域适应
用于溃疡性结肠炎严重程度估计
使用患者级诊断

Takamasa Yamaguchi[1], Brian Kenji Iwana[1], Ryoma Bise[1], Shota Harada[1], Takumi Okuo[1], Kiyohito Tanaka[2], and Kaito Shiku[1]

[1] Kyushu University, Japan
[2] Kyoto Second Red Cross Hospital, Japan
takamasa.yamaguchi@human.ait.kyushu-u.ac.jp

摘要 溃疡性结肠炎（UC）严重程度估计方法的发展具有重要意义。然而，这些方法通常会受到不同医院成像设备和临床环境差异导致的领域转移的影响。尽管已经提出了几种域适应方法来解决领域转移问题，但它们仍然难以应对目标领域的监督不足或标注成本高的问题。为了解决这些问题，我们提出了一种新颖的弱监督域适应方法，该方法利用 UC 诊断中常规记录的患者级诊断结果作为目标领域的弱监督信息。所提方法通过共享聚合标记和最大严重性三元组损失对跨领域内的类别分布进行对齐，从而利用了患者级别的诊断是由每个患者的最严重的区域决定这一特性。实验结果显示，我们的方法优于对比的域适应方法，在领域转移场景下提高了 UC 严重程度估计的准确性。

Keywords: 溃疡性结肠炎 · 弱监督领域适应

1 介绍

溃疡性结肠炎（UC）的诊断对于确定适当的临床治疗至关重要。在当前的 UC 诊断协议中，患者会接受结肠内镜成像，从不同位置拍摄大约 20 到 40 张图像。根据显示病变最严重的那张图像记录 UC 的严重程度分数，尽管在诊断过程中会对所有图像进行审查 [11, 16] 。较高的分数表示更严重的病情。重要的是，单个图像的严重程度没有被记录下来。然而，人们越来越认识到可视化结肠不同区域的严重性分布的重要性，而不仅仅是关注最严重的区域 [16] 。因此，迫切需要自动方法来评估诊断过程中拍摄的每个图像的严重性，并且可视化整个结肠的严重性分布。

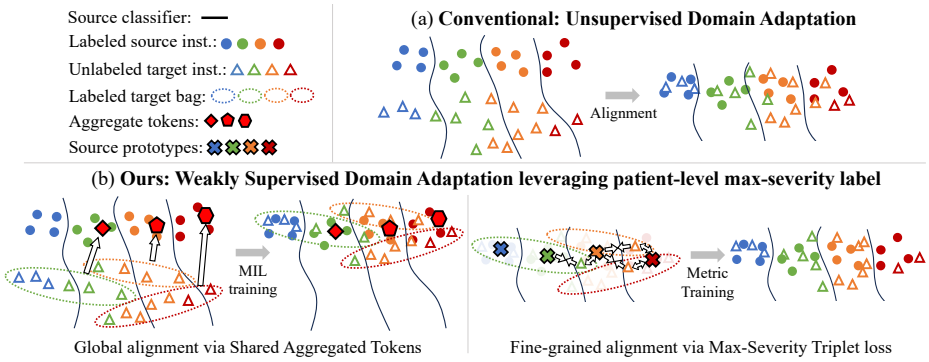

图 1: (a) 传统的无监督领域适应方法，由于目标域中缺乏监督信息而无法实现跨领域的类别对齐。(b) 所提出的弱监督领域适应方法，通过利用患者级别的最大严重性标签，并采用共享聚合标记策略和最大严重性三元组损失实现了类别对齐。

各种在图像层面和患者层面对严重程度进行自动评估的方法已经被探索。对于患者层面的严重性估计，已经将诸如 K-排序算法等顺序分类技术整合到多示例学习（MIL）[13, 14] 中。然而，这些方法无法估计图像级别的严重度。对于图像级别严重性的分类，已经提出了许多处理顺序类别的方法 [4, 5, 15, 17, 18, 9]。例如，Kadota 等人。[4] 提出了一种同时学习回归和排序任务的图像层面严重性评估方法。这些方法假设测试数据集与源数据具有相同的分布。

然而，这些方法遇到了一个重大挑战：领域迁移。当源（训练）域和目标（测试）域的数据分布存在差异时就会发生领域迁移，这种差异通常是由成像设备或医院条件的变化引起的。这会导致在将针对源域数据训练的模型应用于目标域数据时性能下降。

为了解决领域迁移问题，无监督领域适应（UDA）[19, 20] 已被广泛研究。UDA 假设目标域中没有标记数据，并旨在仅使用源域的标记数据来对齐源域和目标域之间的整体数据分布。然而，如图 1(a) 所示，由于 UDA 只关注对齐整体分布，因此难以在领域内对齐类别级分布。另一种方法，半监督领域适应（SSDA）[12, 6, 21]，则利用少量目标域的标记样本来对齐类别级分布。虽然 SSDA 比 UDA 提高了性能，但它仍然需要额外的标注，这通常因需要专业知识而产生高昂的成本。

在本文中，我们提出了一种新颖的弱监督领域适应问题，该问题利用实际临床环境中常规记录的患者级诊断信息作为目标领域的弱监督。如上所述，代表每位患者捕获图像中最严重分数（“最严重程度标签”）的患者级别标签通常会作为诊断记录的一部分存储，并且可以在不需要任何额外注释的情况下加以利用。据我们所知，在溃疡性结肠炎严重程度估计的背景下，尚未探索基于患者级诊断的领域适应。

为了充分利用仅在图像集级别可用的患者级最大严重性标签，并在不同领域对齐类别特征分布，我们提出了一种新方法，该方法通过共享聚合令牌策略进行全局分布对齐并通过最大严重性三元组损失进行细粒度对齐。共享聚合令牌策略利用了最初用于 MIL [14] 中患者级严重性估计的聚合令牌来进行领域对齐。如图 1(b) 左侧所示，这些训练于源域以预测最大标签并捕捉图像级别类别分布的聚合令牌在目标域的最大标签预测训练期间被冻结和重复使用，从而鼓励目标域中的类别分布接近源域的分布，并实现粗略对齐。如图 1(b) 右侧所示，最大严重性三元组损失利用了最大标签的一个特性——在一个包内不存在比该最大标签更严重的图像，并且惩罚那些被错误预测为比最大标签更严重的类别的目标域中的图像，鼓励它们的分布与源域中相应的类别对齐。

在使用两个内窥镜图像数据集的实验中，我们确认了所提出的方法通过利用患者级别的诊断结果，在性能上优于传统的领域适应方法。此外，所提出的方法甚至超过了那些在数据集的 5% 上使用额外标注的半监督方法，尽管它不需要任何额外的标注。

2 弱监督域适应利用患者级别的诊断

2.1 问题设定

利用患者级别的诊断进行弱监督领域适应被公式化为在顺序多实例学习 (MIL) 上的领域适应，其中两个领域的数据集作为训练数据提供。源数据，表示为 $\mathcal{D}^{\mathrm{s}} = \{\mathcal{B}_i^{\mathrm{s}}, Y_i^s\}$，提供了实例级别（即，图像）和袋级别（即，患者）的标签。每个包 $\mathcal{B}_i^{\mathrm{s}} = \{\boldsymbol{x}_{i,j}^s, y_{i,j}^{\mathrm{s}}\}_{j=1}^{|\mathcal{B}_i^{\mathrm{s}}|}$ 被定义为一组 $|\mathcal{B}_i^{\mathrm{s}}|$ 个实例 $\boldsymbol{x}_{i,j}^{\mathrm{s}}$ 和实例标签 $y_{i,j}^{\mathrm{s}}$。目标数据，记作 $\mathcal{D}^{\mathrm{t}} = \{\mathcal{B}_i^{\mathrm{t}}, Y_i^{\mathrm{t}}\}$，在包级别提供了标签，而目标数据中每个包的实例级标签 $\mathcal{B}_i^{\mathrm{t}} = \{\boldsymbol{x}_{i,j}^{\mathrm{t}}\}_{j=1}^{|\mathcal{B}_i^{\mathrm{t}}|}$ 则未提供。本文的目标是训练一个分类模型，利用这些训练数据估计目标域中的实例级标签 $\hat{y}_{i,j}^{\mathrm{t}}$。

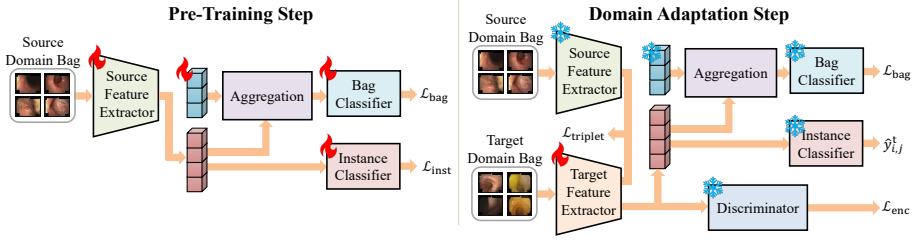

图 2: 所提方法概述

在两个领域中，包级和实例级标签都是用第 K 类定义的，其中 $Y_i, y_{i,j} \in \{1, 2, \ldots, K\}$ 每个类都有一个顺序关系 $K \succ K-1, ..., \succ 1$。包级标签 Y_i 是由包中所有实例中严重程度最高的类别定义的。

2.2 序数多实例学习中的领域适应

所提出的方法分为两个步骤进行训练，如图 2所示。第一步涉及使用源实例级和包级标签对源特征提取器、聚合标记和源分类器进行预训练。第二步通过全局分布对齐（借助共享聚合标记）及细粒度分布对齐（利用最大严重性三元组损失函数），对目标特征提取器进行训练，以使目标数据的领域级和类别级分布与源数据相一致。

预训练步骤。此步骤旨在通过实例级和包级分类训练来训练源特征提取器 $g^{\text{s}}\text{inst}$、源实例和包分类器 $f^{\text{s}}\text{inst}$ 和 $f^{\text{s}}_{\text{bag}}$，以及聚合标记 $\mathcal{T} = \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{K-1}$。

图像级分类使用实例级标签，采用标准的监督顺序分类方法进行，例如 k-rank 方法 [1, 7]。在包级分类训练中，受 [14] 提出的用于患者级别严重程度估计的顺序 MIL 方法启发，通过对使用聚合标记 $\mathcal{T}$ 获得的包表示进行分类。聚合标记 $\mathcal{T}$ 通过测量自身与实例特征之间的相似性来计算实例在包内的注意力分数，并通过加权聚合生成包级表示。每个聚合标记设计用于汇总对应特定严重程度级别的实例特征。第 k 个聚合标记，$\boldsymbol{a}_k$，训练以对严重程度级别为 k 的实例分配高注意力。通过这种训练，每个聚合标记都与源域中的类别分布保持一致。

跨域全局分布对齐。此步骤的目的是训练目标特征提取器，使得源域和目标域之间的领域级和类别级数据分布大致对齐。在此步骤中，源特征提取器 $g^{\text{s}}_{\text{inst}}$ 和分类器 $f^{\text{s}}_{\text{inst}}$ 及 $f^{\text{s}}_{\text{bag}}$ 被冻结，仅训练目标特征提取器 $g^{\text{t}}_{\text{inst}}$。

首先，为了对齐领域内的分布，采用了对抗学习。一个领域判别器 $d(\cdot)$ 用于判断数据来自源域还是目标域：$\mathcal{L}_{\text{disc}} = -\sum_{j=1}^{|\mathcal{B}^{\text{s}}_i|} \log(d(\boldsymbol{e}^{\text{s}}_{i,j})) - \sum_{j=1}^{|\mathcal{B}^{\text{t}}_i|} (\log(1 -$

$d(\boldsymbol{e}^{\mathrm{t}}_{i,j})))$。这里，$\boldsymbol{e}^{s}_{i,j}$ 和 $\boldsymbol{e}^{t}_{i,j}$ 分别表示通过相应实例级特征提取器获得的源域和目标域中的实例特征向量。具体地，$\boldsymbol{e}^{s}_{i,j} = g^{\mathrm{s}}_{\mathrm{inst}}(\boldsymbol{x}^{\mathrm{s}}_{i,j})$ 和 $\boldsymbol{e}^{\mathrm{t}}_{i,j} = g^{\mathrm{t}}_{\mathrm{inst}}(\boldsymbol{x}^{\mathrm{t}}_{i,j})$。接下来，训练目标特征提取器使其能够被误认为是源域使用对抗编码损失：$\mathcal{L}_{\mathrm{enc}} = -\sum_{j=1}^{|\mathcal{B}^{\mathrm{t}}_i|} \log(d(\boldsymbol{e}^{\mathrm{t}}_{i,j}))$。

虽然对抗学习可以在一定程度上对齐域级分布，但在类级别上对齐分布仍然具有挑战性。为了仅使用目标域中的包级标签实现跨域实例的类级别对齐，我们提出了一种共享聚合令牌策略，旨在基于预先训练以捕捉源域中实例类别分布的聚合令牌来对齐目标域中的实例特征。具体来说，在目标域中训练包分类时，冻结在源域上预训练的聚合令牌 $\mathcal{T}$ 的参数，并且仅更新目标域 $g^{\mathrm{t}}_{\mathrm{inst}}$ 的特征提取器。因此，为了正确地对目标域中的包进行分类，实例特征必须与源域的类级别分布对齐，以便固定的聚合令牌可以为目标实例计算准确的关注分数。共享聚合令牌策略通过目标域上的包级分类损失 $\mathcal{L}_{\mathrm{bag}}$ 进行优化。

细粒度分布对齐与最大严重性三元组损失。本模块的目的是在域之间详细对齐类别分布；然而，由于目标域中没有实例标签，实现每种类别的这种对齐是具有挑战性的。为了解决这一问题，本模块的关键思想是通过利用最大严重性标签的性质来对齐实例级别的类别分布，该性质表明在一个包内仅存在标签不超过最大标签的实例。为此，我们提出了最大严重性三元组损失，它由三个部分组成：一个基准点、一个正样本和一个负样本。为了选择正样本和负样本，我们使用带有实例级别严重性标签的源数据为每个类别构建原型 $\{\boldsymbol{p}^{\mathrm{s}}_k\}_{k=1}^{K}$。每个原型计算为属于类 k 的实例特征的平均值：$\boldsymbol{p}^{\mathrm{s}}_k = \frac{1}{J_k}\sum_{j=1}^{J_k} \boldsymbol{e}^{\mathrm{s}}_{k,j}$，其中 J_k 是类 k 中源实例的数量，而 $\boldsymbol{e}^{\mathrm{s}}_{k,j}$ 是类 k 的实例特征。

目标域中预测严重性高于袋标签 Y^{t}_i 的实例被认为是与源域类别分布不一致的数据。因此，我们将其用作锚点并应用三元组损失。在这个过程中，正样本和负样本分别选为对应于 $y^{\mathrm{s}}_{i,j}$ 同等严重性和比 $y^{\mathrm{s}}_{i,j}$ 更高严重性的源原型。三元组损失 $\mathcal{L}_{\mathrm{triplet}}$ (1) 确保锚实例与正例 $\boldsymbol{p}^{\mathrm{s}}_{+}$ 的距离小于到负例 $\boldsymbol{p}^{\mathrm{s}}_{-}$ 的距离，或者：

$$\mathcal{L}_{\mathrm{triplet}}(\mathcal{B}^{\mathrm{t}}_i, \boldsymbol{p}^{\mathrm{s}}_{+}, \boldsymbol{p}^{\mathrm{s}}_{-}) = \sum_{j=1}^{|\mathcal{B}^{\mathrm{t}}_i|} \max(||\boldsymbol{e}^{\mathrm{t}}_{i,j} - \boldsymbol{p}^{\mathrm{s}}_{+}|| - ||\boldsymbol{e}^{\mathrm{t}}_{i,j} - \boldsymbol{p}^{\mathrm{s}}_{-}|| + \xi, 0), \quad (1)$$

$$\text{s.t. } Y^{\mathrm{t}}_i \leq K-1, Y^{\mathrm{t}}_i < \hat{y}^{\mathrm{t}}_{i,j}, \quad (2)$$

其中 $||\cdot||$ 表示欧几里得距离，ξ 是边界。

然后，总损失通过超参数 α 控制三元组损失的权重进行计算，如下所示：

$$\mathcal{L}_{\text{target}} = \mathcal{L}_{\text{bag}} + \mathcal{L}_{\text{enc}} + \alpha\mathcal{L}_{\text{triplet}}. \tag{3}$$

3 实验结果

数据集。为了验证我们提出的方法的有效性，我们使用了两个数据集（**极限值**）[10] 和一个私有数据集（**私有**）作为不同的领域。**极限不存在**包含从 564 名患者收集的 11,276 张图像，每个包中的图像数量从 1 到 105 不等，平均每个包有 20 张图像。该数据集具有图像级标签，并允许根据每位患者的最严重区域获得患者级别的最大严重程度标签。**私有**是从一家匿名医院收集的，它记录了在常规诊断过程中针对每位患者的特定的最大严重程度标签。在这里，在本实验中，为了评估目的对图像进行了额外的标注。每个图像级别都标注了一个代表从 0 级严重性到 3 级严重性的诊断标签。在这两个数据集中，作为目标领域，仅使用患者级别的标签进行训练，排除了图像级别的标签。

实现细节。对于特征提取器，我们使用了在 ImageNet 数据集上预训练的 ResNet18 [3] [2]。提出的网络使用 Adam 优化器进行优化 [8]。在预训练步骤中，特征提取器和实例分类器的学习率分别为 3e-6，袋分类器为 1e-5，共进行了 1,500 个周期，每批大小为 16，并且设置了 100 次的早停条件。为了处理类别不平衡问题，每个步骤都基于实例级别和袋标签的数量进行过采样。在领域适应步骤中，判别器和特征提取器的学习率分别为 1e-4 和 1e-6，固定了 150 个周期，每批大小为 16，并且在 0.01 的 $\mathcal{L}_{\text{target}}$ 中有 α。

评估指标。为了评估我们提出的方法的性能，我们采用了三个指标：分类准确率、对数据不平衡具有鲁棒性的 Macro-F1，以及考虑顺序关系的 Quadratic Weighted Kappa (Kappa) [9]。使用了 5 折交叉验证进行评估。

比较评估。我们比较了所提出的方法与几种领域自适应方法，包括两种 UDA 方法，ADDA [19] 和 DANN [20]，以及三种 SSDA 方法，MME [12]，CDAC [6]，和 S^3D [21]。UDA 方法通过对抗学习对齐整体数据分布，而不使用目标标签，而 SSDA 方法则利用目标域中的一些标记数据。需要注意的是，SSDA 方法在目标域中使用额外的监督标签，而提出的方法则是通过利用常规记录的患者级别的诊断记录进行无注释成本的训练。在这个实验中，SSDA 方法使用每个类别中的 1%、3%和 5%的图像作为标记训练数据。

表 1展示了与比较方法的对比结果。由于目标域中缺乏类别信息，ADDA 和 DANN 难以在不同领域之间对齐类别分布。因此，它们无法达到足够的

表 1: 与 UDA 和 SSDA 方法的比较。“目标标签”表示目标域中标签的类型。“实例标签比例”表示额外标注标签的比例。最佳结果在**加粗**中。

Target Label	Instance Label Ratio	Method	LIMUC 至私有化			私有到 LIMUC		
			Accuracy	Kappa	Macro-F1	Accuracy	Kappa	Macro-F1
无监督	0%	ADDA	0.521	0.575	0.364	0.604	0.645	0.529
		DANN	0.429	0.500	0.352	0.560	0.446	0.423
半监督	1%	MME	0.610	0.593	0.469	0.670	0.695	0.541
		CDAC	0.588	0.545	0.446	0.593	0.571	0.506
		S^3D	0.653	0.641	0.507	0.666	0.684	0.522
	3%	MME	0.588	0.550	0.476	0.679	0.727	0.570
		CDAC	0.670	0.643	0.512	0.607	0.607	0.532
		S^3D	0.674	0.655	0.537	0.677	0.726	0.567
	5%	MME	0.608	0.593	0.496	0.697	0.753	**0.595**
		CDAC	0.651	0.628	0.522	0.633	0.655	0.562
		S^3D	0.668	0.652	0.534	0.692	0.738	0.586
Patient-level labels	0%	Ours	**0.714**	**0.746**	**0.603**	**0.706**	**0.787**	0.594

表 2: 每个模块的有效性。最佳结果在**加粗**中。

方法	高级的	聚合。标记	三元组	LIMUC 私有化 →			私有 →LIMUC		
				Accuracy	Kappa	Macro-F1	Accuracy	Kappa	Macro-F1
Ours	✓	✓	✓	**0.714**	**0.746**	**0.603**	**0.706**	**0.787**	0.594
w/o Triplet	✓	✓		0.658	0.703	0.576	0.705	**0.787**	**0.605**
w/o Triplet&AT	✓			0.521	0.574	0.363	0.604	0.645	0.529
w/o Adv&Triplet&AT				0.237	0.300	0.212	0.580	0.397	0.344

性能表现。MME、CDAC 和 S^3D 由于目标域中存在的额外实例信息，其性能高于 UDA 方法。然而，整个训练数据集的 5%大约对应 400 张图像，需要极高的标注成本由医学专家承担。相比之下，所提出的方法利用常规记录的患者级别的诊断记录，在 LIMUC 到私有实验中达到了所有方法中的最高表现。此外，在私有到 LIMUC 实验中，它也实现了最高的准确率和 kappa 分数，同时在 Macro-F1 性能上与使用 5%标注数据的最佳方法相当。这些结果表明，所提出的方法可以利用不需要标注成本的患者级别的诊断记录作为弱标签，并对齐跨域的类别分布。

每个模块的有效性。我们进行了一项消融研究，通过将我们的方法与三种消融方法比较来评估共享聚合标记和最大严重性三元组损失的有效性。表 2显示了消融研究的结果。我们将所提出的方法与没有三元组损失（无 Triplet）、没有三元组损失和共享聚合标记（无 Triplet&AT），以及一种未使用三元组损失、共享聚合标记和对抗学习训练的方法（无 Adv&Triplet&AT）进行了

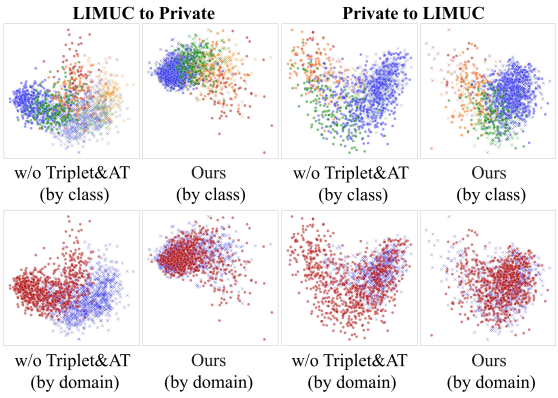

图 3: 使用 PCA 对目标域在提出方法前后的特征空间进行可视化。在上一行中，类别，严重程度 0、1、2 和 3，分别用蓝色、绿色、橙色和红色表示。在下一行中，源域用蓝色表示，目标域用红色表示。

比较，即仅在源域上进行训练并在目标域上测试。结果显示，虽然对抗学习提高了性能，但单独使用是不够的。通过额外引入所提出的共享聚合标记和最大严重性损失以实现跨领域的类别分布对齐，性能显著提升。

为了证明所提出的方法能够在类别级别对齐跨域的特征分布，图 3使用主成分分析（PCA）可视化了数据分布。这些结果表明，仅使用对抗学习训练的模型（无三元组&AT）无法有效对齐跨域的类别级分布，而提出的方法，结合共享聚合标记和最大严重性损失，则实现了改进的类别级分布对齐。

4 结论

在这项研究中，我们提出了一种用于 UC 严重性估计的弱监督领域适应方法，该方法利用常规记录的患者级诊断结果作为目标域中的弱监督标签。为了使用患者级最大严重程度诊断来对齐跨领域的类别分布，我们提出了一种新颖的方法，包括共享聚合标记和最大严重程度三元组损失。实验结果证实了所提方法的有效性。此外，与需要额外标注的半监督方法相比，我们的方法在不产生额外标注成本的情况下实现了高性能。

致谢: 本工作部分得到了 KAKEN JP23K16949, JP2318509, SIP JPJ012425 和 ASPIRE JPMJAP2403 的支持。

参考文献

1. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters **140**, 325–331 (2020)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
4. Kadota, T., Abe, K., Bise, R., Kawamura, T., Sakiyama, N., Tanaka, K., Uchida, S.: Automatic estimation of ulcerative colitis severity by learning to rank with calibration. IEEE Access **10**, 25688–25695 (2022)
5. Kadota, T., Hayashi, H., Bise, R., Tanaka, K., Uchida, S.: Deep bayesian active-learning-to-rank for endoscopic image data. In: MIUA. pp. 609–622 (2022)
6. Li, J., Li, G., Shi, Y., Yu, Y.: Cross-domain adaptive clustering for semi-supervised domain adaptation. In: CVPR. pp. 2505–2514 (2021)
7. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output cnn for age estimation. In: CVPR. pp. 4920–4928 (2016)
8. P, K.D., Jimmy, B.: Adam: A method for stochastic optimization (2014)
9. Polat, G., Ergenc, I., Kani, H.T., Alahdab, Y.O., Atug, O., Temizel, A.: Class distance weighted cross-entropy loss for ulcerative colitis severity estimation. In: MIUA. pp. 157–171 (2022)
10. Polat, G., Kani, H.T., Ergenc, I., Ozen Alahdab, Y., Temizel, A., Atug, O.: Improving the computer-aided estimation of ulcerative colitis severity according to mayo endoscopic score by using regression-based deep learning. Inflammatory Bowel Diseases **29**(9), 1431–1439 (2023)
11. Raimundo Fernandes, S., Santos, P.M., Moura, C.M., Marques da Costa, P., Carvalho, J.R., Valente, A.I., Baldaia, C., Gonçalves, A.R., Moura Santos, P., Araújo-Correia, L., et al.: The use of a segmental endoscopic score may improve the prediction of clinical outcomes in acute severe ulcerative colitis. Revista Española de Enfermedades Digestivas **108**(11), 697–702 (2016)
12. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. ICCV (2019)
13. Schwab, E., Cula, G.O., Standish, K., Yip, S.S., Stojmirovic, A., Ghanem, L., Chehoud, C.: Automatic estimation of ulcerative colitis severity from endoscopy videos using ordinal multi-instance learning. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **10**(4), 425–433 (2022)

14. Shiku, K., Nishimura, K., Suehiro, D., Tanaka, K., Bise, R.: Ordinal multiple-instance learning for ulcerative colitis severity estimation with selective aggregated transformer. WACV (2025)

15. Stidham, R.W., Liu, W., Bishu, S., Rice, M.D., Higgins, P.D., Zhu, J., Nallamothu, B.K., Waljee, A.K.: Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. JAMA Network Open **2**(5), e193963–e193963 (2019)

16. Takabayashi, K., Kobayashi, T., Matsuoka, K., Barrett, L.G., Kawamura, T., Tanaka, K., Kadota, T., Bise, R., Uchida, S., Kanai, T., et al.: Development of artificial intelligence to represent quantitative inflammation assessment of ulcerative colitis diagnosed by inflammatory bowel disease expert endoscopists. SSRN (2022)

17. Takenaka, K., Ohtsuka, K., Fujii, T., Negi, M., Suzuki, K., Shimizu, H., Oshima, S., Akiyama, S., Motobayashi, M., Nagahori, M., et al.: Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. Gastroenterology **158**(8), 2150–2157 (2020)

18. Takezaki, S., Tanaka, K., Uchida, S., Kadota, T.: Disease Severity Regression with Continuous Data Augmentation. In: ISBI. pp. 1–5 (2023)

19. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR. pp. 7167–7176 (2017)

20. Yaroslav, G., Evgeniya, U., Hana, A., Pascal, G., Hugo, L., François, L., Mario, M., Victor, L.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**, 1–35 (2016)

21. Yoon, J., Kang, D., Cho, M., of Science, P.U., Technolgy(POSTECH), Korea, S.: Semi-supervised domain adaptation via sample-to-sample self-distillation. In: WACV (2022)