	T7: PM	*mQWK*
	Kaggle-CAT	DDR	REFUGE2	Kaggle-CAT	Average	Kaggle-CAT	ADAM	Kaggle-HR	PALM	
MultiNets	56.5 ±2.6	88.1 ±0.1	51.4 ±2.6	79.0 ±4.0	65.2 ±3.7	76.0 ±4.1	70.5 ±3.5	61.2 ±7.4	93.5 ±0.9	73.0 ±1.4
MultiHeads	73.6 ±1.6	87.8 ±0.8	54.5 ±2.8	71.0 ±0.7	62.7 ±1.2	77.5 ±6.5	68.7 ±2.5	62.3 ±3.5	92.6 ±0.3	75.0 ±1.3
SST *AAAI2022* [2]	75.2 ±1.6	88.4 ±0.2	47.6 ±2.1	74.6 ±3.1	61.1 ±1.4	84.7 ±5.5	67.5 ±4.2	55.2 ±6.8	90.4 ±1.3	74.7 ±2.8
SARB *AAAI2022* [26]	72.5 ±6.7	87.3 ±0.9	54.0 ±4.7	77.1 ±1.5	63.9 ±3.0	80.0 ±1.8	71.7 ±1.7	62.7 ±4.3	92.0 ±0.9	75.7 ±1.8
BoostLU *CVPR2023* [15]	60.8 ±8.2	84.0 ±2.2	46.1 ±3.6	71.7 ±6.0	57.6 ±5.2	80.1 ±4.0	54.7 ±4.1	57.6 ±3.0	91.8 ±0.8	69.5 ±1.1
HST *IJCV2024* [3]	74.1 ±0.9	88.8 ±0.1	45.0 ±4.5	78.0 ±6.2	61.5 ±1.6	85.3 ±3.5	62.5 ±4.3	59.3 ±1.8	91.3 ±1.0	74.7 ±0.8
CALDNR *TMM2024* [27]	65.3 ±5.9	87.0 ±0.8	50.0 ±7.3	77.2 ±3.2	63.6 ±2.4	86.3 ±3.1	52.4 ±13.2	63.7 ±7.5	91.5 ±0.5	72.8 ±2.2
PSScreen *Ours*	73.0 ±3.4	87.4 ±0.2	55.4 ±3.0	78.1 ±6.2	66.8 ±4.0	78.3 ±4.7	69.1 ±2.7	70.1 ±3.0	93.0 ±0.9	76.8 ±0.8

Table 6: 每种疾病在元数据集上的 QWK 比较。最佳和次佳用粗体和下划线显示。平均值和标准差基于三次试验报告。

Methods	T1: Normal		T2: DR		T3: Glaucoma		T4: AMD		T5: PM		Average	
	RFMID		APTOS		ORIGA		RFMID		HPMI			
	F-score	QWK	F-score	QWK	F-score	QWK	F-score	QWK	F-score	QWK	*mF*	*mQWK*
MultiNets	68.1 ±2.3	36.7 ±4.2	41.3 ±2.2	70.2 ±5.8	71.3 ±0.7	42.8 ±1.5	48.3 ±2.4	7.3 ±1.1	83.1 ±1.1	66.8 ±2.1	62.4 ±1.2	44.8 ±2.3
MultiHeads	67.4 ±4.1	36.3 ±7.5	44.7 ±2.7	77.7 ±4.7	67.9 ±1.8	36.7 ±3.0	52.3 ±2.0	9.9 ±2.1	79.6 ±0.4	60.4 ±0.8	62.4 ±1.5	44.2 ±2.0
SST *AAAI2022* [2]	76.8 ±4.2	53.6 ±8.5	42.0 ±0.5	74.9 ±2.1	63.9 ±1.6	29.0 ±2.7	52.1 ±1.3	12.4 ±2.6	81.3 ±1.0	63.5 ±1.1	63.2 ±0.9	46.7 ±1.8
SARB *AAAI2022* [26]	63.4 ±4.3	28.2 ±7.4	41.7 ±0.8	70.8 ±3.3	65.8 ±0.7	32.0 ±1.8	57.4 ±3.3	18.9 ±2.8	74.3 ±2.6	51.3 ±4.2	60.5 ±1.0	40.3 ±1.2
BoostLU *CVPR2023* [15]	46.0 ±1.2	1.2 ±1.5	37.3 ±1.2	58.1 ±2.9	52.8 ±1.9	13.7 ±2.7	53.9 ±1.8	10.3 ±3.5	32.6 ±3.8	7.2 ±1.2	44.5 ±0.8	29.8 ±2.3
HST *IJCV2024* [3]	74.6 ±5.9	48.4 ±11.5	42.9 ±2.3	74.0 ±5.9	64.7 ±2.0	29.7 ±3.8	50.9 ±0.7	18.0 ±0.6	78.8 ±1.8	59.0 ±3.2	62.3 ±1.7	44.8 ±3.8
CALDNR *TMM2024* [27]	55.8 ±4.4	14.9 ±9.7	32.1 ±6.3	42.1 ±18.5	58.8 ±1.1	22.3 ±1.7	50.6 ±3.2	3.3 ±5.7	41.0 ±3.7	12.5 ±6.2	47.6 ±3.1	19.0 ±6.7
PSScreen *Ours*	80.2 ±1.4	60.5 ±2.5	46.7 ±0.6	84.5 ±0.9	68.0 ±1.2	36.9 ±2.2	59.1 ±2.6	19.6 ±4.6	75.4 ±0.8	53.1 ±1.3	65.9 ±0.1	50.9 ±0.1

Table 7: 对未见数据集上的每种疾病进行 F-score 和 QWK 的比较。最佳和次佳结果分别以粗体和下划线突出显示。均值和标准差基于三次试验报告。

B .2 每种疾病在未见数据集和 ODIR 数据集上的表现

我们在表格 7 中展示了每种疾病在未见数据集上的表现。如图所示，PSScreen 在 *mF* 中比第二好的 SST [2] 高出 2.7 %，在 *mQWK* 中高出 4.2 %。更具体地说，PSScreen 在正常状态检测、AMD 检测和 DR 分级任务中实现了最佳表现。值得注意的是，它在 QWK 上取得了显著的提升，在正常状态检测中比第二好的 SST [2] 高出 6.9 %，在 DR 分级中比第二好的 MultiHeads 高出 6.8 %。此外，在青光眼检测任务中，PSScreen 排名第二。

我们还在表格 8 和 9 中分别报告了在 ODIR 数据集 [17] 上的 F-score 和 QWK。作为上限性能的参考，全监督模型的性能也在表格 8 和表格 9 的第一行中报告。可以观察到，PSScreen 相比于第二好的 SARB [26] 实现了显著提升，*mF* 提高了 2.7 %，*mQWK* 提高了 6.0 %。更具体地，我们发现除了 HR 检测，PSScreen 在所有任务中都取得了最佳表现。值得注意的是，在正常状态检测、DR 分级和青光眼检测中，PSScreen 在 QWK 指标上相对于分别是第二好的方法 SST [2] 、BoostLU [15] 和 CALDNR [27] 有显著提升，分别增益了 5.7 %、5.2 % 和 6.7 %。特别是在青光眼检测和 DR 分级中，我们模型的 F-score 甚至超过了全监督模型，进一步验证了 PSScreen 的强域泛化能力。此外，在 HR 检测中，所有方法表现都不好，甚至全监督模型也难以做出准确预测。

Methods	T1: Normal	T2: DR	T3: Glaucoma	T4: Cataract	T5: AMD	T6: HR	T7: PM	*mF*
Full supervise	70.8 ±0.2	32.1 ±1.7	66.8 ±0.9	91.8 ±0.8	80.2 ±1.4	58.8 ±1.1	87.8 ±1.2	69.8 ±0.2
MultiNets	59.3 ±2.6	28.7 ±0.7	60.8 ±1.7	66.2 ±1.9	59.2 ±5.7	47.7 ±2.9	72.7 ±1.4	56.4 ±1.4
MultiHeads	53.9 ±1.1	29.7 ±1.3	62.8 ±2.4	67.3 ±0.5	57.4 ±2.4	47.7 ±1.1	78.2 ±1.9	56.7 ±0.6
SST *AAAI2022* [2]	62.8 ±3.2	29.1 ±1.6	58.7 ±1.0	70.9 ±3.9	53.0 ±1.1	44.1 ±1.5	70.6 ±4.2	55.6 ±1.7
SARB *AAAI2022* [26]	60.8 ±2.2	31.7 ±2.3	59.1 ±1.9	84.6 ±2.2	61.2 ±4.2	51.0 ±1.3	81.3 ±0.9	61.4 ±0.9
BoostLU *CVPR2023* [15]	36.1 ±0.0	29.6 ±2.8	63.5 ±3.4	48.8 ±0.0	61.7 ±5.5	49.0 ±0.0	66.8 ±3.1	50.6 ±1.0
HST *IJCV2024* [3]	60.9 ±2.4	29.0 ±1.9	60.4 ±1.2	76.0 ±2.8	50.4 ±2.0	43.2 ±1.7	71.9 ±1.7	56.0 ±0.8
CALDNR *TMM2024* [27]	36.1 ±0.0	33.5 ±3.1	63.5 ±3.4	55.9 ±10.6	58.9 ±3.7	49.3 ±0.0	67.9 ±1.8	52.1 ±0.6
PSScreen *Ours*	66.8 ±1.4	33.1 ±1.9	66.9 ±2.3	85.9 ±1.6	63.0 ±0.8	51.1 ±2.3	81.6 ±2.6	64.1 ±1.0

Table 8: ODIR 数据集中每种疾病的 F-score 比较。最佳和次佳的结果分别以加粗和下划线标出。统计结果为三次试验的均值和标准差。

Methods	T1: Normal	T2: DR	T3: Glaucoma	T4: Cataract	T5: AMD	T6: HR	T7: PM	*mQWK*
Full supervise	41.6 ±0.4	53.6 ±3.1	33.8 ±1.7	83.5 ±1.5	60.5 ±2.8	19.2 ±0.8	75.6 ±2.3	52.6 ±0.7
MultiNets	19.2 ±4.4	29.7 ±1.6	24.8 ±2.2	35.4 ±3.2	22.9 ±8.3	3.9 ±2.5	46.6 ±2.5	26.1 ±2.1
MultiHeads	17.9 ±1.9	34.0 ±2.5	26.9 ±4.1	36.5 ±1.3	18.8 ±3.6	3.2 ±1.7	56.8 ±3.6	27.7 ±0.4
SST *AAAI2022* [2]	28.9 ±4.3	32.3 ±1.0	21.7 ±1.5	43.3 ±6.8	14.0 ±1.8	0.3 ±1.1	42.8 ±7.5	26.2 ±2.5
SARB *AAAI2022* [26]	24.3 ±3.6	31.9 ±2.1	21.0 ±2.9	69.3 ±4.5	24.5 ±7.2	2.6 ±2.2	62.9 ±1.8	33.8 ±1.6
BoostLU *CVPR2023* [15]	0.0 ±0.0	39.6 ±7.3	24.7 ±1.0	0.0 ±0.0	24.5 ±10.7	0.0 ±0.0	34.0 ±6.2	17.5 ±2.1
HST *IJCV2024* [3]	25.4 ±2.9	33.1 ±0.4	23.9 ±1.5	52.7 ±5.2	11.8 ±1.9	0.6 ±0.4	45.3 ±3.2	27.5 ±0.8
CALDNR *TMM2024* [27]	0.0 ±0.0	38.7 ±3.9	27.3 ±6.6	13.5 ±20.3	18.7 ±7.1	0.0 ±0.0	36.1 ±3.6	19.1 ±0.8
PSScreen *Ours*	34.6 ±2.1	44.8 ±2.5	34.0 ±4.4	71.9 ±3.1	27.6 ±1.4	2.3 ±4.6	63.4 ±5.2	39.8 ±1.3

Table 9: 在 ODIR 数据集上，每种疾病的 QWK 比较。最佳和第二佳用加粗和下划线标出。均值和标准差是在三次实验中得出的。

λ_1	λ_2	λ_3	Meta	Unseen	ODIR
0.1	1.0	0.6	83.2± 0.2	65.4± 0.1	63.6± 0.5
0.05	1.0	0.6	84.2± 0.3	65.9± 0.1	64.1± 1.0
0.025	1.0	0.6	84.1± 0.1	65.7± 0.4	63.9± 0.5
0.05	0.5	0.6	83.2± 0.1	65.2± 0.2	63.4± 0.4
0.05	1.0	0.6	84.2± 0.3	65.9± 0.1	64.1± 1.0
0.05	2.0	0.6	82.9± 0.2	64.5± 0.8	63.6± 0.2
0.05	1.0	0.4	84.1± 0.5	64.7± 0.6	63.9± 0.4
0.05	1.0	0.6	84.2± 0.3	65.9± 0.1	64.1± 1.0
0.05	1.0	0.8	84.2± 0.3	65.4± 0.9	63.2± 0.1

(a) F-score under different loss weight settings.

Methods	Meta	Unseen	ODIR
ResNet-101 [11]	83.6± 0.7	62.4± 1.5	56.7± 0.6
+PSScreen	84.2± 0.3	65.9± 0.1	64.1± 1.0
ConvNeXt-T [22]	83.6± 0.3	64.4± 0.6	55.8± 1.8
+PSScreen	84.5± 0.2	65.6± 0.7	62.9± 0.8
ConvNeXt V2-T [33]	84.4± 0.4	65.1± 0.6	58.1± 0.9
+PSScreen	84.5± 0.1	65.3± 0.5	63.7± 0.7
Swin-T [21]	83.6± 0.2	65.5± 0.5	56.1± 0.3
+PSScreen	85.3± 0.5	66.2± 0.3	61.8± 0.5
VMamba-T [20]	84.0± 0.8	64.3± 0.4	56.8± 0.4
+PSScreen	84.9± 0.1	65.3± 1.3	62.6± 1.0

(b) Performances with different backbones.

Table 10: (a) 在不同的 λ_1 、λ_2 和 λ_3 设置下的 *mF* 。(b) 通过不同骨干网络的 MultiHeads 和 PSScreen 得到的 *mF* 。

τ	Meta		Unseen		ODIR	
	mF	*mQWK*	*mF*	*mQWK*	*mF*	*mQWK*
0.99	83.9± 0.2	76.1± 0.4	64.7± 0.3	49.4± 0.5	62.9± 0.3	38.7± 0.6
0.95	84.2± 0.3	76.8± 0.8	65.9± 0.1	50.9± 0.1	64.1± 1.0	39.8± 1.3
0.90	84.0± 0.3	76.6± 0.5	65.0± 0.3	50.1± 0.5	63.7± 0.2	39.3± 0.3
0.85	83.8± 0.1	76.1± 0.1	64.4± 0.2	48.8± 0.4	63.0± 0.8	38.0± 1.2

Table 11: 在不同的 τ 设置下对 *mF* 和 *mQWK* 的性能分析。

C 额外的消融研究结果

C .1 损失权重 λ_1 、λ_2 和 λ_3 的影响

我们变化了 λ_1 、λ_2 和 λ_3 ，并在表 10 (a) 中报告了 *mF* ，这表明 $\lambda_1 = 0.05$ 、$\lambda_2 = 1.0$ 和 $\lambda_3 = 0.6$ 达到了最佳性能。

为了进一步验证我们 PSScreen 的兼容性，我们报告了 PSScreen 与各种骨干网络的 *mF* ，包括 ConvNeXt-T [22] 、ConvNeXt V2-T [33] 、Swin-T [21] 和 VMamba-T [20] ，并将它们与基础版本进行比较，即在表格 10 (b) 中具有各种骨干网络的 MultiHeads。结果表明，我们的 PSScreen 在所有三个数据集上与各种骨干网络一致地提高了性能。

在方程 6 中，τ 是获取伪标签的阈值，我们改变其值并在表 11 中报告性能，以展示其影响。我们观察到，当 τ 设置得太大时，很少有样本满足阈值条件，限制了监督

Disease	Description
normal	"healthy", "no findings", "no lesion signs", "no glaucoma", "no retinopathy"
glaucoma	"optic nerve abnormalities", "abnormal size of the optic cup", "anomalous size in the optic disc"
cataract	"opacity in the macular area"
age-related macular degeneration	"many small drusen", "few medium-sized drusen", "large drusen", "macular degeneration"
hypertensive retinopathy	"possible signs of haemorraghe with blot, dot, or flame-shaped", "possible presence of microaneurysm, cotton-wool spot, or hard exudate", "arteriolar narrowing", "vascular wall changes", "optic disk edema"
pathologic myopia	"anomalous disc, macular atrophy and possible tessellation"
no diabetic retinopathy	"no diabetic retinopathy", "no microaneurysms"
mild diabetic retinopathy	"only few microaneurysms"
moderate diabetic retinopathy	"many exudates near the macula", "many haemorrhages near the macula", "retinal thickening near the macula", "hard exudates", "cotton wool spots", "few severe haemorrhages"
severe diabetic retinopathy	"venous beading", "many severe haemorrhages", "intraretinal microvascular abnormality"
proliferative diabetic retinopathy	"preretinal or vitreous haemorrhage", "neovascularization"

Table 12: 每种视网膜疾病的专家知识描述，直接借鉴自 [31] 。

信号；而当 τ 太小时，许多样本被错误地分配了伪标签，这会损害训练的稳定性并降低性能。当 $\tau = 0.95$ 时，PSScreen 达到最佳性能；因此，我们在实验中将 τ 设置为 0.95。值得注意的是，PSScreen 在 τ 值的一定范围内仍然保持稳健，只有当阈值设置过低时，导致大量噪声伪标签时，性能才会出现明显下降。

使用文本引导的语义解耦模块时，我们对每个疾病的多个专家知识描述进行编码，并计算其嵌入向量的平均值作为该疾病的最终文本嵌入。表 12 列出了与每个疾病对应的多个专家知识描述，这些描述直接借自 [31] 。