arXiv:2301.00973v2 [cs.CV] 8 Jun 2024

从眼底图像检测糖尿病视网膜病变严重程度：基于变换器网络的综述

Tejas Karkera[a]、Chandranath Adak[b,*]、Soumi Chattopadhyay[c]、Muhammad Saqib[d]

[a]美国东北大学，波士顿，*02115*，美国[b]印度理工学院巴特那分校计算机科学与工程系，比哈尔邦，*801106*，印度[c]印度理工学院印多尔分校计算机科学与工程系，中央邦，*453552*，印度[d]*Data61*，澳大利亚联邦科学与工业研究组织，新南威尔士州，*2122*，澳大利亚

摘要

糖尿病视网膜病变（糖尿病视网膜病变）被认为是全球重要的健康问题之一，主要因为它会导致许多糖尿病患者视力丧失。眼科医生通常通过眼底摄影获得的视网膜图像，手动评估糖尿病视网膜病变的严重程度。本文探讨对糖尿病视网膜病变严重程度进行自动化理解。文献中的研究者主要采用传统的机器学习算法和卷积架构来实现这一自动化。然而，过往研究鲜少关注视网膜图像中的关键区域来提升模型性能。本研究采用并微调基于变换器的学习模型，以提取视网膜图像中的关键特征，从而更细致地理解糖尿病视网膜病变的严重程度。此外，我们还探究了图像变换器根据眼底照片推断糖尿病视网膜病变严重程度的效果。实验采用公开的 APTOS-2019 失明检测数据集，基于 Transformer 的模型表现颇为理想。

关键词：失明检测，糖尿病视网膜病变，深度学习，变换器网络

*通讯作者电子邮箱：tejkar10@gmail.com (Tejas Karkera), adak32@gmail.com (Chandranath Adak),soumi61@gmail.com (Soumi Chattopadhyay), saqib.uet1@gmail.com (Muhammad Saqib)

Neurocomputing，爱思唯尔（已接收） *2024*年*6*月*11*日

1. 引言

糖尿病（Diabetes Mellitus），通常称为糖尿病，是一种患者血糖水平长期升高的疾病。糖尿病视网膜病变（DR）是糖尿病相关的微血管并发症，会造成视网膜血管损伤，若不及时治疗，可能导致视力受损甚至失明 [1, 2]。研究估计，约 99%（或 60%）的 1 型（或 2 型）糖尿病患者可能在糖尿病发病后的二十年内出现糖尿病视网膜病变 [1]。2010 年，全球糖尿病视网膜病变患者约有 1.266 亿人；目前估计到 2030 [3, 4]这一数字将增至约 1.91 亿人。然而，通过及时治疗并监测病情严重程度，约 56% 的新增糖尿病视网膜病变病例可以减少 [5]。眼科医生会分析眼底图像，检查微动脉瘤、硬性／软性渗出物和出血等病灶症状，以了解糖尿病视网膜病变的严重程度分期[1, 2]。糖尿病视网膜病变分为以下几个阶段 [5]：（1）轻度表示最早期阶段，其特征是出现微动脉瘤；（2）中度表示血管开始丧失输送能力的阶段；（3）重度表示血管发生阻塞，进而促使新生血管生长；（4）增殖期表示病情进展阶段，其特征是新生血管开始生长。图1 展示了一些代表糖尿病视网膜病变严重程度阶段的眼底图像。人工评估眼底图像以判定糖尿病视网膜病变的严重程度，可能因患者数量庞大、训练有素的临床医生不足、诊断时间过长、病灶不明确等因素而产生不一致。此外，眼科医生在选择正确的严重程度等级时也可能存在分歧 [6]。因此，计算机辅助技术应运而生，有助于提高诊断水平，并拓展早期检测的前景 [2]。

基于眼底照片的糖尿病视网膜病变严重程度分期自动检测在过去二十五年间不断发展。早期曾使用一些图像处理工具 [8, 9], ，但基于机器学习（ML）的糖尿病视网膜病变方



图 1：来自 APTOS-2019 [7] 的不同糖尿病视网膜病变严重程度阶段眼底图像

法在本世纪初 21st 开始流行。基于机器学习的技术大多依赖人工设计的特征：这些特征经过仔细提取后，再输入分类器，例如随机森林[10], K近邻 [11], 支持向量机 [12],和人工神经网络 [13]。尽管支持向量机和人工神经网络模型在糖尿病视网膜病变领域备受推崇，但基于人工设计特征的机器学习模型需要高效的先验特征提取，而处理复杂眼底图像时可能因此产生错误 [2, 1]。另一方面，基于深度学习的模型通过卷积运算自动提取特征 [14, 15]。此外，自 2012 年以来，计算机视觉领域深度学习架构的蓬勃发展显著影响了基于眼底图像的糖尿病视网膜病变严重程度分析 [1]。以往基于深度学习的技术大多采用卷积神经网络 [16, 1]。然而，传统卷积神经网络几乎不具备关注特定区域或特征并弱化其余部分的能力。此外，基于卷积神经网络的分类器可能会损害所学习特征之间的空间关系。为克服卷积神经网络的局限性，研究者已采用胶囊网络 [17], ，或通过引入注意力机制进行分类 [18, 19]。尽管文献中已有多项研究 [1, 2]，并且研究者也尝试检测糖尿病视网膜病变在发展初期是否存在，但通过采用更先进的深度学习模型、实现更高程度的自动特征提取，仍有望进一步提升检测性能。

本文采用并微调Transformer模型，以利用其多头自注意力（MSA） [20] 聚焦于眼底图像中揭示糖尿病视网膜病变严重程度迹象的特定区域。此外，近年来，Transformer模型在各种计算机视觉任务 [21, 20]中展现出优异的性能。最初，鉴于ViT（视觉Transformer）[20] 在图像分类任务中的卓越表现，我们采用它来检测糖尿病视网膜病变的严重程度。ViT将输入图像拆分为一系列图像块，并应用全局注意力 [20]。由于标准ViT需要大量数据，我们还探索了其他一些图像变换器模型，例如CaiT（图像Transformer中的类别注意力） [22], 、DeiT（数据高效图像Transformer） [23], 和BEiT（图像Transformer双向编码器表示） [24]。CaiT是ViT的改进版本，并采用了专门的类别注意力 [22]。DeiT运用知识蒸馏，将一个网络的知识迁移到另一个网络，并构建教师-学生层次网络 [23]。BEiT从BERT（来自Transformer的双向编码器表示） [25]汲取灵感，以实现对图像块的掩码

，并对其进行建模，以预训练ViT [24]。在实验中，我们使用了公开的APTOS-2019失明检测数据集 [7]，，其中各个图像变换器的表现并不理想。因此，我们将上述经过微调的图像变换器进行集成，以期获得更好的预测性能。集成图像变换器在糖尿病视网膜病变严重程度分期检测方面取得了相当令人鼓舞的结果。这是最早尝试采用并集成图像变换器进行糖尿病视网膜病变严重程度分期检测的研究之一，也是本文的主要**贡献**。

本文其余部分安排如下。第2 节讨论糖尿病视网膜病变的相关文献，第3 节介绍所提出的方法。随后，第4节分析并讨论实验结果。最后，第5节总结全文。

2. 相关工作

本节简要介绍基于眼底图像进行糖尿病视网膜病变严重程度检测的相关文献。现代糖尿病视网膜病变严重程度分期可追溯至 ETDRS 研究组的报告 [26]。过去曾发表了一些基于图像处理的策略（例如小波变换[8]，拉东变换 [9]）。过去二十年来，基于机器学习和深度学习的方法占据了主导地位。我们将相关工作大致分为（a）手工设计特征模型 [27, 11, 28]，和（b）基于深度特征的模型 [2]，，并将在下文中讨论。

2.1. 手工设计特征模型

手工设计特征模型主要采用随机森林 [27],K近邻 [29]，支持向量机 [28]，人工神经网络 [30] 来检测糖尿病视网膜病变严重程度阶段。Acharya等人 [27] 采用基于离散小波/余弦变换特征的决策树，这些特征从视网膜图像中提取。Casanova等人 [10] 引入随机森林用于糖尿病视网膜病变严重程度阶段分类。在 [31]，中，随机森林也用于评估糖尿病视网膜病变风险。 [11]采用K近邻分类器检测玻璃疣、渗出物和棉絮斑，以诊断糖尿病视网膜病变。Tang等人 [29]使用K近邻从眼底照片中检测视网膜出血。在 [28]，中，使用支持向量机检测糖尿病视网膜病变引起的视网膜变化。Akram等人 [12] 使用支持向量机和高斯混合模型（GMM），结合形状、强度和受影响区域统计信息等增强特征，识别微动脉瘤，以便早期检测糖尿病视网膜病变。 [13] 采用人工神经网络对病灶进行分类，以检测糖尿病视网膜病变严重程度。Osareh等人 [32]采用基于模糊C均值的分割

并采用基于遗传算法的特征选择与人工神经网络来检测糖尿病视网膜病变中的渗出物。在 [30], 中，使用粒子群优化进行特征选择，随后采用基于人工神经网络的糖尿病视网膜病变严重程度分类。

2.2. 基于深度特征的模型

过去的深度架构大多使用卷积神经网络来处理糖尿病视网膜病变的严重程度问题。例如，Yu 等人 [16] 使用卷积神经网络检测糖尿病视网膜病变中的渗出物；Chudzik 等人 [33] 利用卷积神经网络结合迁移学习和层冻结进行微动脉瘤检测；Gargeya 和 Leng [34] 采用基于卷积神经网络的深度残差学习来识别患有糖尿病视网膜病变的眼底图像。在 [4], 中，卷积神经网络也用于识别糖尿病视网膜病变严重程度阶段及一些相关眼病，例如青光眼和年龄相关性黄斑变性（AMD）。在 [35], 中，研究人员采用了一些经典的卷积神经网络架构（例如 AlexNet、VGG Net、GoogLeNet、ResNet）进行糖尿病视网膜病变严重程度分期检测。Wang 等人 [18] 提出了 Zoom-in-Net，将卷积神经网络、注意力机制和贪心算法相结合，以放大感兴趣区域来处理糖尿病视网膜病变。 [19] 采用改进的 DenseNet169 架构结合注意力机制，提取用于糖尿病视网膜病变严重程度分级的精细特征。 [36], 采用改进的 Xception架构进行糖尿病视网膜病变分类。 [37]提出了 TAN（纹理注意力网络），利用风格（纹理特征）和内容（语义及上下文特征）重新校准机制。Tymchenko 等人 [5] 集成了三种卷积神经网络架构（EfficientNet-B4 [38],、EfficientNet-B5 和 SE-ResNeXt50 [39]），用于糖尿病视网膜病变严重程度检测。最近又出现了一些基于 Transformer 的模型，例如 CoT-XNet [40], 它结合了上下文 Transformer 和 Xception架构；以及 SSiT [41], 它采用显著图引导的自监督图像 Transformer。

3. 方法

本节首先对问题陈述进行形式化定义，然后提出解决方案架构。

3.1. 问题表述

在本研究中，我们将眼底摄影拍摄的图像 I 输入架构。任务是根据 I 预测糖尿病视网膜病变的严重程度阶段，分为阴性、轻度、中度、重度和增殖期。我们将该任务表述为一个多分类问题 [15]。在此，从 I中提取特征并输入分类器，以预测糖尿病视网膜病变严重程度类别

标签 c，其中 $c = \{0, 1, 2, 3, 4\}$ 分别对应于{阴性、轻度、中度、重度、增殖期}。

3.2. 解决方案架构

为了从眼底照片中检测糖尿病视网膜病变的严重程度阶段，我们采用图像变换器，即 ViT [20], BEiT [24], CaiT [22], 和 DeiT [23], 并将它们集成。不过，在将原始眼底图像输入 Transformer 之前，我们会先对其进行预处理，下面首先讨论这一点。

3.2.1. 预处理

深度学习模型的性能容易受到输入模型的数据质量和数量的影响。由于图像中可能存在预先存在的噪声和不一致性，仅将原始数据作为输入，很难使模型达到最佳性能。因此，明确的预处理流程对于更好地训练模型至关重要 [15]。

我们现在讨论应用于原始眼底照片的各种预处理和增强技术 [42,15] ，以便更好地学习。在数据集中，眼底图像的尺寸可能各不相同；因此，我们将图像I 调整为 256×256 尺寸的图像 I_z。深度网络对数据需求量大，因此已采用多种增强方法来解决深度网络 [43], 面临的各种问题，包括模型过拟合，并提升模型的鲁棒性。因此，我们对训练集（DB_{tr}）进行数据增强，使用中心裁剪，保留中央_比例$= 0.5$的区域、水平/垂直翻转、在 $[0^o, 45^o]$范围内随机旋转、最大_delta $= 0.95$的随机亮度变化，以及在区间 $[0.1, 0.9]$内的随机对比度变化。我们还对DB_{tr}中30%的样本应用限制对比度自适应直方图均衡化（CLAHE） [44] ，以确保仅在较小区域内过度增强对比度，而非作用于整幅图像。医学图像中的强度通常不均匀，可能影响自动图像分析方法的性能。文献中已针对不同类型的图像采用一些归一化方法，以获得良好的性能 [45]。此外，图像噪声可能影响计算机化方法，因此针对不同类型的图像采用了不同的去噪算法 [46]。然而，这些方法可能会增加计算成本。因此，本文提出的方法未采用这些处理。

3.2.2. Transformer 网络

长期以来，计算机视觉任务中的深度学习模型一直以卷积神经网络为主，通过一系列卷积运算处理图像来提取高级特征图，再将其输入多层感知机（多层感知机）进行分类 [47]。近年来，Transformer 模型在自然语言处理（自然语言处理）领域表现出显著增长，其性能更为出色 [21]。为了通过 Transformer 实现类似的高性能，它们已被引入图像分类及其他一些面向计算机视觉的任务 [20]。此外，与卷积神经网络相比，Transformer模型具有较少的图像特定归纳偏置 [20]。

为了从眼底图像中识别糖尿病视网膜病变的严重程度分期，我们在此高效地采用并集成了一些图像变换器，例如 ViT [20], BEiT[24], CaiT [22], 和 DeiT [23]。在介绍集成的 Transformer 模型之前，我们先讨论各个图像变换器如何适配我们的任务，并从 ViT 开始。

3.2.2.1. 视觉 *Transformer (ViT)*：ViT 模型采用了基于文本的 Transformer 模型的思路 [48], ，即将输入图像视为一系列图像块，而非文本词语，然后提取特征并输入多层感知机 [20]。

ViT 的示意图见图 2。在这里，输入图像 I_z 被转换为一系列展平的图像块 x_p^i （其中 i =1、2、.....、n_p），每个图像块的大小为 $w_p \times w_p \times c_p$，其中 c_p 表示 I_z的通道数。由于 I_z 是红绿蓝眼底图像，因此这里 $c_p = 3$。在我们的任务中，I_z 的大小为 256×256，根据经验，我们选择 $w_p = 64$，由此得到$n_p = (\frac{256}{64})^2 = 16$。每个图像块 x_p^i 会进一步展平，并通过可训练线性投影映射为一个D维潜在向量（即图像块嵌入 z_0），如下所示。

$$z_0 = [x_{class}\ ;\ x_p^1\,\mathbb{E}\ ;\ x_p^2\,\mathbb{E}\ ;\ \ldots\ ;\ x_p^{n_p}\,\mathbb{E}] + \mathbb{E}_{pos} \tag{1}$$

其中，$\mathbb{E}$ 是图像块嵌入投影，$\mathbb{E} \in \mathbb{R}^{w_p \times w_p \times C \times D}$；$\mathbb{E}_{pos}$ 是添加到图像块嵌入中的位置嵌入，用于保留图像块的位置信息，$\mathbb{E}_{pos} \in \mathbb{R}^{(n_p+1)\ \times D}$；$x_{class} = z_0^0$是一个可学习的嵌入[25]。

将图像块映射到嵌入空间并加入位置信息后，我们再添加一系列 Transformer 编码器 [49, 20]。Transformer 编码器的内部结构如图3所示，其中包含两个模块 A_s 和 F_n。A_s 和 F_n 包含 MSA （多头自注意力）

[20] 和 MLP [15] 模块。 LN （层归一化） [50] 和残差连接 [15] 分别应用于这些模块之前和之后。公式2 以一般语义展示了这一点。这里，MLP 模块由两层组成，分别具有 $4D$ 和 D 个神经元，并采用与 [20]类似的高斯误差线性单元非线性激活函数。

$$\begin{aligned} z'_l &= MSA(LN(z_{l-1})) + z_{l-1}; \\ z_l &= MLP(LN(z'_l)) + z'_l; \ l = 1, 2, \ldots, L \end{aligned} \tag{2}$$

其中， L 是 Transformer 块的总数。Transformer编码器的核心组件是 MSA ，包含 h 个注意力头，每个注意力头都包含 SA（缩放点积注意力） [20, 49]。每个注意力头 $i \in \{1, 2, ..., h\}$ 的 MSA都会计算一个由查询、键和值组成的元组 [20], ，即（Q^i， K^i， V^i），如下所示。

$$Q^i = XW_Q^i \ ; \ K^i = XW_K^i \ ; \ V^i = XW_V^i \tag{3}$$

其中， X 是输入嵌入， W_Q、 W_K、 W_V 是

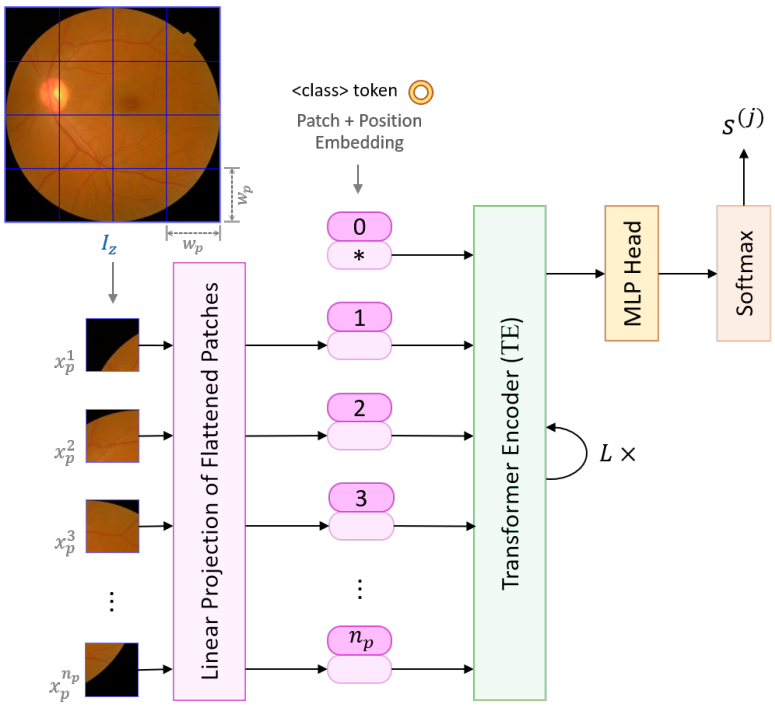

图 2：ViT 的工作流程

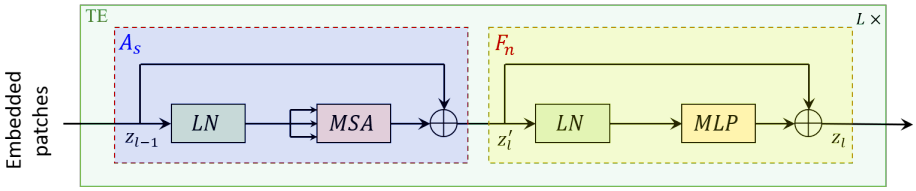

图3：Transformer编码器（TE）的内部视图

线性变换中使用的权重矩阵。元组（Q， K， V）被输入到 SA 中，以计算对输入图像块所需的注意力，如下所示。

$$SA(Q, K, V) = \psi\left(\frac{QK^T}{\sqrt{D_h}}\right)V \tag{4}$$

其中， ψ 是Softmax函数，且 $D_h = D/h$。所有头的 SA结果会在 MSA中拼接，如下所示。

$$MSA(Q, K, V) = [SA^1\ ;\ SA^2\ ;\ \ldots\ ;\ SA^h]W_L \tag{5}$$

其中， W_L 是一个权重矩阵。

经过多个Transformer编码器块后， <类别> 标记 [25] 会融入上下文信息。Transformer编码器（z_L^0）输出时可学习嵌入的状态充当图像表示 y[20]。

$$y = LN(z_L^0) \tag{6}$$

现在，如图2所示，我们添加一个包含128个神经元的隐藏层的MLP头。为捕获非线性，我们在此处使用Mish [51] 。在输出层中，我们保留五个神经元，并采用Softmax激活函数，以获得概率分布 $s^{(j)}$ ，从而将眼底照片分类为上述五个糖尿病视网膜病变严重程度阶段之一。

3.2.2.2. 数据高效图像 *Transformer*（*DeiT*）：训练数据较少时，ViT 的泛化能力较差。在这种情况下，DeiT 仍能取得不错的表现，且占用的内存更少 [23]。DeiT 采用 ViT 特有的策略，并通过知识蒸馏与教师-学生方案相结合 [52]。DeiT 的关键在于知识蒸馏机制，本质上是将知识从一个模型（教师）迁移到另一个模型（学生） [23]。在这里，我们使用 EfficientNet-B5 [38] 作为预先训练好的教师模型。学生模型采用 Transformer，

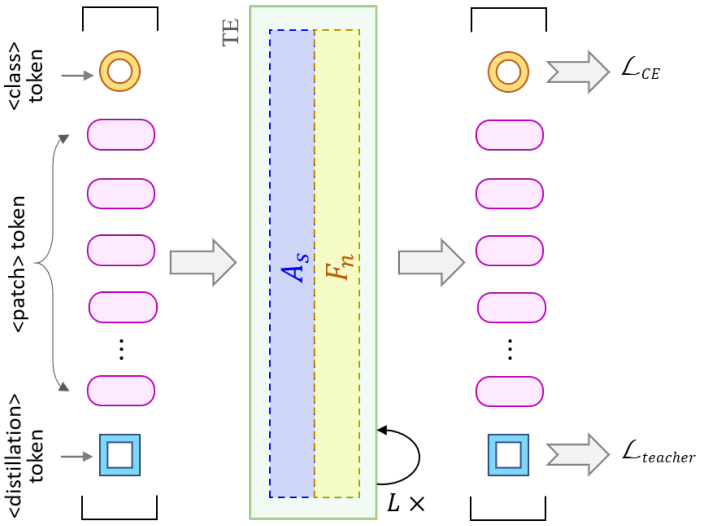

图 4：DeiT 的蒸馏过程

通过注意力机制，根据蒸馏标记从教师模型的输出中学习 [23]。在本研究中，我们采用硬标签蒸馏[23], ，将教师模型的硬判决视为真实标签，即 $y_t = \text{argmax}_c Z_t(c)$。硬标签蒸馏目标定义如下。

$$\mathcal{L}_{global}^{hard} = 0.5\ \mathcal{L}_{CE}(\psi(Z_s), y) + 0.5\ \mathcal{L}_{CE}(\psi(Z_s), y_t) \tag{7}$$

其中， $\mathcal{L}_{CE}$ 是针对真实标签的交叉熵损失， y， ψ 是 softmax 函数， Z_s 和 Z_t 分别是学生模型和教师模型的逻辑值。通过标签平滑，硬标签可以转换为软标签[23]。

在图 4中，我们展示了 DeiT 的蒸馏过程。这里，我们向 Transformer 中添加<蒸馏> 标记，它通过 Transformer 编码器与 <类别>和 <图像块> 标记进行交互。此处使用的 Transformer 编码器与 ViT 的类似，包含 A_s 和 F_n 个模块，如图 3所示。 <蒸馏>标记的目标是复现教师模型的预测标签，而不是复现真实标签。<蒸馏> 和 <类别>标记通过反向传播进行学习 [15]。

DeiT 使用线性分类器，而不是 ViT 的多层感知机头 [23, 20]，以便在计算资源有限的情况下高效运行。

3.2.2.3. 图像 *Transformer* 中的类别注意力（*CaiT*）： CaiT 通常

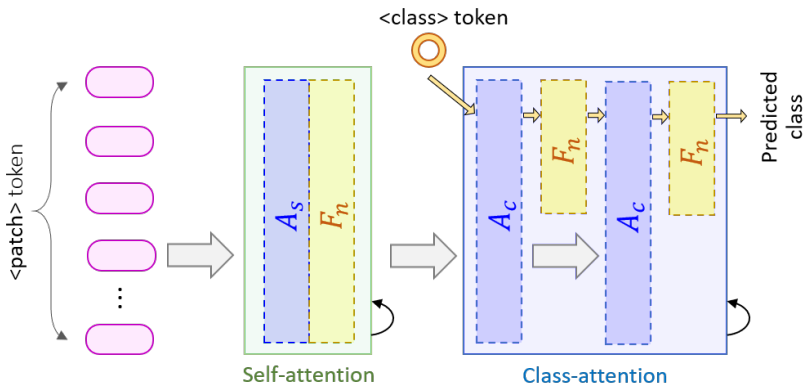

图 5：CaiT 的工作流程

在浮点运算次数和学习参数较少的情况下，其性能通常优于 ViT 和 DeiT [15], ，尤其是在需要增加 Transformer 深度时 [22]。CaiT 基本上是 ViT 的升级版本，它采用了具有特定类别注意力和 LayerScale 的层 [22]。在图5中，我们展示了 CaiT 的工作流程。

LayerScale 通过在 A_s 和 F_n 模块的输出上分别乘以对角矩阵 M_λ ，帮助 CaiT 在更深的网络中运行。

$$\begin{aligned} z'_l &= M_\lambda(\lambda^l_1, \ldots, \lambda^l_D) \times MSA(LN(z_{l-1})) + z_{l-1}; \\ z_l &= M_\lambda(\lambda'^l_1, \ldots, \lambda'^l_D) \times MLP(LN(z'_l)) + z'_l \end{aligned} \tag{8}$$

其中， λ^l_i 和 λ'^l_i 是学习参数，其他符号的含义与上述 ViT 中相同。

在 CaiT 中，负责图像块间自注意力的 Transformer 层与类别注意力层相互分离。类别注意力层专门用于将图像块的内容提取为一个向量，并将其传递给线性分类器 [22]。后期阶段才会插入 <类别> 标记，因此初始层可以专门执行图像块间的自注意力。在类别注意力阶段，我们交替使用多头类别注意力（A_c） [22]和 F_n，如图5所示，并且只更新类别嵌入。

3.2.2.4. 图像双向编码器表示*Transformer (BEiT)*：BEiT 是一种自监督模型，其根源于 BERT（来自Transformer的双向编码器表示） [25], ，并利用双向编码和预训练 [24]。在图 6中，我们展示了 BEiT 的预训练流程。

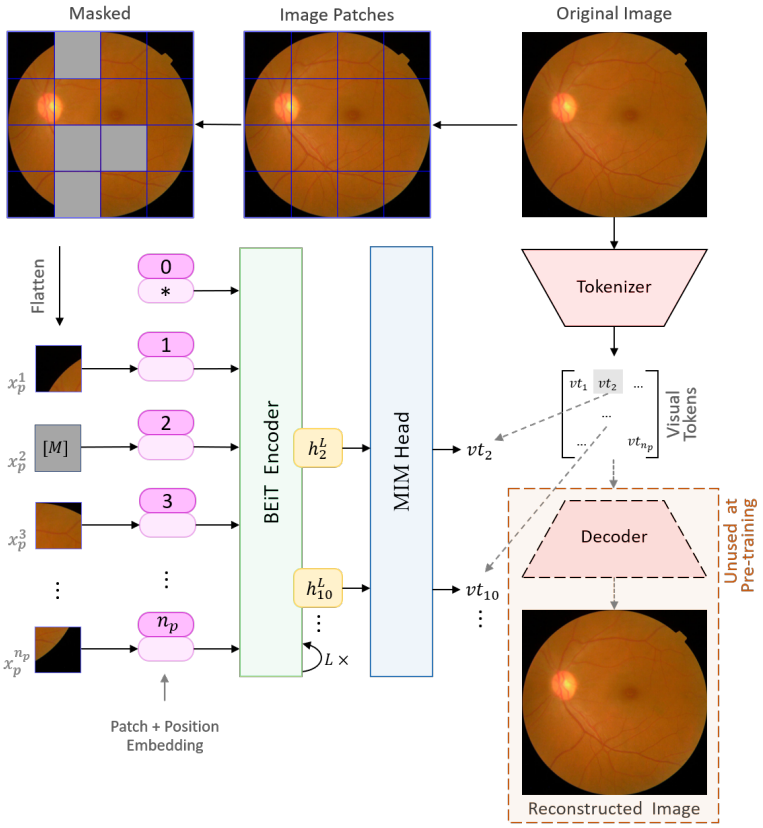

图 6：BEiT 预训练流程

输入图像 I_z 被分割成图像块 x_p^i 并展平为向量，与前文提到的 ViT 类似。在 BEiT 中，使用骨干变换器，我们采用 ViT [20]。另一方面， I_z 则表示为一系列视觉标记 $vt = [vt_1, vt_2, \ldots, vt_{n_p}]$，这些标记由离散变分自编码器（变分自编码器） [53]生成。为了学习视觉标记，我们采用一个分词器 $\mathcal{T}_\phi(vt \mid x)$ 将图像像素 x 映射为标记 vt，并使用解码器 $\mathcal{D}_\theta(x \mid vt)$从 x 重建输入图像像素 vt [24]。

这里执行一项掩码图像建模（MIM） [24]任务，以对图像变换器进行预训练：随机掩码部分图像块，然后预测相应的视觉词元。被掩码的图像块会替换为可学习嵌入 $e_{[M]}$。我们将受损的图像块 $x^{\mathcal{M}} = \{x_p^i : i \notin \mathcal{M}\} \bigcup \{e_{[M]} : i \in \mathcal{M}\}$ 输入 Transformer编码器。这里， $\mathcal{M}$ 是掩码位置的索引集合。

编码表示 h_i^L是最后一个Transformer层 L 中对应于图像块 i^{th} 的隐藏向量。对于每个掩码位置，使用Softmax分类器 ψ 预测相应的视觉词元，即 $p_{\text{MIM}}(vt' \mid x^{\mathcal{M}}) = \psi(W_M h_i^L + b_M)$；其中， W_M 和 b_M 包含线性变换的学习参数。BEiT的预训练目标是最大化给定 $x^{\mathcal{M}}$时正确词元 vt_i 的对数似然，如下所示：

$$max \sum_{x\in \text{ DB}_{tr}} \mathbb{E}_{\mathcal{M}} \left[\sum_{i\in\mathcal{M}} \log p_{\text{MIM}} \left(vt_i \mid x^{\mathcal{M}} \right) \right]$$

其中，数据库$_{tr}$ 是训练数据集。BEiT 预训练可视为变分自编码器训练 [24, 53], ，其中包含两个阶段，即 阶段 *1*：最小化视觉标记重建的损失；阶段 *2*：对掩码图像进行建模，即学习先验分布 p_{MIM} ，同时固定 $\mathcal{T}_\phi$ 和 $\mathcal{D}_\theta$ 。可写为如下形式：

$$\sum_{\substack{(x_i, x_i^{\mathcal{M}}) \\ \in \text{ DB}_{tr}}} \left(\underbrace{\mathbb{E}_{vt_i \sim \mathcal{T}_\phi(vt|x_i)} \left[\log \mathcal{D}_\theta(x_i|vt_i) \right]}_{stage\text{-}1} + \underbrace{\log p_{\text{MIM}} \left(\hat{vt}_i | x_i^{\mathcal{M}} \right)}_{stage\text{-}2} \right)$$

其中，ˆ vt_i =argmax$_{vt}$ $\mathcal{T}_\phi(vt \mid x_i)$。

3.2.3. 集成Transformer

上述四种图像变换器，即 ViT [20], DeiT [23],CaiT [22], 和 BEiT [24] 均在训练集数据库$_{tr}$上进行预训练。现在，我们将这些 Transformer 进行集成，以预测测试集数据库$_t$眼底图像的严重程度分期，因为多个学习算法的集成表现优于各个组成算法单独使用 [54]。集成式 Transformer 网络的示意图见图7。

对于来自数据库的图像样本$_t$，我们获得 softmax 概率分布 $s^{(j)}$：$\{P_1^j, P_2^j, \ldots, P_{n_c}^j\}$在 j^{th}Transformer [15],中，对于 $j = 1$，2，……，n_T；其中， n_c 是类别（严重程度阶段）的总数，而 n_T 是所采用的图像变换器数量。这里， $\sum_{i=1}^{n_c} P_i^j = 1$， $n_c = 5$ （参见第 3.1 小节），并且 $n_T = 4$ 由于我们使用了四个分别训练的不同图像变换器，如前文所述。

我们分别使用两种组合方法加权平均和多数投票来获得严重程度分期/类别_标签 $\hat{c}|_{wm}$ 和 $\hat{c}|_{mv}$ ，分-

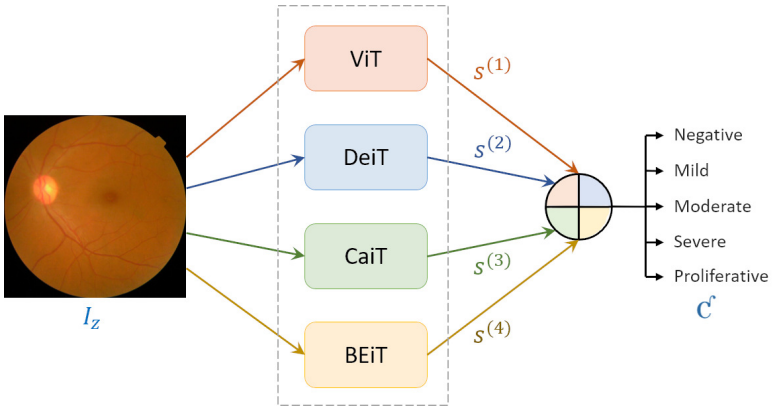

图 7：集成式 Transformer 网络。

别地。

$$\hat{c}|_{wm} = \text{argmax}_i \; P_i^{\mu} \; ; \; P_i^{\mu} = \frac{\sum_{j=1}^{n_T} \alpha_j P_i^j}{\sum_{j=1}^{n_T} \alpha_j} \; ; \; \text{for } i = 1, 2, \ldots, n_c \tag{9}$$

在这项任务中，我们选择 $\sum_{j=1}^{n_T} \alpha_j = 1$。

$$\begin{aligned} \hat{c}|_{mv} &= \text{mode}\left(\text{argmax}_i(P_i^1), \text{argmax}_i(P_i^2), \ldots, \text{argmax}_i(P_i^{n_T})\right) \\ &= \text{mode}\left(\text{argmax}_i(s^{(1)}), \text{argmax}_i(s^{(2)}), \ldots, \text{argmax}_i(s^{(n_T)})\right); \\ &\qquad\qquad \text{for } i = 1, 2, \ldots, n_c \end{aligned} \tag{10}$$

在这项任务中，我们在所采用的 [42] 图像变换器中使用交叉熵作为损失函数。此处使用 AdamW 优化器，以利用其权重衰减正则化效果来应对过拟合 [55]。超参数调优的训练细节见第4.2节。

4. 实验与讨论

本节介绍所采用的数据库，随后展示实验结果并进行讨论。

4.1. 所用数据库

在计算实验中，我们使用了 Kaggle APTOS（亚太远程眼科协会）公开提供的训练样本

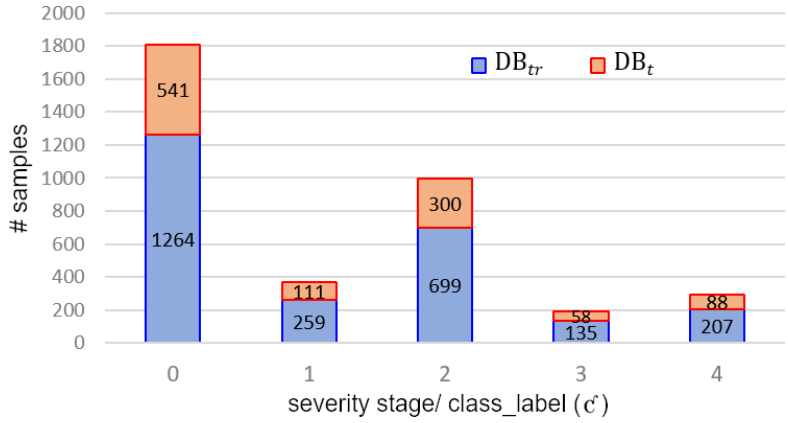

图 8：APTOS-2019 [7]。

2019年失明检测数据集 [7], ，即 APTOS-2019。该数据库（DB）包含五种糖尿病视网膜病变严重程度分期的眼底图像样本，分别为阴性、轻度、中度、重度和增殖期。图1展示了该数据集中的一些样本图像。数据库中共有3662张眼底图像，我们按7：3的比例将其划分为训练集（DB$_{tr}$）和测试集（DB$_t$），并确保 DB$_{tr}$ 和 DB$_t$ 中各层（糖尿病视网膜病变严重程度类别）的样本比例相同。实际上，DB$_{tr}$ 和 DB$_t$ 互不重叠。为验证模型，我们将 DB$_{tr}$ 中10%的数据留作验证集 DB$_v$。图8分别展示了 DB$_{tr}$ 和 DB$_t$ 中各严重程度阶段／类别_标签（c）的样本数量。其中，49.3%的样本属于阴性糖尿病视网膜病变类别（$c = 0$）。在阳性类别中，中度阶段（$c = 2$）所占比例最大，占样本总数的27.28%；重度阶段（$c = 3$）所占比例最小，仅占样本总数的5.27%。图8显示，由于不同严重程度阶段的样本数量不均，数据库中的数据存在不平衡。为解决这一问题，我们在模型训练期间采用了数据增强，如第3.2.1小节所述。数据增强也有效降低了过拟合风险 [15]。

4.2. 实验结果

本节讨论所开展的实验，分析模型结果，并将其与主要先进方法进行比较。我们首先介绍实验设置。

4.2.1. 实验设置

我们在 TensorFlow-2 框架上开展实验，使用 Python 3.7.13，运行环境的配置如下：Intel(R) Xeon(R) CPU @ 2.00GHz、52 GB RAM 和 Tesla T4 16 GB 图形处理器。此处展示的所有结果均来自数据库$_{to}$。

在训练期间，我们根据数据库$_{v}$样本上的性能对框架的超参数进行了调优并固定。对于此处使用的所有图像变换器（即 ViT、DeiT、CaiT 和 BEiT），我们通过经验将以下超参数设定为：Transformer_层数（L） $=$ 12，嵌入_维度（D） $=$ 384，头数（h） $=$ 6。AdamW [55]采用了以下超参数：初始_学习_率$=$ 10^{-3}； 1st 和 2nd 矩估计的指数衰减率，即 $\beta 1 = 0.9$、 $\beta 2 = 0.999$；分母为零时的消除参数（ε） $=$ 10^{-8}；以及权重_衰减 $=$ $10^{-3}/4$。模型训练时，小批量大小固定为 32。

4.2.2. 模型性能

在表 1中，我们展示了集成图像变换器（EiT）采用加权平均（wm）和多数投票（mv）组合方案的性能，分别通过 EiT_{wm}和 EiT_{mv}获得了 94.63% 和 91.26% 的准确率。我们还对所用 Transformer 的多种组合进行了集成，并在该表中展示了它们的性能。在这里， wm方案的表现优于 mv。从表中可以明显看出，集成不同类型的 Transformer 提升了性能。在单个 Transformer 中（对于 $n_T = 1$），CaiT 的表现最佳。对于 $n_T = 2$ 和$n_T = 3$，“BEiT $+$ CaiT”和“DeiT $+$ BEiT $+$ CaiT”的表现分别优于其他组合。总体而言， EiT_{wm} 在此处取得了最高准确率。

图 A.10 见附录 A，其中展示了 Grad-CAM 生成的粗定位图 [56]，这些图由所采用的单个图像变换器生成，用于突出显示理解严重程度分期的关键区域。

4.2.2.1. 多种评估指标： 除准确率外，表 2还展示了 EiT 在其他一些评估指标上的表现，例如 Kappa 系数、精确率、召回率（或敏感度）、F_1 分数、特异度、平衡准确率 [57]。科恩二次加权卡帕系数用于衡量人工标注分数（即糖尿病视网膜病变严重程度阶段）与 EiT预测分数之间的一致性。精确率衡量所有阳性预测中真阳性样本所占的比例。召回率或敏感度用于计算真阳性

表1：不同Transformer集成方式的性能

n_T	集成Transformer	准确率（%）	
		加权 mean	多数 投票
1	ViT	82.21	
	DeiT	85.65	
	BEiT	86.74	
	CaiT	86.91	
2	ViT + DeiT	87.03	86.55
	ViT + BEiT	87.48	87.03
	ViT + CaiT	87.77	87.21
	DeiT + BEiT	88.18	87.69
	DeiT + CaiT	88.86	87.93
	BEiT + CaiT	89.28	88.12
3	ViT + DeiT +BEiT	90.53	88.87
	ViT + DeiT +CaiT	91.39	89.56
	ViT + BEiT + CaiT	92.14	90.28
	DeiT + BEiT +CaiT	93.46	90.91
4	ViT + DeiT + BEiT + CaiT (EiT)	94.63	91.26

率。同样，特异度用于计算真阴性率。F_1 分数是精确率与召回率的调和平均数。由于所用数据库存在类别不平衡，我们还计算了平衡准确率，即敏感度与特异度的算术平均数。从表中可以看出，对于 EiT_{wm}和 EiT_{mv}，卡帕系数均大于 0.81，表明人工评定者与 EiT [57]之间具有“几乎完全一致性”。此处，宏平均 表示所有类别的精确率、召回率和 F_1 分数的算术平均数。

表 2： EiT 在各项评估指标上的表现

指标	Weighted mean (EiT_{wm})	Majority vot ing (EiT_{mv})
准确率（%）	94.63	91.26
Kappa 系数	0.92	0.87
宏平均精确率（%）	90.55	84.65
宏平均召回率（%）	92.88	88.81
宏平均 F_1 分数（%）	91.67	86.55
宏平均特异度（%）	98.62	97.74
平衡准确率（%）	95.75	93.27

4.2.2.2. 各类别表现：表3展示了各个类别的

在检测每个糖尿病视网膜病变严重程度阶段时的性能 EiT_{wm} 和 EiT_{mv}。从该表可以看出，我们的模型在识别 阴性 糖尿病视网膜病变（c= 0）类别时，精确率和召回率最高，而在 重度 糖尿病视网膜病变类别（c= 3）上的表现最差。重度类别的表现较差，原因是该类别可用样本稀缺（参见第 4.1节）。

表 3： EiT 在糖尿病视网膜病变各严重程度阶段上的表现

	类别 _ 标签（c）	0	1	2	3	4
EiT_{wm}	精确率（%）	**98.48**	86.67	*95.00*	83.61	89.01
	召回率 (%)	**95.75**	93.69	*95.00*	87.93	92.05
	F_1-分数（%）	**97.09**	90.04	*95.00*	85.71	90.50
	特异度（%）	98.56	98.38	98.12	**99.04**	*99.01*
EiT_{mv}	精确率（%）	**96.74**	79.67	*94.14*	70.59	82.11
	召回率 (%)	**93.35**	88.29	*91.00*	82.76	88.64
	F_1-分数（%）	**95.01**	83.76	*92.54*	76.19	85.25
	特异度（%）	96.95	97.47	97.87	*98.08*	**98.32**

每行中，最佳结果以**粗体**标出，次佳结果以斜体标出，最低结果加下划线。

4.2.3. 比较

在表4中，我们将其与一些主要的当代基于 CNN 的架构进行了比较分析，例如 ResNet50 [58], InceptionV3 [59], MobileNetV2 [60], Xception [61] 及其改进版本（Kassani 等人 [36]）、EfficientNet [38], SE-ResNeXt50 [39], 以及集成 CNN（Tymchenko 等人[5]）。基于 CNN 的分类器往往无法聚焦于眼底图像中的特定区域或特征，同时忽略其他部分，这可能导致在学习用于糖尿病视网膜病变严重程度检测的特征时丢失空间关系。此前融入注意力机制的尝试，例如 Farag 等人 [19] 采用了带注意力机制的 DenseNet169；TAN（纹理注意力网络） [37] 旨在重新校准纹理和内容特征。然而，这些基于注意力的模型并未始终表现出更优的性能。为应对这些局限，近期的进展引入了采用多头自注意力的基于 Transformer 的模型。例如，CoT-XNet [40] 将上下文 Transformer 与 Xception 相结合，而 SSiT [41] 则采用由显著图引导的自监督图像 Transformer；这些方法展现了利用 Transformer 网络根据眼底图像检测糖尿病视网膜病变严重程度的潜力。我们的基于加权平均的集成 Transformer 网络 EiT_{wm} 在准确率、平衡准确率、敏感度和特异度方面均优于主要的先进方法。我们的模型 EiT_{mv} 在.....方面也表现相当出色

表 4：比较研究

方法	准确率 (%)	敏感度 (%)	特异度 (%)	平衡准确率 (%)
ResNet50 [58]	74.64	56.52	85.71	71.12
InceptionV3 [59]	78.72	63.64	85.37	74.51
MobileNetV2 [60]	79.01	76.47	84.62	80.55
Xception [61]	79.59	82.35	86.32	84.34
Kassani 等人 [36]	83.09	88.24	87.00	87.62
EfficientNet-B4 [38]	90.30	81.20	97.60	89.40
EfficientNet-B5 [38]	90.70	80.70	97.70	89.20
SE-ResNeXt50 [39]	92.40	87.10	98.20	92.65
Tymchenko 等人 [5]	92.90	86.00	<u>98.30</u>	92.15
Farag 等人 [19]	82.00	-	-	-
TAN [37]	85.10	90.30	92.00	-
CoT-XNet [40]	84.18	-	95.74	-
SSiT [41]	<u>92.97</u>	-	-	-
EiT_{mv} [我们的方法]	91.26	<u>88.81</u>	97.74	<u>93.28</u>
EiT_{wm} [我们的方法]	**94.63**	**92.88**	**98.62**	**95.75**

每列中，最佳结果以**粗体**标示，次佳结果带下划线。

平衡准确率。

4.2.4. 超参数的影响

我们调整了超参数，并观察其对实验的影响。

a) MSA 头数：我们分析了多头自注意力中头数（h）对Transformer编码器性能的影响，并在图9中展示了结果。如图所示，EiT_{mv} 和 EiT_{wm} 的性能（准确率）均随着 h 增加而提高，直到 $h = 6$，此后开始下降。

b) 权重 α_j 的 EiT_{wm}：我们调整了权重 α_j （参见公式9），以考察其对 EiT_{wm}性能的影响。此处， α_j通过网格搜索技术进行了调整 [62]。我们从 EiT_{wm} 中获得了最高准确率 94.63%，对应 $\alpha_1 = \alpha_2 = 0.1$，以及 $\alpha_3 = \alpha_4 = 0.4$。调整 EiT_{wm}的 α_j期间的性能见表5。在附录 B.7 中附录 B，我们还列出了经过调整的 α_j，这些参数帮助获得了表1中性能最佳的集成Transformer。

4.2.5. 消融研究

我们在此介绍所开展的消融研究，通过系统地逐一消融各个 Transformer 来评估其影响。我们的 EiT 实际上是由四种不同的图像变换器集成而成，即 ViT、DeiT、CaiT 和 BEiT。我们逐一消融各个 Transformer，并观察到性能下降

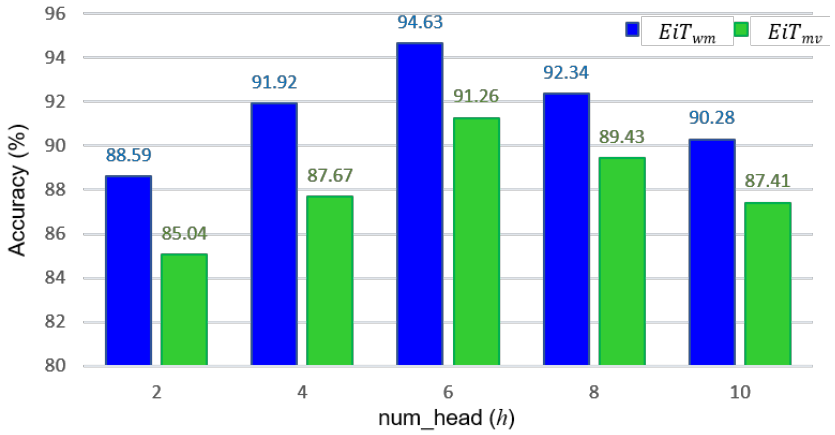

图 9：MSA 中头数（h）对模型性能的影响

表 5：通过调整权重 α_j 得到的 EiT_{wm} 性能

α_1	α_2	α_3	α_4	准确率（%）
0.25	0.25	0.25	0.25	89.53
0.85	0.05	0.05	0.05	82.29
0.05	0.85	0.05	0.05	85.78
0.05	0.05	0.85	0.05	86.92
0.05	0.05	0.05	0.85	87.05
0.7	0.1	0.1	0.1	82.35
0.1	0.7	0.1	0.1	85.91
0.1	0.1	0.7	0.1	87.04
0.1	0.1	0.1	0.7	87.20
0.5	0.167	0.167	0.166	82.88
0.166	0.5	0.167	0.167	86.35
0.167	0.166	0.5	0.167	87.62
0.167	0.167	0.166	0.5	87.74
0.3	0.3	0.2	0.2	88.16
0.3	0.2	0.3	0.2	89.58
0.3	0.2	0.2	0.3	90.27
0.2	0.3	0.3	0.2	90.85
0.2	0.3	0.2	0.3	91.67
0.2	0.2	0.3	0.3	92.72
0.4	0.4	0.1	0.1	91.18
0.4	0.1	0.4	0.1	91.49
0.4	0.1	0.1	0.4	92.15
0.1	0.4	0.4	0.1	92.84
0.1	0.4	0.1	0.4	93.47
0.1	0.1	0.4	0.4	**94.63**

幅度大于 EiT。例如，采用加权平均方案时，从 EiT 中消融 CaiT 后，准确率下降了 4.1%。同样，消融 BEiT 和 CaiT 后，准确率下降了 7.6%。对于我们的任务，最佳

单个 Transformer（CaiT）的准确率低了 7.72%，而 EiT_{wm}。更多示例见表 1。

4.2.6. 使用其他数据集进行预训练

我们通过使用其他数据集进行预训练，检查了我们的 EiT 模型的性能。我们从 MESSIDOR 中选取了 1200 张带有 [64] 判定等级的图像（称为 DB_M）。我们还使用了来自 IDRiD [65], 的“疾病分级”数据集，其中包含 516 张图像（称为 DB_I）。在此，我们从 DB_M中分别取 25%、50%、75% 和 100% 的样本，构建了四种训练集设置。同样，我们也从 DB_I中构建了四种训练设置。如第4.1节所述，我们按 7：3 的比例将 APTOS-2019 数据库（DB）划分为训练集（DB_{tr}）和测试集（DB_t）。表6展示了 EiT 在 DB_t上的性能，其中使用 DB_M 和 DB_I进行预训练，并使用 DB_{tr}进行训练。可以看出，使用其他数据集中的更多数据进行预训练后， EiT的性能略有提升。

表6： EiT 采用预训练后的准确率（%）

	预训练数据	25%	50%	75%	100%
EiT_{wm}	DB_M	94.71	94.78	94.83	94.88
	DB_I	94.65	94.67	94.7	94.79
	数据库$_M$ +数据库$_I$	94.73	94.85	94.98	95.13
	N.A.	94.63			
EiT_{mv}	DB_M	91.35	91.48	91.56	91.61
	DB_I	91.27	91.32	91.34	91.35
	数据库$_M$ +数据库$_I$	91.42	91.6	91.68	91.75
	N.A.	91.26			

不适用：不使用预训练数据

5. 结论

本文针对从眼底图像中自动检测糖尿病视网膜病变严重程度阶段这一问题展开研究。为此，我们采用加权平均和多数投票两种组合方案，提出了两种集成图像 Transformer： EiT_{wm} 和 EiT_{mv}。我们选用了四种 Transformer 模型，即 ViT、DeiT、CaiT 和 BEiT。在实验中，我们使用了公开可用的 APTOS-2019 失明检测数据集，在该数据集上，EiT_{wm} 和 EiT_{mv} 的准确率分别达到 94.63% 和 91.26%。尽管所用数据集存在类别不平衡问题，我们的模型仍表现出色。我们的 EiT_{wm} 优于主流的最先进技术。我们

我们还进行了消融研究，并观察到集成方法优于单个 Transformer。

未来，我们将努力运用一些类别不平衡学习技术来提升模型性能。目前，我们的模型尚未进行任何病灶分割，我们也将尝试开展这项工作，以探索眼底图像中由糖尿病视网膜病变引起的一些隐含特征。

附录 A. 定性可视化

如第 4.2.2 小节所述，我们在图 A.10 中展示了所采用的各个图像变换器的 Grad-CAM 热图。

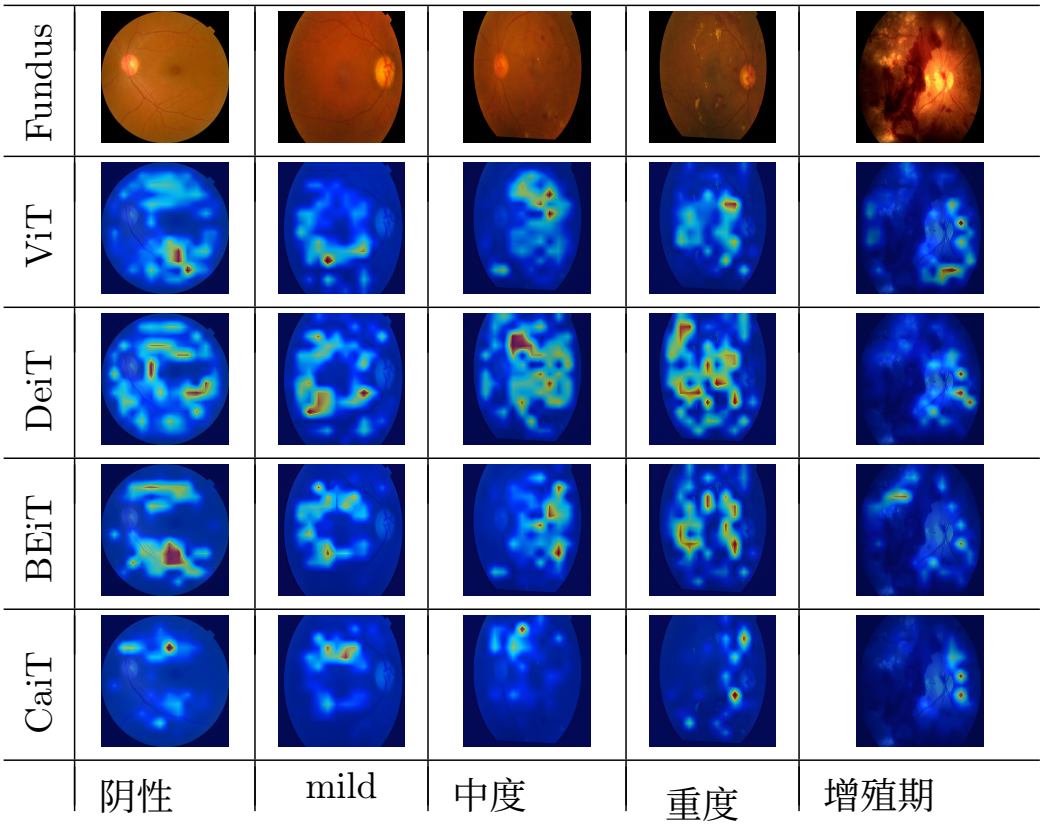

图 A.10：眼底图像（1st 行）及 ViT、DeiT、BEiT、CaiT 的 Grad-CAM热图，分别如 2nd、3rd、4th、5th 行所示

附录 B. 调优加权平均集成 Transformer

如第 4.2.4 小节所述，我们在表 B.7 中提供了经调优的 α_j 值，这些值有助于通过加权平均方案（EiT_{wm}）集成 Transformer，从而实现最佳性能。

表 B.7：调优权重 α_j，用于 EiT_{wm}

Transformer_{wm}	α_1	α_2	α_3	α_4
ViT +DeiT	0.25	0.75	-	-
ViT +BEiT	0.4	0.6	-	-
ViT + CaiT	0.4	0.6	-	-
DeiT +BEiT	0.4	0.6	-	-
DeiT +CaiT	0.3	0.7	-	-
BEiT +CaiT	0.5	0.5	-	-
ViT + DeiT + BEiT	0.2	0.3	0.5	-
ViT + DeiT +CaiT	0.2	0.3	0.5	-
ViT + BEiT + CaiT	0.2	0.4	0.4	-
DeiT + BEiT +CaiT	0.3	0.3	0.4	-
ViT + DeiT + BEiT + CaiT	0.1	0.1	0.4	0.4

参考文献

[1] S. Stolte, R. Fang, 《糖尿病视网膜病变医学图像分析综述》，《医学图像分析》64（2020年）101742。

[2] N. Asiri 等，《糖尿病视网膜病变深度学习计算机辅助诊断系统综述》，《医学人工智能》99（2019年）101701。

[3] Y. Zheng 等，《糖尿病视网膜病变的全球流行》，《印度眼科学杂志》60（2012年）428。

[4] D. S. W. Ting 等，《利用多族裔糖尿病人群的视网膜图像开发并验证用于检测糖尿病视网膜病变及相关眼病的深度学习系统》，《美国医学会杂志》318（2017年）2211–2223。

[5] B. Tymchenko 等，《糖尿病视网膜病变检测的深度学习方法》，载于：ICPRAM，2020年，第 501–509 页。

[6] J. Krause 等，《评估糖尿病视网膜病变机器学习模型时评估者间差异与参考标准的重要性》，《眼科学》125卷（2018年），1264–1272。

[7]APTOS 2019 失明检测（2019年），最后检索日期：2023年3月8日。网址：
h
t
t
p
s://www.kaggle.com/competitions/aptos2019-blindness-detection。
[8] G. Quellec 等，《用于检测视网膜照片中微动脉瘤的最佳小波变换》，《IEEE医学影像汇刊》27卷（2008年），1230–1241。[9] L. Giancardo 等，《基于拉东变换分类的视网膜图像微动脉瘤检测》，载于：EMBS，2011年，第5939–5942页。[10] R. Casanova 等，《随机森林方法在糖尿病视网膜病变分类分析中的应用》，PLOS ONE 9卷（2014年），e98587。[11] M. Niemeijer 等，《为糖尿病视网膜病变诊断而在数字彩色眼底照片中自动检测并区分玻璃疣、渗出物和棉絮斑》，《眼科研究与视觉科学》48卷（2007年），2260–2267。[12] M. U. Akram 等，《用于糖尿病视网膜病变早期检测的微动脉瘤识别与分类》，模式识别46卷（2013年），107–116。[13] D. Usher 等，《数字视网膜图像中糖尿病视网膜病变的自动检测：糖尿病视网膜病变筛查工具》，《糖尿病医学》21卷（2004年），84–90。[14] J. D. Bodapati 等，《深度卷积特征聚合：在糖尿病视网膜病变严重程度预测中的应用》，《信号、图像与视频处理》15卷（2021年），923–930。[15] A. Zhang、Z. C. Lipton、M. Li、A. J. Smola，《动手学深度学习》，arXiv:2106.11342（2021年）。[16] S. Yu 等，《利用卷积神经网络检测糖尿病视网膜病变中的渗出物》，载于：EMBC，2017年，第1744–1747页。[17] E. Goceri，《用于图像分类的胶囊网络分析》，载于：第15届国际计算机图形学、可视化、计算机视觉与图像处理会议（CVGCVIP），2021年，第53–60页。

[18] Z. Wang 等，Zoom-in-Net：通过深度挖掘病灶检测糖尿病视网膜病变，载于：MICCAI，2017年，第267–275页。[19]M. M. Farag 等，基于 DenseNet 和卷积块注意力模块的糖尿病视网膜病变自动严重程度分类，《IEEE Access》10（2022年）38299–38308。[20] A. Dosovitskiy 等，一张图像胜过16×16个词：用于大规模图像识别的 Transformer，arXiv:2010.11929（2020年）。[21] A. M. Braşoveanu、R. Andonie，NLP Transformer 可视化：简要综述，载于：国际信息可视化会议（IV），IEEE，2020年，第270–279页。[22] H. Touvron 等，深入探索图像 Transformer，载于：ICCV，2021年，第32–42页。[23] H. Touvron 等，通过注意力训练数据高效图像 Transformer 并进行蒸馏，载于：ICML，2021年，第10347–10357页。[24] H. Bao、L. Dong、F. Wei，BEiT：图像 Transformer 的 BERT 预训练，载于：ICLR，arXiv:2106.08254，2022年。[25] J. Devlin 等，BERT：用于语言理解的深度双向 Transformer 预训练，arXiv:1810.04805（2018年）。[26]根据立体彩色眼底照片对糖尿病视网膜病变进行分级——修订版艾利之家分类法的扩展：ETDRS报告第10号，《眼科学》98（1991年）786–806。[27] U. R. Acharya 等，利用离散小波变换、离散余弦变换特征和黄斑病变指数的糖尿病性黄斑水肿（DME）自动分级系统，《计算机生物学与医学》84（2017年）59–68。[28] K. M. Adal 等，一种自动检测和分类纵向眼底图像中由红色病灶导致的视网膜变化的系统，《IEEE生物医学工程汇刊》65（2017年）1382–1390。[29] L. Tang 等，Splat特征分类及其在眼底图像视网膜出血检测中的应用，《IEEE医学影像汇刊》32（2013年）364–375。

[30] A. Herliana 等，利用粒子群优化和神经网络进行糖尿病视网膜病变特征选择，载于：CITSM，IEEE，2018年，第1–4页。[31] S. Sanromà 等，利用随机森林评估糖尿病视网膜病变风险，载于：ESANN，2016年。[32] A. Osareh 等，基于计算智能的方法检测糖尿病视网膜病变图像中的渗出物，《IEEE生物医学信息技术汇刊》13（2009年）535–545。[33] P. Chudzik 等，利用深度学习和交错冻结进行微动脉瘤检测，载于：《医学成像2018：图像处理》，第10574卷，SPIE，2018年，第379–387页。[34] R. Gargeya、T. Leng，利用深度学习自动识别糖尿病视网膜病变，《眼科学》124（2017年）962–969。[35] S. Wan 等，利用图像分类的深度卷积神经网络检测糖尿病视网膜病变，《计算机与电气工程》72（2018年）274–282。[36] S. H. Kassani 等，使用改进的 Xception 架构进行糖尿病视网膜病变分类，载于：ISSPIT，2019年，第1–6页。[37] M. D. Alahmadi，糖尿病视网膜病变分类的纹理注意力网络，《IEEE Access》10（2022年）55522–55532。[38] M. Tan、Q. Le，EfficientNet：重新思考卷积神经网络的模型缩放，载于：ICML，2019年，第6105–6114页。[39] J. Hu 等，压缩与激励网络，载于：CVPR，2018年，第7132–7141页。[40] S. Zhao 等，CoT-XNet：用于糖尿病视网膜病变分级的上下文 Transformer 与 Xception 网络，《医学与生物学物理》（2022年）。[41] Y. Huang 等，SSiT：用于糖尿病视网膜病变分级的显著性引导自监督图像 Transformer，arXiv:2210.10969（2022年）。[42] I. Goodfellow 等，《深度学习》，MIT出版社，2016年。

[43] E. Goceri, 《医学图像数据增强：技术、比较与解读》，《人工智能评论》（2023年）1–45页。[44] V. Stimper 等, 《多维对比度受限自适应直方图均衡化》，IEEE Access 7（2019年）165437–165447页。[45] E. Goceri, 《利用空间变化分布匹配对脑部磁共振图像进行强度归一化》，载于：CVGCVIP，2017年，第300–304页。[46] E. Goceri, 《评估去噪技术以去除皮肤镜图像中的斑点噪声和高斯噪声》，《生物与医学中的计算机》152（2023年）106474。[47] D. Sarvamangala 等, 《医学图像理解中的卷积神经网络：综述》，Evolutionary intel.（2021年）1–22页。[48] T. Wolf 等, 《Transformer：自然语言处理的最新进展》，载于：EMNLP：系统演示，2020年，第38–45页。[49] A. Vaswani 等, 《注意力就是一切》，《神经信息处理系统进展》第30卷（2017年）。[50]J. L. Ba、J. R. Kiros、G. E. Hinton, 《层归一化》，arXiv:1607.06450（2016年）。[51] D. Misra, 《Mish：一种自正则化非单调神经激活函数》，BMVC，论文编号#928（2020年）。[52] G. Hinton、O. Vinyals、J. Dean, 《神经网络中的知识蒸馏》，载于：NIPS深度学习与表征学习研讨会，arXiv:1503.02531，2015年。[53] A. Ramesh 等, 《零样本文本到图像生成》，载于：ICML，2021年，第8821–8831页。[54] L. Rokach, 《基于集成的分类器》，《人工智能评论》33（2010年）1–39页。[55] I. Loshchilov、F. Hutter, 《解耦权重衰减正则化》，ICLR, arXiv:1711.05101（2019年）。

[56] R. Selvaraju 等，《Grad-CAM：通过基于梯度的定位对深度网络进行可视化解释》，载于：ICCV，2017年，第618–626页。[57] M. Grandini、E. Bagli、G. Visani，《多分类指标：概述》，arXiv预印本 arXiv:2008.05756（2020年）。[58] K. He 等，《用于图像识别的深度残差学习》，载于：CVPR，2016年，第770–778页。[59] C. Szegedy 等，《重新思考用于计算机视觉的Inception架构》，载于：CVPR，2016年，第2818–2826页。[60] M. Sandler 等，《MobileNetV2：倒置残差与线性瓶颈》，载于：CVPR，2018年，第4510–4520页。[61] F. Chollet，《Xception：采用深度可分离卷积的深度学习》，载于：CVPR，2017年，第1251–1258页。[62] L. Zahedi 等，《自动化超参数调优的搜索算法》，载于：国际数据科学会议（ICDATA），arXiv:2104.14677，2021年，第1–10页。[63] E. Decencière 等，《对公开发布的图像数据库的反馈：Messidor数据库》，《图像分析与立体学》33（2014年）231–234页。[64] J. Krause 等，《糖尿病视网膜病变机器学习模型评估中评估者间差异及参考标准的重要性》，《眼科学》125（2018年）1264–1272页。[65] P. Porwal 等，《印度糖尿病视网膜病变图像数据集（IDRiD）》，IEEE Dataport（2018年）。DOI：10.21227/H25W98。