

不确定感知多专家知识蒸馏在疾病分级中的应用

Shuo Tong^{1,4,5,*}, Shangde Gao^{2,3,4,5,*}, Ke Liu², Zihang Huang⁶, Hongxia Xu³, Haochao Ying^{1,4,✉}, and Jian Wu^{1,3,4,5}

¹ School of Public Health, Zhejiang University

² College of Computer Science and Technology, Zhejiang University

³ Liangzhu Laboratory and WeDoctor Cloud

⁴ State Key Laboratory of Transvascular Implantation Devices, The Second Affiliated Hospital Zhejiang University School of Medicine

⁵ Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence

⁶ School of Electronic Information and Communications, Huazhong University of Science and Technology

{tongshuo, gaosde, lk2017, Einstein, haochaoying, wujian2000}@zju.edu.cn

摘要 自动疾病图像分级是人工智能在医疗健康领域的一项重要应用，能够实现更快更准确的患者评估。然而，由数据不平衡加剧的域转移问题会将偏差引入模型中，在临床应用部署过程中造成困难。为了解决这一问题，我们提出了一种新的 **U** 不确定性感知 **M** 多专家 **K** 知识 **D** 蒸馏 (UMKD) 框架，以从多个专家模型向单一学生模型转移知识。具体来说，为了提取辨别特征，UMKD 在特征空间中通过浅层且紧凑的特征对齐解耦任务无关和任务特定特征。在输出空间中，一种不确定性感知的分离蒸馏 (UDD) 机制根据专家模型的不确定性动态调整知识传输权重，确保稳健可靠的知识迁移。此外，UMKD 还解决了模型架构异质性和源域与目标域之间分布差异的问题，这些问题之前的知识蒸馏方法未能充分解决。在组织学前列腺分级 (*SICAPv2*) 和眼底图像分级 (*APTOS*) 上的大量实验表明，UMKD 在数据不平衡的源域和目标域场景中均达到了新的先进水平，为实际疾病图像分级提供了稳健且实用的解决方案。

Keywords: 知识蒸馏 · 疾病分级 · 数据不平衡

1 介绍

图像驱动的疾病分级系统对于提高临床决策效率至关重要 [14,22]，特别是对于糖尿病视网膜病变 (DR) 和前列腺癌。早期精确分级显著改善患者

*: 同等贡献作者。

✉: 通讯作者。

预后 (例如, 及时干预可将 DR 患者的失明风险降低 90%) [17,1]。然而, 传统分级受到主观专家判断差异以及难以识别细微病理特征等挑战的限制。临床证据报告格里森评分存在 40% 的观察者间变异性, 并且早期 DR 微动脉瘤误诊率超过 25%。

最近, 由于基于人工智能的自动疾病分级系统的快速发展, 医学效率得到了提升 [2,6]。对于 DR 分级, 戴等人士。[4] 开发了 DeepDR 来检测糖尿病视网膜病变从早期到晚期的不同阶段。王等人。[20] 将 DR 分级重新表述为序列预测, 有效解决了模糊边界问题。另一方面, 对于前列腺癌的分级, Morpho-Grader [3] 分离了腺体形态和间质纹理。BayeSeg [8] 采用变分推断将结构不变量与纹理变化分离, 显著提高了模型的鲁棒性。

尽管疾病分级方法取得了进展 [15,1], 其在临床应用中的部署仍然受到领域转移的限制。特别是, 不平衡的数据加剧了领域转移 (源域和目标域分布之间的差异)。如 SICAPv2 数据集 (图 1 左侧) 所示, III 期前列腺癌样本占队列的 8%, 导致模型过度拟合多数类。这种不平衡在 DR 分级中同样显著, 早期病变稀少降低了微血管瘤检测敏感性超过 30% [12]。多专家知识蒸馏 (MKD) [5,9,7], 是一种通过转移专家模型的知识来改善学生模型对少数类样本泛化的技术。由于其在领域转移方面的鲁棒性, MKD 已被应用于解决自然图像中的类别不平衡问题 [21,13], 但在疾病影像分级中的研究仍较少。

本文中, 我们提出了一种新型的 **U** 不确定性感知 **M** 多专家 **K** 知识 **D** 蒸馏 (UMKD) 框架来解决类别不平衡的问题。具体来说, 为了分离图像表示中的结构和语义信息, 我们设计了两种特征对齐机制: 浅层特征对齐 (SFA) 和紧凑特征对齐 (CFA)。SFA 通过多尺度低通滤波泛化专家与学生模型之间的特征对齐, 从而保留疾病图像的结构信息 (任务无关特征)。CFA 将专家模型和学生模型的特征映射到一个共同的球形空间中, 使学生模型能够从每个专家那里学习分级相关的特征知识。我们在输出空间设计了一个不确定性感知的解耦蒸馏 (UDD) 机制, 自动检测由类别不平衡导致的专家模型中的不确定性。通过不确定性指标, 学生模型动态调整知识转移权重, 减少偏差传播, 确保更加健壮和可靠的知识转移过程。实验结果显示, 我们的方法在视网膜和前列腺疾病图像分级任务中显著优于现有的多专家蒸馏方法。特别是对于类别不平衡和异构模型, UMKD 实现了更可靠的知识转移, 如图 1 右所示。

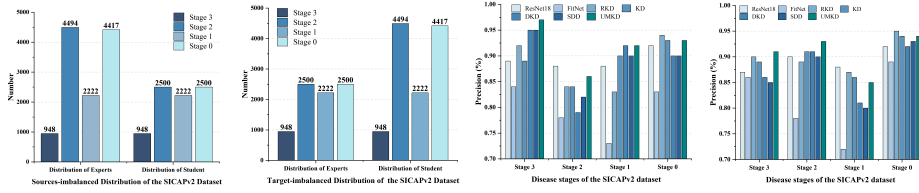


图 1: 源数据和目标数据之间的域转移（左）以及方法在来源不平衡和目标不平衡知识蒸馏任务上的性能（右）。

2 方法

不确定性感知的多专家知识蒸馏模型。我们旨在将多专家模型的知识提炼到目标学生模型中，用于处理不平衡的疾病分级任务。如图 2 所示，我们的框架包括浅层特征对齐（SFA）、紧凑特征对齐（CFA）和不确定性感知解耦蒸馏（UDD）。首先，我们将介绍特征对齐损失函数，该函数将被 SFA 和 CFA 模块重复使用。接下来，将提供每个模块的详细描述。

最大均值差异和重构损失。为了衡量学生特征与每位专家的特征在特征空间中的分布差异以实现特征对齐，我们采用最大均值差异（MMD）[5]。MMD 距离计算如下：

$$\mathcal{L}_{\text{MMD}} = \frac{1}{B} \sum_{t=1}^N \left\| \sum_{i=1}^B \phi(\hat{F}_{T_t}^i) - \sum_{j=1}^B \phi(\hat{F}_S^j) \right\|_2^2, \quad (1)$$

其中， ϕ 是一个显式映射函数， $\hat{F}_{T_t}$ 和 $\hat{F}_S$ 分别表示专家 T_t 和学生 S 投影后的特征， B 是批次大小， N 是专家数量。同时，为了确保专家模型在隐私约束下保持不变，使用重构损失 $\mathcal{L}_{\text{MSE}}$ 来衡量特征对齐前后专家模型的变化：

$$\mathcal{L}_{\text{MSE}} = \sum_{t=1}^N \left\| F_{T_t} - \hat{F}_{T_t} \right\|_2^2, \quad (2)$$

其中 F_{T_t} 表示对齐前专家模型 T_t 的原始特征，而 $\hat{F}_{T_t}$ 表示对齐后专家模型 T_t 的解码特征。通过聚合对齐损失和重构损失，特征对齐的总损失可以表示为：

$$\mathcal{L}_{\text{FA}} = \mathcal{L}_{\text{MMD}} + \mathcal{L}_{\text{MSE}}. \quad (3)$$

浅层特征对齐。SFA 在去除噪声和高频细节的同时保留给定图像的主要结构信息，确保专家模型与学生模型之间泛化特征学习的一致性 [9]。具体来

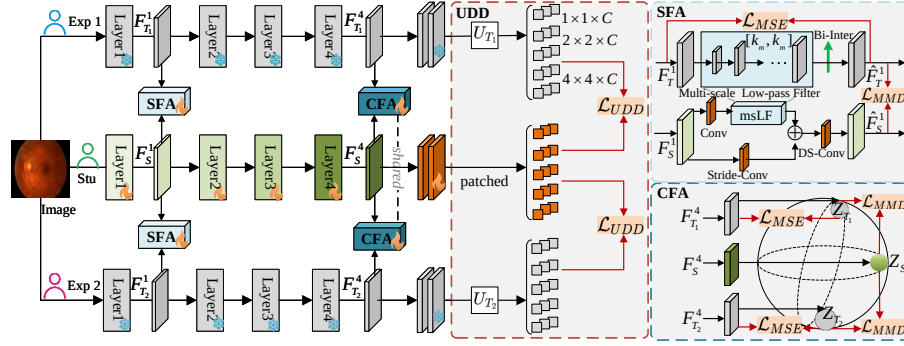


图 2: 不确定感知多专家知识蒸馏模型。

说，在异构的专家-学生模型之间的浅层特征表示中，我们提出采用多尺度低通滤波从频率域对齐特征。对于每个专家特征 F_{T_i} ，我们采用传统的平均池化作为低通滤波器，并通过调整不同组中的内核大小和步幅来构建多尺度滤波器以适应不同的截止频率。对于第 m 组，通过多尺度低通滤波器 (msLF) 的频域特征 $\hat{F}_{T_i}$ 可以表示为：

$$\hat{F}_{T_i} = \text{msLF}(F_{T_i}) = \Phi(\text{AvgPool}_{k_m \times k_m}(F_{T_i})), \quad (4)$$

其中， $\text{AvgPool}_{k_m \times k_m}$ 表示核大小为 $k_m \times k_m$ 的平均池化函数， $\Phi(\cdot)$ 代表双线性插值操作。

对于学生特征 F_S ，我们设计了一个可学习的低通滤波器，它由多尺度低通滤波器、卷积下采样模块和深度可分离卷积（DSConv）组成。频率域中的可学习学生特征 $\hat{F}_S$ 表示为：

$$\hat{F}_S = \text{Conv}_{3 \times 3}(\text{Concat}[\text{DownSample}_{s \times s}(F_S), \text{msLF}(F_S)]), \quad (5)$$

其中， $\text{DownSample}_{s \times s}$ 、 Concat 和 $\text{Conv}_{3 \times 3}$ 分别表示卷积下采样模块、特征拼接操作和 3×3 卷积操作。获得使用 msLF 变换的频域教师专家和学生特征后，SFA 使用的损失为方程 (3) 中的损失重写为： $\mathcal{L}_{\text{SFA}} = \mathcal{L}_{\text{MMD}} + \mathcal{L}_{\text{MSE}}$ 。

紧凑特征对齐。 CFA 将所有模型第四层（全连接层之前的倒数第二层）的特征集投影到一个紧凑的高维球形空间 $\mathcal{Z}$ 中。在这个空间中，学生模型可以通过空间域特征对齐从不同的预训练专家那里学习与退化相关的分层知识。考虑到模型之间的异质性，它们编码器的输出特征维度可能不同。在执行 CFA 之前，在每个编码器的末尾添加一个 1×1 卷积核以调整不同编码

器的输出特征到相同的维度。在球形空间 $\mathcal{Z}$ 中, 使用 MMD 和 MSE 损失计算空间 CFA 的总损失为: $\mathcal{L}_{\text{SFA}} = \mathcal{L}_{\text{MMD}} + \mathcal{L}_{\text{MSE}}$ 。

感知不确定性的解耦蒸馏。为了解决专家偏见对输出预测的影响, 我们提出了一种不确定性感知的蒸馏模块, 该模块能够动态地将每个专家的全局和局部知识转移到学生网络。给定来自专家 T_t 和学生 S 的 logits 图 $L_{T_t} \in \mathbb{R}^{C \times H \times W}$ 和 $L_S \in \mathbb{R}^{C \times H \times W}$, 我们在多个尺度 $w \in W = \{1, 2, 4, \dots, w_{\max}\}$ 上应用空间划分 $\mathcal{P}(w, w)$ 。对于每个尺度为 w 的划分单元格 $Z(w, n)$ (其中 $n \in N_w = \{1, 4, 16, \dots, w^2\}$), 累积的对数几率是:

$$\psi_{T_t}(w, n) = \frac{1}{w^2} \sum_{(j,k) \in Z(w,n)} L_{T_t}(j, k), \quad \psi_S(w, n) = \frac{1}{w^2} \sum_{(j,k) \in Z(w,n)} L_S(j, k). \quad (6)$$

然后, 我们设计了一个不确定性系数 U_{T_t} , 它包含了教师模型的预测置信度。对于每个尺度区域对 (w, n) , 基于解耦知识蒸馏范式 [23], UDD 损失可以定义为:

$$\mathcal{L}_{\text{UDD}}(w, n) = (2 + U_{T_t}) \cdot \mathcal{L}_{\text{TCKD}} + (1 - U_{T_t}) \cdot \mathcal{L}_{\text{NCKD}}. \quad (7)$$

这里, 损失成分被定义为: $\mathcal{L}_{\text{TCKD}} = \|\sigma(\psi_{T_t}(w, n)) - \sigma(\psi_S(w, n))\|_2^2$, $\mathcal{L}_{\text{NCKD}} = \|\psi_{T_t}(w, n) - \psi_S(w, n)\|_2^2$ 。 σ 表示 *softmax* 函数。 $U_{T_t} = 1 - \max(\sigma(\psi_{T_t}(w, n))) \in [0, 1]$, 通过测量与 one-hot 分布的偏差来量化预测模糊性。 $(2 + U_{T_t})$ 项放大了对模糊区域 ($U_{T_t} \rightarrow 1$) 的监督, 在这些区域中专家预测往往不可靠, 而 $(1 - U_{T_t})$ 则保持任务无关区域 ($U_{T_t} \rightarrow 0$) 的精确 logit 对齐。

训练损失函数。总训练损失可以描述为:

$$\mathcal{L}_{\text{Total}} = \underbrace{\mathcal{L}_{\text{cls}}}_{\text{Classification}} + \underbrace{\alpha \cdot (\mathcal{L}_{\text{SFA}} + \mathcal{L}_{\text{CFA}})}_{\text{feature alignment}} + \underbrace{\beta \cdot \sum_{w \in W} \sum_{n \in N_w} \mathcal{L}_{\text{UDD}}(w, n)}_{\text{uncertainty-aware decoupled distillation}}, \quad (8)$$

其中 $\mathcal{L}_{\text{cls}}$ 是交叉熵损失, 而 α, β 平衡损失组件。

3 实验

3.1 设置

数据集和度量标准。我们在两个广泛使用的数据集上评估了我们的 UMKD: SICAPv2 [19] 用于前列腺组织学分级, APTOS [12] 用于眼底图像分级。采用了四个基本指标进行性能比较: 总体精度 (OA)、平均精度

(mAcc)、加权 F1 分数 (F1) 和平均绝对误差 (MAE)。**粗体**和 下划线 分别表示在每个子数据集测试中的最佳和次佳性能。

实现细节。我们在两个现实世界的挑战任务上进行所有实验：来源不平衡蒸馏和目标不平衡蒸馏。在来源不平衡蒸馏中，专家模型是在不平衡的源数据集上训练的，并且在类别平衡的目标数据集上进行蒸馏。在目标不平衡蒸馏中，专家模型是在平衡的源数据集上训练的，并且在类别不平衡的目标数据集上进行蒸馏。具体来说，我们通过随机抽样从原始不平衡的数据集中生成平衡子集，包括 SICAPv2-balanced (2500, 2222, 2500, 948) 和 APTOS-balanced (600, 370, 300, 193, 295)，用于每个分级类别。所有数据集的训练、验证和测试集按 8:1:1 的比例划分，并采用随机裁剪和翻转等数据增强技术来扩展数据集。值得注意的是，由于病理图像对颜色变化敏感，因此排除了色彩抖动，因为随机颜色注入可能会破坏病理特征。对于模型训练，我们使用两个在 ImageNet 上预训练的 ResNet50 模型作为专家模型，并一个在 ImageNet 上预训练的 ResNet18 模型作为学生模型。

基线。我们将 UMKD 与多种最具代表性的 SOTA 方法进行了比较，包括基于特征的 (FitNet [18] 和 RKD [16])，以及基于 logits 的 (KD [11]、DKD [23] 和最新的 SDD [13,21])。此外，我们还报告了在每个数据集上训练的 ResNet [10] 的结果作为基准。

3.2 比较结果

SICAPv2 评分结果。我们的 UMKD 在所有指标上都超过了所有以前基于 logits 和基于特征的 KD 基线，达到了新的最先进 (SOTA) 性能，如表 1 所示。

在来源不平衡 KD 任务中，UMKD 达到了最高的整体精度 (OA = 91.02%)、平均精度 (mAcc = 90.23%) 和加权 F1 分数 (F1 = 90.94%)，同时达到了最低的平均绝对误差 (MAE = 0.1294)。具体来说，与 SDD [21] 相比，UMKD 将平均精度提高了 3.57%，所有性能提升均在绿色中突出显示。同样，在目标不平衡 KD 任务中，UMKD 达到了 SOTA 性能，平均精度超过 90.72%。UMKD 在这两项任务中的持续优势突显了其在不同蒸馏设置下的稳健性和泛化能力。

值得注意的是，由于专家模型的知识存在固有偏见（这些模型是在不平衡的数据集上进行训练的），来源不平衡 KD 比目标不平衡任务更具挑战性。我们的 UMKD 明确量化并减轻了这种不平衡偏差，确保了一个更强大和可

表 1: 使用单独训练的 ResNet 模型（顶部）、基于特征的 KD 模型（中部）和 KD 模型（底部）对前列腺癌进行分级。

方法	源不平衡 KD (%)				目标不平衡 KD (%)			
	OA↑	mAcc↑	F1 ↑	MAE ↓	OA↑	mAcc↑	F1 ↑	MAE↓
Resnet50 (<i>Exp₁</i>)	91.53	89.48	91.47	0.1098	89.19	89.44	89.13	0.1332
Resnet50 (<i>Exp₂</i>)	92.05	89.88	91.93	0.1100	89.71	89.78	89.61	0.1322
Resnet18 (<i>Stu</i>)	89.58	89.06	89.49	0.1463	90.36	88.11	90.23	0.1318
FitNet [18]	78.78	78.41	78.42	0.3136	79.06	55.77	77.12	0.3694
RKD [16]	88.54	88.24	88.44	0.1690	90.79	88.30	90.67	0.1369
KD [11]	<u>89.06</u>	<u>88.39</u>	<u>88.97</u>	0.1624	<u>90.97</u>	<u>89.44</u>	<u>90.90</u>	<u>0.1368</u>
DKD [23]	86.91	85.01	86.68	0.1739	89.19	87.11	89.12	0.1476
SDD [21]	87.82	86.66	87.67	<u>0.1594</u>	89.93	88.81	89.85	0.1447
UMKD	91.02	90.23	90.94	0.1294	91.75	90.72	91.72	0.1199
△	+3.20	+3.57	+3.27	+0.0300	+1.82	+1.91	+1.87	+0.0248

靠的蒸馏过程。总之，我们的 UMKD 不仅减少了模型参数的数量（ResNet18 对比 ResNet50），还显著提升了模型性能，使其成为提高前列腺癌分级和诊断准确性和可靠性的一种有前景的工具。

APTOS 分级的结果。我们评估了 UMKD 在更具挑战性的 APTOS 数据集上的性能，该数据集中数据更加不平衡，并且专家标注存在偏差，这是由于视网膜图像固有的复杂性所致。

如表 2 所示，UMKD 在来源不平衡和目标不平衡的 KD 任务中相较于现有方法取得了更优的表现。具体而言，在来源不平衡的 KD 任务中，我们提出的 UMKD 持续优于所有基准模型，达到了最高的整体准确率（OA = 74.61%）、平均准确率（mAcc = 67.33%）和加权 F1 分数（F1 = 74.43%），同时获得了最低的平均绝对误差（MAE = 0.3589）。这些结果证明了 UMKD 量化专家模型因训练数据不平衡而产生的预测偏差的能力。通过利用专家模型输出的不确定性，学生模型可以自适应地调整知识转移的权重，从而确保一个更稳健和可靠的蒸馏过程，并显著提升模型性能。

在目标不平衡蒸馏任务中，UMKD 相比所有基于 logits 的方法实现了 SOTA 性能，达到了最高的平均准确率（mAcc = 74.38%），并在其他指标

表 2: 使用单独训练的 ResNet 模型（顶部）、基于特征的 KD 模型（中间）和 KD 模型（底部）对眼底图像进行分级。

方法	来源不平衡的 KD (%)				目标不平衡 KD (%)			
	OA↑	mAcc↑	F1 ↑	MAE ↓	OA↑	mAcc↑	F1 ↑	MAE↓
Resnet50 (<i>Exp</i> ₁)	82.34	66.33	80.63	0.2478	72.66	63.74	72.18	0.4001
Resnet50 (<i>Exp</i> ₂)	82.81	65.77	81.89	0.2421	72.66	63.74	72.51	0.3936
Resnet18 (<i>Sup</i>)	74.21	67.18	73.94	0.4192	82.34	70.56	81.04	0.2521
FitNet [18]	67.57	59.12	66.57	0.5704	79.06	55.77	77.12	0.3694
RKD [16]	<u>74.61</u>	<u>67.15</u>	<u>74.37</u>	0.4203	85.00	69.75	84.38	<u>0.2482</u>
KD [11]	73.04	66.07	72.77	0.4448	83.43	71.56	83.50	0.2578
DKD [23]	70.70	61.74	69.83	0.4018	81.87	73.95	82.43	0.2743
SDD [21]	73.44	65.07	72.62	<u>0.4017</u>	83.12	73.55	82.83	0.2662
UMKD	74.61	67.33	74.43	0.3589	<u>83.91</u>	74.38	<u>84.03</u>	0.2476
△	+1.17	+2.26	+1.81	+0.0428	+0.79	+0.83	+1.20	+0.0186

上也保持了强劲的表现 (OA = 83.91%, F1 = 84.03%)。如第 7 行所示, 在 OA 和 F1 指标上, UMKD 相比基于特征的 RKD [16] 得到了次优结果, 但在 mACC 上显著优于 RKD (74.38% 对 69.75%)。我们将其归因于以下原因: 首先, RKD 优化了角动量以计算样本三元组之间的 mini-batch 距离, 鼓励同一类别的样本聚得更近。虽然这种方法实现了最高的整体准确率, 但由于批次内固有的类别不平衡问题, 在目标不平衡蒸馏任务中它倾向于偏向多数类别。相比之下, UMKD 通过显式地测量并使用不确定性来缓解这种不平衡, 从而在整体准确率和平均准确率之间取得了更好的平衡。

3.3 消融研究

我们进行了一项消融研究, 以分析 UMKD 三个组件的贡献。在 SICAPv2 数据集上的结果, 如表 3 所示, 表明所有组件对于实现高表现均至关重要, 在源不平衡和目标不平衡分级任务中都是如此。缺少 SFA 和 CFA 组件会导致 UMKD 性能显著下降。这是因为任务无关的结构特征和任务特定的语义特征没有得到有效分离。在疾病图像分级中, 病理特征通常集中在特定的图像区域内部。仅依赖全局特征可能会掩盖这些关键分类相关特征, 降低模型准确评估疾病的性能。

表 3: SFA、CFA 和 UDD 模块在 SICAPv2 数据集上的消融研究。

方法			来源不平衡的 KD (%)				目标不平衡 KD (%)			
SFA	CFA	UDD	OA↑	mAcc↑	F1 ↑	MAE ↓	OA↑	mAcc↑	F1 ↑	MAE↓
			87.82	86.66	87.67	0.1594	89.93	88.81	89.85	0.1447
	✓	✓	90.69	89.92	90.58	0.1314	91.62	90.54	91.58	0.1258
✓		✓	90.36	89.68	90.25	0.1355	91.44	90.93	91.40	0.1275
✓	✓		88.15	86.84	87.97	0.1566	90.06	89.15	90.03	0.1416
✓	✓	✓	91.02	90.23	90.94	0.1294	91.75	90.72	91.72	0.1199

类似地，没有 UDD，学生模型无法动态调整知识转移权重，导致最差的蒸馏性能。这突显了不确定性感知机制在减轻偏差传播和确保稳健知识转移方面的重要性。值得一提的是，我们在 APTOS 数据集上的消融实验得出了一致的结论，但由于篇幅限制未予报告。

4 结论

我们提出了一种 UMKD 框架来解决分级任务中的数据不平衡问题。通过集成频域 SFA 和空域 CFA 模块，UMKD 有效地将与任务无关的结构特征与特定任务的语义特征解耦。UDD 机制进一步增强了鲁棒性，通过基于专家知识不确定性动态调整知识转移权重，并减轻了由数据不平衡和模型异质性引起的偏差。在视网膜和前列腺癌数据集上的广泛实验表明，UMKD 在源域不平衡和目标域不平衡场景中均实现了最先进的性能。这些结果突显了 UMKD 确保可靠且公平的知识转移的能力，使其成为现实世界疾病图像分级的实际且可扩展的解决方案。未来的工作将重点扩展 UMKD 至其他医学成像任务并提高其临床应用的可解释性。

参考文献

1. Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Van Boven, H., Vink, R., et al.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine **28**(1), 154–163 (2022)

2. Cheng, Y., Ying, H., Hu, R., Wang, J., Zheng, W., Zhang, X., Chen, D., Wu, J.: Robust image ordinal regression with controllable image generation. arXiv preprint arXiv:2305.04213 (2023)
3. Claudio Quiros, A., Coudray, N., Yeaton, A., Yang, X., Liu, B., Le, H., Chiriboga, L., Karimkhan, A., Narula, N., Moore, D.A., et al.: Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unannotated pathology slides. *Nature Communications* **15**(1), 4596 (2024)
4. Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., Liu, R., Wang, X., Hou, X., Liu, Y., et al.: A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nature communications* **12**(1), 3242 (2021)
5. Gao, S., Fu, Y., Liu, K., Gao, W., Xu, H., Wu, J., Han, Y.: Collaborative knowledge amalgamation: Preserving discriminability and transferability in unsupervised learning. *Information Sciences* **669**, 120564 (2024)
6. Gao, S., Fu, Y., Liu, K., Han, Y.: Contrastive knowledge amalgamation for unsupervised image classification. In: *International Conference on Artificial Neural Networks*. pp. 192–204. Springer (2023)
7. Gao, S., Fu, Y., Liu, K., Xu, H., Wu, J.: Ka² er: Knowledge adaptive amalgamation of experts for medical images segmentation. arXiv preprint arXiv:2410.21085 (2024)
8. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. *Medical Image Analysis* **89**, 102889 (2023)
9. Hao, Z., Guo, J., Han, K., Tang, Y., Hu, H., Wang, Y., Xu, C.: One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems* **36** (2024)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
12. Karthik, Maggie, Dane, S.: Aptos 2019 blindness detection. <https://kaggle.com/competitions/aptos2019-blindness-detection> (2019), kaggle
13. Li, L., Li, X.C., Ye, H.J., Zhan, D.C.: Enhancing class-imbalanced learning with pre-trained guidance through class-conditional knowledge distillation. In: *Forty-first International Conference on Machine Learning* (2024)

14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60–88 (2017)
15. Mohan, N.J., Murugan, R., Goel, T., Roy, P.: Drfl: federated learning in diabetic retinopathy grading using fundus images. *IEEE Transactions on Parallel and Distributed Systems* **34**(6), 1789–1801 (2023)
16. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3967–3976 (2019)
17. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al.: Idrid: Diabetic retinopathy–segmentation and grading challenge. *Medical image analysis* **59**, 101561 (2020)
18. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* (2014)
19. Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V.: Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine* **195**, 105637 (2020)
20. Wang, J., Cheng, Y., Chen, J., Chen, T., Chen, D., Wu, J.: Ord2seq: Regarding ordinal regression as label sequence prediction. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5865–5875 (2023)
21. Wei, S., Luo, C., Luo, Y.: Scaled decoupled distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15975–15983 (2024)
22. Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., Yu, S.: A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis* **69**, 101985 (2021)
23. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. pp. 11953–11962 (2022)