发布日期 xxxx 00, 0000，当前版本日期 xxxx 00, 0000。数字对象标识符
10.1109/ACCESS.2017.DOI

高效 Scopeformer：面向颅内出血检测的可扩展丰富特征提取

YASSINE BARHOUMI[1]，[3][2]，**NIDHAL C. BOUAYNAYA 和 GHULAM RASOOL**[3]。[1]美国新泽西州罗文大学电气与计算机科学系（电子邮件：barhou29@students.rowan.edu）[2]美国新泽西州罗文大学电气与计算机科学系（电子邮件：bouaynaya@rowan.edu）[3]美国佛罗里达州坦帕市莫菲特癌症中心机器学习部（Ghulam.Rasool@moffitt.org）

通讯作者：Yassine Barhoumi（电子邮件：barhou29@students.rowan.edu）。

arXiv:2302.00220v1 [cs.CV] 1 Feb 2023

摘要 卷积神经网络（CNN）和视觉 Transformer（ViT）提取的特征图的质量与丰富度直接关系到模型的稳健性能。在医学计算机视觉领域，这些信息丰富的特征对于从大型数据集中检测罕见病例至关重要。本文提出了“Scopeformer”，一种用于计算机断层扫描（CT）图像颅内出血分类的新型多 CNN-ViT 模型。Scopeformer 架构具有可扩展性和模块化特性，能够采用各种 CNN 架构作为骨干网络，并利用多样化的输出特征和预训练策略。我们提出了有效的特征投影方法，以减少 CNN 生成特征之间的冗余，并控制 ViT 的输入大小。对各种 Scopeformer 模型进行的大量实验表明，模型性能与特征提取器中使用的卷积块数量成正比。通过采用多种策略，包括使 CNN 的预训练范式多样化、使用不同的预训练数据集以及应用风格迁移技术，我们证明了在不同计算预算下模型性能均得到整体提升。随后，我们提出了更小、更节省计算资源的 Scopeformer 版本，涵盖三种不同类型的 ViT 输入和输出配置。高效 Scopeformer 使用四种不同的预训练 CNN 架构作为特征提取器，以提升特征丰富度。我们表现最佳的高效 Scopeformer 模型达到了 96.94% 的准确率和 0.083 的加权对数损失；与基础 Scopeformer 相比，其可训练参数数量减少了八倍。另一种高效 Scopeformer 模型进一步将参数空间缩减了近 17 倍，同时性能几乎没有下降。总之，我们的研究表明，由 CNN 和 ViT 组成的混合架构或可提供所需的特征丰富度，从而构建准确的医学计算机视觉模型。

索引词 计算机断层扫描(CT)；颅内出血；医学成像；卷积神经网络；视觉变换器；特征图

I. 引言

卒中是全球范围内导致死亡的主要原因之一 [1]。这种疾病的严重性使得早期检测脑出血对于患者预后至关重要。利用头部计算机断层扫描（CT）扫描对颅内出血进行早期检测和分类，可以降低发生严重医学并发症和大范围脑损伤的风险，尤其是在最初24小时内 [1], [2]。早期检测和分类通常需要具备资质的医师评估并检测颅骨内是否存在出血迹象，或脑组织内是否存在病灶。机器学习算法经过训练后能够自主识别和分类脑出血，从而缩短检测时间，实现更快速、更有效的治疗。新兴的计算机视觉技术能够提供更快速、更稳健的模型，用于对患者进行分诊，并帮助专业医师和放射科医师高效地安排时间 [3]–[5]。

近年来，卷积神经网络（CNNs）和视觉变换器（ViTs）凭借处理和分析大量视觉数据的能力，在计算机视觉领域受到了广泛关注。卷积神经网络利用卷积运算和池化层从图像中提取

具有判别性且富有意义的特征。随后，这些特征被用于图像分类或目标识别。另一方面，视觉 Transformer（ViTs）引入了借鉴自自然语言处理（NLP）中 Transformer 的自注意力机制，在进行预测时动态衡量图像各个部分的重要性。这种方法已被证明在图像分类和目标识别任务中十分有效，展现了视觉 Transformer 解决更复杂计算机视觉问题的潜力。

受卷积神经网络和视觉 Transformer 性能的启发，我们提出了一种混合架构，由多个卷积神经网络和一个多编码器 ViT模型组成。该模型称为 Scopeformer，可容纳 (n) 个专用于特征提取的卷积神经网络，以及多个堆叠的 ViT 编码器，用于从全局特征图中差异化地提取权重。这些权重代表模型学习到的特征间相关性，并被认为与出血分类问题相关。结果表明，在 Scopeformer训练中，用于提取特征的卷积神经网络数量越多，分类准确率越高，但计算需求也随之增加。本文介绍了 Scopeformer模型，并探讨了选择性特征工程方法，以生成更丰富的分类内容。我们还解决了增加用于特征提取的卷积神经网络数量所导致的大型可训练参数空间问题。我们采用降维技术处理卷积特征空间，以控制 ViT 编码器内的注意力复杂度。通过实验，我们构建了一种可扩展且高效的基于多卷积网络的混合 ViT 模型（n-CNN-ViT），用于解决出血检测问题。我们的贡献可概括如下：

- 提出了一种名为 Scopeformer 的混合架构，将多个卷积神经网络与多编码器 ViT模型相结合，用于检测 CT 图像中的颅内出血。
- 结果表明，分类准确率与用于特征提取的卷积神经网络数量成正比，但这也导致计算需求更高。
- 通过对卷积特征空间采用降维技术来控制 ViT 编码器中的注意力复杂度，从而解决可训练参数空间过大的问题。
- 针对出血检测问题，开发了一种可扩展且高效的基于多卷积的混合 ViT模型，从而提高了分类准确率。

II. 背景与文献综述

A. RSNA 颅内出血检测

2019年，北美放射学会（RSNA）提供了大量健康参与者和患有各种类型脑内出血的患者的脑部 CT 扫描。RSNA 举办了一项机器学习挑战赛，以促进用于多类别出血分类的自主算法开发 [6]。这些计算机化的多标签分类器旨在确定输入 CT 图像的每个二维（2D）切片中是否存在脑出血，并给出一个包含六个分量、分别对应分类目标的概率向量。

B. 卷积神经网络（*CNN*）

直到最近，在计算机视觉应用中，卷积神经网络（卷积神经网络）一直是为下游任务（例如分类）提取高分辨率特征的事实标准模型 [7]–[12]。发布在 Kaggle 在线社区平台上的 RSNA 挑战赛官方排名靠前的解决方案，采用了包含基于卷积的特征提取阶段的多阶段分类模型 [22]。卷积层的不同堆叠方式和排列方式会产生不同的特征。这些特征会受到各种配置实现变体的影响，例如架构结构、控制视觉信息流的参数以及模型深度 [26]。基于扩大感知场、提高特征提取效率以及减少可训练参数空间以实现更快速和高效计算的目标，人们提出了若干现成架构 [18], [30]–[33]。增加架构深度可以更好地逼近目标，并且由于更高的非线性和更大的感受野，能够提供更优的特征表示。相关改进基于对卷积层、激活函数、损失函数、正则化方法和优化过程的设计进行引入与优化 [27]。

C. 视觉变换器（视觉 *Transformer*）

视觉变换器（ViT）正越来越多地应用于各种计算机视觉识别任务[35], [36]，并已在多种视觉任务中取得成功，例如 ImageNet 分类挑战赛 [37]。ViT 的基本工作组件是 Transformer 块[13]，，最初由 Vaswani 等人 [14] 在自然语言处理（NLP）领域提出。将 Transformer 模型 [14] 成功应用于图像，形成视觉 Transformer（ViT），是计算机视觉领域的一个里程碑 [15]。医学领域提出了多种成功的 ViT 架构，其表现大幅优于标准的卷积模型 [16]。ViT 模型 [15] 将自然图像划分为大小相同的三通道正方形图像块。这些图像块被展平，表示为一维词元。每个图像块代表原始图像的局部语义信息，模型则学习从这些图像块的相关性中提取模式。更精细的图像块能够提取更高的局部相关性并改善语义表现，但这是以模型巨大的二次复杂度为代价的。然而，这种复杂度会导致计算成本高昂且数据需求量大。研究表明，只有在预训练或训练阶段处于高数据量场景时，这类模型才会胜过标准卷积神经网络 [15]。

D. 卷积神经网络与视觉变换器

卷积神经网络（CNN）和视觉 Transformer（ViT）是计算机视觉中用于图像识别的两种模型，它们采用不同的特征提取方法。CNN 使用卷积层学习局部模式，从而形成金字塔状结构 [23], ；而 ViT 使用自注意力机制处理全局模式，从而形成柱状结构 [24]。CNN 通过池化层处理图像尺度，速度更快且内存效率更高；ViT 则先进行线性投影，再通过自注意力学习图像的定长表示，因此在学习全局模式方面更有效，且对输入尺寸的适应性更强。

卷积神经网络 [30], [39] 通过结合多次卷积运算，整合输入图像中的局部信息；视觉变换器 [15] 则利用编码器块的多头自注意力（MHSA）功能，从空间信息和非局部依赖关系中学习模式 [14]。这些模式使视觉变换器模型能够获得日益丰富的全局上下文，而无需手动构建逐层的局部特征。如 [35], 所示，对图像中的所有像素应用注意力会增强全局特征相关性的影响，使模型能够提取更相关的隐含模式。堆叠多个 ViT 编码器通常可以提升模型性能；采用适当的训练方法后，由 12 个 ViT 编码器块构成的模型，在 ImageNet分类任务上的表现优于由 30 多个瓶颈卷积块组成的 ResNet模型 [35]。然而，研究表明，通过堆叠 Transformer 块来增加视觉 Transformer 的深度，并不一定能提升模型性能 [19]。事实上，当堆叠的编码器数量超过一定阈值后，ViT模型的性能会趋于平稳并开始下降 [19]。Zhou 等人 [19]发现了注意力崩溃问题，并提出了一种名为再注意力的新机制，用于考虑注意力头之间的相关性。所提出的 DeepViT 模型呈现出延迟趋于平稳的特性，从而能够堆叠更多模块以获得更高的性能。

在我们先前的论文 [17], 中，我们介绍了 Scopeformer，这是一种结合卷积神经网络（CNN）与视觉 Transformer 优势的混合架构。该架构旨在从各种输入中提取高层特征。Scopeformer 利用 CNN 捕捉数据中局部模式的能力，以及视觉 Transformer 捕捉全局依赖关系和输入中不同部分之间关系的能力。将这两种架构结合起来，便得到一种高效且有效的模型，能够执行复杂的特征提取任务。Scopeformer 的提出标志着该领域向前迈出了重要一步，并为未来的研究与开发开辟了新途径。

III. 材料与方法

本文介绍了 Scopeformer，这是一种混合多卷积神经网络视觉变换器模型，以及其改进版本 Efficient Scopeformer。该模型通过卷积神经网络和视觉 Transformer 处理伪 RGB CT 扫描图像，以提取用于分类的高层特征。我们的重点是通过修改架构来提升性能和效率。本文分析并呈现了这些修改的结果。

A. SCOPEFORMER

我们在图 1 中展示了混合 n-CNN-ViT 模型。该模型由 n 个 CNN模型堆叠而成，以构建特征提取骨干网络。

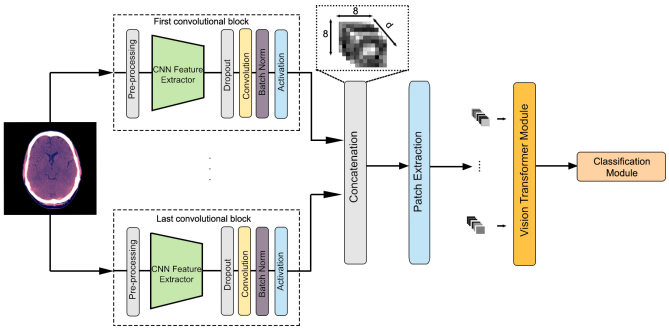


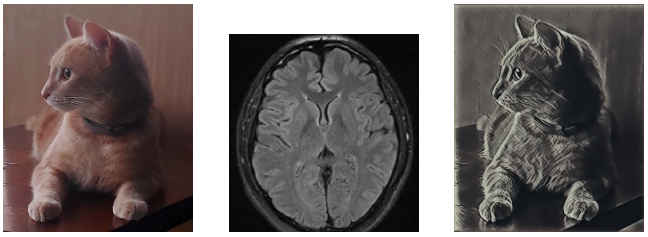
图 1：图中展示了 Scopeformer 架构的示意图。所提出的模型由四个主要模块组成：(1) Scopeformer 主干网络，(2) 图块提取，(3) 视觉 Transformer（ViT）编码器，以及 (4) 分类头。将单张输入图像送入多个 CNN模型，以提取各种特征并构建特征图。这些特征图经图块提取模块处理并向量化。生成的向量作为 Transformer 编码器的输入，模型输出则取自分类模块。

我们将 n-CNN-ViT 模型称为“Scopeformer”，其名称源自“Transformer”（-former）和“Scope”一词，后者指由具有深层感受野的卷积块生成的选择性特征提取骨干网络。Scopeformer 模型推动视觉 Transformer 和卷积神经网络取得了显著进展。Scopeformer 模型与 ViT 的主要区别在于，前者将包含更多语义信息的高层特征用作 Transformer 编码器的输入；而最初提出的 ViT 模型则以小图像块的形式输入原始自然图像。Scopeformer 模型以较小但更深的图块尺寸接收全局卷积特征图。ViT 图块提取方法沿高度和宽度方向将自然图像划分为图像块，然后将每个图像块展平，并将所有通道合并为一个一维标记。类似地，我们沿特征图的高度和宽度方向逐像素地将其划分为 $p \times p$ 个图像块，其中 p 为特征图块大小（本研究所有实验均采用 $p = 1$ ）。

模型的输入是一个维度为 $H \times W \times C$ 的张量，其中 H 表示高度， W 表示宽度， C 是拼接通道的数量

，这些通道源自 RSNA DICOM文件。模型会将输入图像并行输入不同的 CNN 架构进行前向传播，并存储输出特征 f。这些特征沿通道轴拼接。最终得到的全局特征图维度为 $h\times w\times c$，其中 h 表示特征的高度，w 表示特征的宽度，c 是特征总数，包含 $c=n\times f$。

首个 Scopeformer 架构采用 Xception卷积神经网络 [18] 和多个 ViT 层。Xception 模型由多个 Inception模块组成，这些模块包含深度卷积和逐点卷积。在我们的 Scopeformer模型中，我们在特征提取骨干网络中堆叠（n）个采用不同预训练方式的 Xception 模型 [18]，并在训练期间冻结其权重，不进行更新。我们使用嵌入 Xception 模型中的最后几个 Inception 层作为特征生成器。经过 ImageNet 预训练的 Xception卷积神经网络能够提供高层特征给 ViT模块。为此，我们认为 ViT模块的主要作用是从深度补丁中提取相关性。如图 1 所示，可以在同一个 Scopeformer 中堆叠一个或多个 Xception 模块，以生成全局特征图。我们的初步实验考虑直接堆叠卷积神经网络模块的原始特征，而不进行任何进一步处理。



（a）内容图像　（b）风格图像　(c) 输出图像

图 2：应用于 ImageNet 数据集的风格迁移方法。(a) 内容图像，(b) 风格图像，(c) 输出图像

在我们的方案中，我们倾向于对每个 Xception 卷积神经网络采用不同的预训练方法，从而为每种架构生成特定的特征。在模型训练的第一阶段，我们使用 Keras API [40] 将 ImageNet 预训练权重加载到所有卷积神经网络中。在训练的第二阶段，我们训练 Xception 卷积神经网络执行不同的分类任务，包括在 RSNA出血数据集上进行分类。我们对其中一个卷积神经网络采用硬数据增强，对其他网络采用软数据增强。我们在 ImageNet 数据集上应用风格迁移 [38]，以引入灰度类脑图像的形状偏差，如图 2 所示。生成的数据集用于对第三个卷积神经网络进行预训练。在实验中，我们测试了 Scopeformer 架构中预训练卷积神经网络的多种组合。

B. 高效 *SCOPEFORMER*

在对 Scopeformer 模型进行了大量测试后，我们提出并采用了特征提取器中的多项创新

卷积神经网络和视觉变换器模块。我们定义了四个模块，如图 1 所示。第一个模块是 Scopeformer 主干网络，由多个卷积神经网络堆叠而成，共同生成全局特征图。第二个模块用于从卷积神经网络特征中提取图块，以生成视觉变换器令牌。第三个模块由视觉变换器流程组成。最后，第四个模块代表分类头。我们将在以下各节讨论这些模块。

1) 模块 1：Scopeformer 主干网络

Efficient Scopeformer 使用多种卷积神经网络构建特征提取块。骨干卷积神经网络包括在 ImageNet 上预训练的 ResNet 152 V2 [30],EfficientNet B5[31], DenseNet 201 [32], 和 Xception [18]。每个卷积神经网络生成的特征沿通道轴拼接，形成全局特征图。然而，构建此类特征图要求每个卷积神经网络生成的特征图具有相同的高度和宽度。我们提出为每个卷积神经网络添加一个可训练的1×1 卷积层，将特征投影到合适的空间中。Efficient Scopeformer 的输入是一个维度为 $H\times$ W $\times 3$ 的张量，其中H 表示高度，W 表示宽度，3 表示通道数。图像同时输入四个卷积神经网络，以生成高层特征图。全部四个特征图的通道维度将使用 1×1 卷积层缩减至 $8\times8\times\frac{d}{4}$，其中 d 是全局特征图的尺寸。

2) 模块 2：图块提取

第二个模块的输入维度取决于第一个模块设定的全局特征图大小。在我们的实验中，得到的全局特征图是一个形状为 $8\times8\times d$ 的 3D 张量。图块提取模块沿高度和宽度方向按通道拆分特征，并提取 $N=\frac{8\times8}{p^2}$ d 维向量。我们将图块大小设为1×1，得到 $N=64$ 个词元，表示所有 d 个特征中一个局部像素位置的特征。维度 d 由前一模块使用的投影方法控制，代表架构中的瓶颈。每个图块都包含来自四个卷积神经网络生成的所有特征中，局部像素位置的语义信息。随后，将展平图块序列 $X_p\in\mathbb{R}^{64\times d}$ 用作 ViT模块的输入集。

3) 模块 3：Scopeformer ViT

我们评估了三种适用于所提出架构的不同 ViT 配置，如图 3 所示。这些配置包括：(1) Deep Scopeformer、(2) deep Scopeformer TR（转置）和 (3) Efficient Scopeformer。

基线 Scopeformer 配置： 在此配置中，我们将图像块提取层生成的一组向量输入视觉变换器编码器。我们使用了可训练的位置编码向量，并将其与向量化图像块及可训练的类别（CLS）标记结合。ViT编码器模块的输入维度为 $Y\in\mathbb{R}^{N\times d+1}$。我们使用了两种自注意力

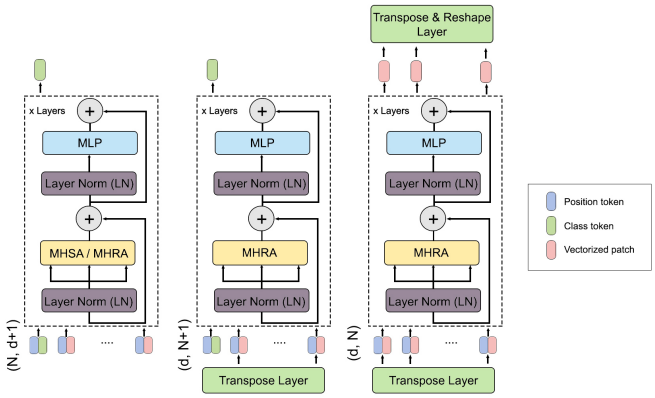


图 3：ViT Scopeformer配置。（左）基线 Scopeformer 配置：第一个配置是一个 ViT模块，其输入为从卷积神经网络特征中提取的向量化图像块。**（中）Deep Scopeformer TR 配置：**第二个配置引入了一个转置层，将按通道划分的图像块转换为按特征划分的图像块。**（右）EfficientScopeformer 配置：**第三个配置移除了令牌类别，并使用全部特征令牌作为输入。第三个模块的输出将经过转置，以恢复卷积神经网络特征的维度，然后将其输入分类模块。

变体。第一种称为多头自注意力（MHSA） [15]，第二种称为多头再注意力（MHRA） [19]。它们的关键区别在于引入了一个可训练变换矩阵。这些变体如下所示：

$$\mathrm{MHSA}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

$$\mathrm{MHRA}(Q,K,V) = \mathrm{Norm}\left(M^T\left(\mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)\right)V, \tag{2}$$

其中，$M \in \mathbb{R}^{h\times h}$ 是一个可学习的变换矩阵，h是自注意力头的数量。

Deep Scopeformer TR 配置：第二种 Scopeformer ViT 配置对图像块提取层生成的向量集合应用转置操作。转置层的输出与位置编码向量相加，并与 CLS 令牌拼接。所得向量集合的维度为 $Y_T \in \mathbb{R}^{d\times N+1}$。在实验中，我们仅使用了 MHRA自注意力变体（公式 2）。

高效Scopeformer配置： 第三个 Scopeformer 模块舍弃了先前配置中使用的 CLS 概念。在这些设置中，我们使用 ViT 编码器生成的全部特征进行分类。因此，ViT 编码器的输入和输出维度保持一致，均为 $Y_T \in \mathbb{R}^{d\times N}$。我们在 ViT 输出端使用转置与重塑层，以获得适合特征图的维度。我们使用 MHRA 自注意力变体来计算自注意力。

4) 模块 4：分类模块

基线配置和深度 Scopeformer TR 配置中的分类模块接收单个 CLS 令牌。该令牌的输出通过一个带有 Sigmoid 激活函数和单个隐藏层的多层感知机（MLP）转化为预测结果。在高效 Scopeformer 配置中，分类模块接收一组经过重塑的特征 $x_t \in \mathbb{R}^{8\times8\times d}$。分类模块先应用二维平均池化层，再接展平层。最后，通过带有 Sigmoid 激活函数的全连接层进行类别推断。

C. 数据集

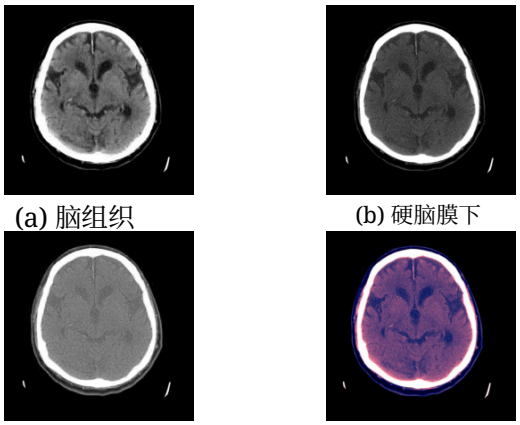
RSNA 数据集由 Adam 等人 [6] 从世界各地不同机构使用的多种扫描仪收集而来。该数据集被认为是目前公开可用的最大数据集，旨在捕捉出血亚型在现实世界中的复杂特征。该数据集于2019年在 Kaggle 平台举办的颅内出血（ICH）检测挑战赛中公开发布。该数据集包含 870,301 张经过标注的 16 位灰度计算机断层扫描（CT）图像，均以 DICOM 格式保存，并标注了五种出血类型。受训医生对每张 CT 切片标注一种或多种脑出血类型。本次竞赛需要识别五种不同类型的出血，另设一个类别表示所提供切片中存在任意类型的出血。这些类别分别标记为硬膜外出血（EDH）、脑实质内出血（IPH）、脑室内出血（IVH）、蛛网膜下腔出血（SAH）和硬膜下出血（SDH）。

D. 预处理

单张图像由像素构成，其取值范围为 0 到 2^{16}，分辨率为 256^2,，以亨氏单位（HU）表示 [7]。HU 表示扫描物质的密度。衰减 HU 值能够反映扫描内容 [21]。例如，骨骼的衰减值范围为 250 到 1000，脂肪和肌肉的衰减值（AV）范围为 50 到 100。RSNA CT扫描 DICOM 文件的元数据标签中包含 CT扫描配准过程中使用的亨氏单位范围。我们使用这些标签，确保在应用亨氏单位窗宽处理之前，数据集中的范围保持标准统一 [25]。如图 4 所示，我们将三个 HU 窗口作为 Scopeformer 模型输入的通道。我们的 HU 窗口设置为：脑部 AV$\in [40,80]$ HU、硬膜下窗 AV $\in [80,200]$ HU，以及软组织窗 AV $\in [40,380]$HU [7]。

E. 设计配置与实验

表1展示了各种Scopeformer超参数配置和架构的详细信息。我们



(a) 脑组织 (b) 硬脑膜下

(c) 软组织 (d) 堆叠伪图像

图 4：亨氏单位 CT 切片转换及其对应的堆叠三通道图像。

表1：Scopeformer 的各种设计配置——超参数和可学习参数。

模型	CNN 模块	层	特征 Size	MLP	注意力头	参数
Scopeformer (S)	4	8	516	3072	12	34 M
Scopeformer (B)	4	8	512	4096	16	42 M
Scopeformer (M)	4	8	512	5120	16	43 M
Scopeformer （L型）/4	4	4	1024	4096	16	51 M
Scopeformer (L)/8	4	8	1024	4096	16	102 M
Scopeformer (L)/16	4	16	1024	4096	16	203 M
Deep Scope-former (L型)/8	4	8	1024	4096	16	102 M
Deep Scope-former TR (L)/8	3	8	384	4096	16	**6 M**
高效 Scope-former	3	8	384	4096	16	**6 M**
Scopeformer [17]	3	12	3072	3072	8	755 M
缩放 Scope-former [17]	4	8	4096	4096	16	870 M

介绍了所提出的不同Scopeformer变体，以及特征提取骨干网络中所使用的卷积模型数量、ViT层数、全局特征图大小、每个ViT模块中的MLP维度和注意力头数量，以及可训练参数总数的详细信息。我们将高效Scopeformer的实现与初始模型实现进行了比较，并根据配置超参数提出了更小的可训练参数空间。我们的实验主要包括四个部分。

在第一组实验中，我们评估 Scopeformer 各种变体的尺寸对分类准确率的影响，共评估四种变体：小型（S）、基础型（B）、中型（M）和大型（L）。我们固定 ViT 层的数量（为 8），并通过配置 ViT 模块中 S、B 和 M 变体的 MLP 大小，以及增大 L 变体的特征尺寸，来提高模型的复杂度。可训练参数的数量从最小型（S）变体到最大型（L）变体大幅增加。

在第二组实验中，我们研究了 ViT 编码器模块数量对模型性能的影响。基于第一组实验中获得的初步结果，我们在大型 Scopeformer 变体（L）上开展消融研究，其特征尺寸为 1024，MLP 维度为 4096。我们设计了三组实验，以端到端的方式逐步堆叠 4、8 和 16 个 ViT 编码器，分别构建出三个模型：Scopeformer (L)/4、Scopeformer (L)/8 和 Scopeformer (L)/16。鉴于最大部分的模型参数位于 ViT架构中，可训练参数总数会随所使用的 ViT 模块数量呈线性缩放。

第三组实验考察了从最初提出的 ViT模型 [15] 向名为 DeepViT [19]的不同版本过渡的情况。我们在前两组实验中性能最高的模型上测试了这一配置：Scopeformer (L)/8，其全局特征图尺寸为 1024，包含 8 层 ViT 编码器，MLP 维度为 4096。该模型版本名为 Deep Scopeformer (L)/8，具有略高数量的可训练参数。

最后一项实验为我们的 Scopeformer 架构引入了三种不同的 ViT 配置，如图 6 所示。我们将这些配置应用于本研究前三部分中表现最佳的模型：Deep Scopeformer (L)/8，其全局特征图大小为 1024，包含 8 层 ViT 编码器，MLP 维度为 4096。我们引入并比较了第 2.2.3 节所述的三种 Scopeformer 配置：基线 Scopeformer 配置、Deep Scopeformer-TR配置和 Efficient Scopeformer 配置。

F. 高效 *Scopeformer* 预训练

在所有实验中，我们首先使用 ImageNet-1k 数据集 [20] 对 Scopeformer模型进行预训练。随后，我们使用 RSNA 数据集 [6] 训练全部模型。在第一个模块（卷积骨干网络）中，我们冻结每个卷积神经网络中 $\approx$ 70% 的层权重，并保留顶部 $\approx$ 30% 的权重参与训练，同时训练新引入的 1×1 卷积层。在最后一项使用 Efficient Scopeformer 模型的实验中，我们在 ImageNet-1k 默认预训练的基础上，使用 RSNA 数据集对骨干神经网络进行出血分类预训练，共训练 150 个轮次。在这项实验中，该模型记为 Efficient Scopeformer (p)，训练期间我们冻结特征提取块的权重。

G. 损失函数

遵循 RSNA 颅内出血挑战赛（ICH）的指南，我们在模型训练中采用了加权版的多标签对数损失函数。该加权方式旨在提高对第一类（代表所有类型出血）进行分类的重要性，其系数为 2，而其余类别的系数为 1。

单个样本的损失值是对各类别分别计算的二元损失进行加权平均所得。颅内出血属于多标签分类问题，即输入图像可以被分为多个类别，并通过对每个类别进行二元标记来表示其是否存在。在我们的设定中，我们对数据集采用多标签独热编码，为每个 CT 切片的每个类别赋予一个二元值。多标签对数损失函数定义如下：

$$L_{\text{multi-BCE}}(y,\tilde{y}) = -\sum_{n=1}^{6} \alpha_{\text{n}}(y_n \log \tilde{y_n} + (1-y_n)\log(1-\tilde{y_n})), \tag{3}$$

其中，α_{n} 表示目标类别的系数，y_n 表示每个类别 n 的真实值，$\tilde{y_n}$ 表示相应的预测概率。

H. 评估指标

RSNA IHC 官方采用的模型评估指标是加权准确率。我们根据三项指标评估模型的整体性能：（1）模型在 RSNA 数据集上的分类准确率；（2）对卷积主干网络生成的嵌入层全局特征丰富度进行可视化评估；（3）模型大小函数与可训练参数总数的比值。

IV. 结果与讨论

A. 骨干神经网络模型大小和预训练技术的影响

我们在特征提取骨干网络中逐步堆叠 n 各种预训练 Xception模型。我们冻结骨干网络中的所有这些架构，以防止它们的权重在训练过程中更新。我们采用多种不同的预训练方案对卷积神经网络模型进行预训练，包括 ImageNet-1k 自然图像数据集（I）和生成的风格迁移基础数据集（S）。表2比较了不同模型及其在出血分类任务上的相应表现。n-CNN-ViT模型在卷积主干网络生成的卷积特征上进行训练，而视觉变换器模型则在原始数据集上进行训练。视觉变换器模块的输入维度表示全分辨率图像，或拆分为图像块之前的特征集合。结果表明，使用卷积模型提取特征来训练视觉变换器模型，比使用原始数据集更有效。Scopeformer模型利用预训练生成适用于视觉变换器架构的高层特征。使用卷积神经网络可减少对高数据量场景的需求，因为视觉变换器模型用于拟合这些高层特征并提取语义相关性，而不是在训练中学习空间特征。此外，结果表明，分类准确率与Scopeformer训练中使用的卷积神经网络模型数量成正比；也就是说，随着我们在模型骨干网络中堆叠更多特征提取架构，出血分类的表现会更好。我们还通过有选择地改变各CNN架构的预训练方法，进一步提升了这些表现。我们推测，增大视觉变换器输入中的特征图尺寸，可使视觉变换器模块提取更多语义相关性。此外，不同预训练CNN架构所产生的归纳偏置各不相同；对这些归纳偏置进行多样化处理，可能会得到不同的特征图集合，从而形成更丰富的特征图，并带来观察到的性能提升。

表2：基于视觉变换器的模型在 RSNA 验证数据集上的分类性能。(I) 表示在 ImageNet 上进行预训练。(S) 表示在 ImageNet 和风格迁移图像上进行预训练。

模型	视觉变换器输入维度	验证准确率	Loss
ViT	256×256×3	94.33%	0.1822
1-CNN-ViT (S)	7×7×1024	96.95%	0.08272
2-CNN-ViT (I-I)	7×7×2048	97.22%	0.07984
2-CNN-ViT (S-S)	7×7×2048	97.26%	0.07934
2-CNN-ViT (I-S)	7×7×2048	97.46%	0.07754
3-卷积神经网络-视觉变换器 (I-I-S)	7×7×3072	98.04%	0.07050

B. Scopeformer 尺寸的影响

表3和表4展示了使用 Scopeformer 模型不同变体进行实验的结果。表4呈现了 S、B 和 M 型模型在各个类别上的不同结果。我们提出了四种尺寸的 Scopeformer 模型：S、B、M 和 L。与最初采用多个基于 Xception 的卷积神经网络的 Scopeformer 模型实现相比，这些模型的可训练参数数量有所减少。参数减少的关键在于：在特征提取骨干网络中，每个卷积架构之后、拼接之前都加入了可训练的 1×1 卷积层。在这项实验中，我们通过调整 ViT 模块中的 MLP 维度和自注意力头数量，逐步增加 S、B 和 M 型变体的模型复杂度，具体如表1所示。

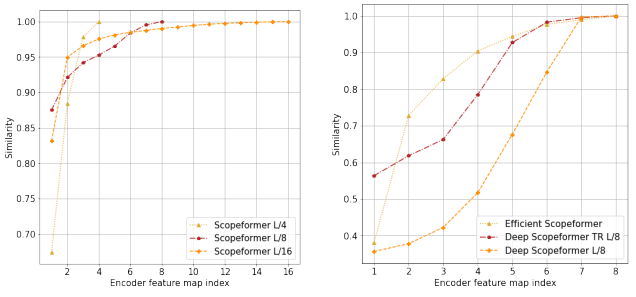
表3：不同 Scopeformer 变体的性能。

模型	准确率	Loss	召回率	可训练参数-ters
小型（S）	93.00%	0.1703	84.95%	34M
基础型（B）	93.92%	0.1461	89.29%	42M
中 (M)	93.88%	0.2285	88.44%	43M
大型 (L型)/4	93.12%	0.1378	87.81%	51M
大型 (L型)/8	**94.69%**	**0.1197**	**89.33%**	102M
大型 (L型)/16	92.57%	0.1395	87.34%	203M

在表3中，我们注意到基础模型的表现优于小型和中型变体。然而，从表4可以看出，基础模型在 IPH、IVH 和 SAH 类别上的表现更好，而小型模型在硬膜外、SDH 及全部类别上的准确率更高。根据这些观察，我们推测，MLP 维度增大后性能有所提升，这表明模型能够涵盖更多信息，并提取有助于分类的语义信息。然而，当 MLP 维度达到 5120 时，模型表现出过拟合迹象。基于这些结果，我们采用基础变体的配置，并将全局特征维度设为 $d = 1024$，构建了大型 Scopeformer (L)/8 模型。特征尺寸的增加使模型的可训练参数数量相应增加。在所提出的各个变体中，大型模型 (L)/8 的表现最佳。随着输入特征嵌入空间增大，在更大 ViT 尺寸下观察到的性能提升表明，新增特征带来了更丰富的信息，使模型得以提取有助于分类的语义信息。正如表4所示，扩大特征空间提升了部分类别的表现，但以牺牲其他类别为代价。

C. ViT 编码器数量的影响

我们使用 4 个、8 个和 16 个 ViT 编码器，评估 ViT 编码器数量对 Scopeformer (L)/8 模型的影响。如表1和表3所示，参数数量随编码器数量呈线性增长。我们注意到，使用 8 个 ViT 编码器比使用 4 个 ViT 编码器的浅层模型取得了更好的结果。然而，使用 16 个 ViT 编码器的深层模型会大幅降低模型性能。我们得出结论：增加 ViT模型的深度并不会呈线性扩展，并且存在一个使模型性能达到最优的 ViT 编码器临界数量。



(a) Scopeformer 变体　(b) 深层 Scopeformer 变体

图5：ViT 编码器特征图相对于最后一个编码器特征图的余弦相似度。我们观察到，随着 Scopeformer模型深度的增加，各 ViT 编码器特征之间的相似度也随之提高。

在图5中，我们绘制了各个 ViT 编码器生成的特征与模型最后一层之间的余弦相似度。我们观察到，对于所有提出的模型，各个 ViT 编码器生成的特征之间的相似度都迅速增加。ViT 编码器数量更多的模型中，这些相似度还会进一步提高。我们认为，这种增加的 Scopeformer(L)/16 模型的特征之间的相似度可能导致了表3中观察到的性能下降。同样，Scopeformer(L)/4 上观察到的 ViT 特征相似度降低，或许可以解释其表现不佳的原因。根据这些结果，我们得出结论，余弦相似度可以作为衡量模型性能的有效指标，因为相似度降低或升高都可能表明 Scopeformer模型的性能不佳。相似度较低的浅层模型可能暗示，叠加更多 ViT层能够提升性能；而较深的模型可能需要更多数据来降低 ViT 特征之间的相似度，从而实现最佳性能。结果还表明，Scopeformer模型的 ViT 编码器数量存在一个最优值，该值取决于数据集的复杂度以及卷积骨干网络的有效性。

D. 两种不同自注意力变体的影响

Deep Scopeformer (L)/8 在 Scopeformer (L)/8 模型的基础上，将多头自注意力层替换为多头关系注意力层。额外的可训练矩阵 M 仅为 Scopeformer (L)/8 模型增加了极少量参数。图5(b)中，Deep Scopeformer (L)/8 模型的视觉变换器编码器特征之间呈现出显著差异。这一结果可能表明，模型通过 MHRA 注意力头之间额外的相互关联，获得了更高的特征丰富度。如表5所示，该配置使准确率提高了 $+1.11\%$。

E. ViT Scopeformer 配置

我们在 ViT 模块前引入转置层，以解决自注意力计算复杂度问题。Deep Scopeformer (L)/8 中的注意力权重矩阵维度为 1024^2。在第二种和第三种 ViT 配置中，注意力权重矩阵的维度分别为 65^2 和 64^2。如表1所示，使用转置层显著减少了可训练参数的数量，这得益于 MHRA 在计算复杂度上的二次方缩减。此外，对输入序列进行转置，有效保留了特征提取器模块提取的特征内容，并维持了分类性能。表5展示了提出的三种配置的性能。与 Deep Scopeformer (L)/8 相比，提出的 Efficient Scopeformer 变体在可训练参数空间更小的情况下表现相对更好。我们推测，在该配置中，ViT 模块的作用是改进此前由卷积主干网络优化的全局特征图。全局特征图的改进源于利用注意力计算生成新特征，这些特征体现了卷积网络所生成的所有特征之间的相互关联。

我们的高效Transformer模块改善了全局特征图之间的相关性，从而提升了性能。我们注意到，对于模型 Efficient Scopeformer (P)

在目标数据集上预训练卷积块，并在训练期间冻结整个卷积块，其性能优于端到端训练；后者使用了 Efficient Scopeformer 卷积块约 30% 的可训练参数。我们认为，骨干卷积神经网络与视觉 Transformer 呈现出不同的动态特性，因此需要采用不同的模型训练设置。

表4：各目标类别上的模型性能

	准确率			
	大型	中	Base	小
All	**71.34%**	60.26%	70.5%	70.83%
硬膜外	96.98%	90.18%	95.73%	**98.08%**
IPH	85.94%	71.10%	**87.28%**	85.95%
IVH	90.5%	70.73%	**91.72%**	90.13%
SAH	**78.69%**	65.49%	78.57%	77.04%
SDH	**77.08%**	60.78%	74.35%	74.54%

表5：不同 Scopeformer 模态下的模型性能

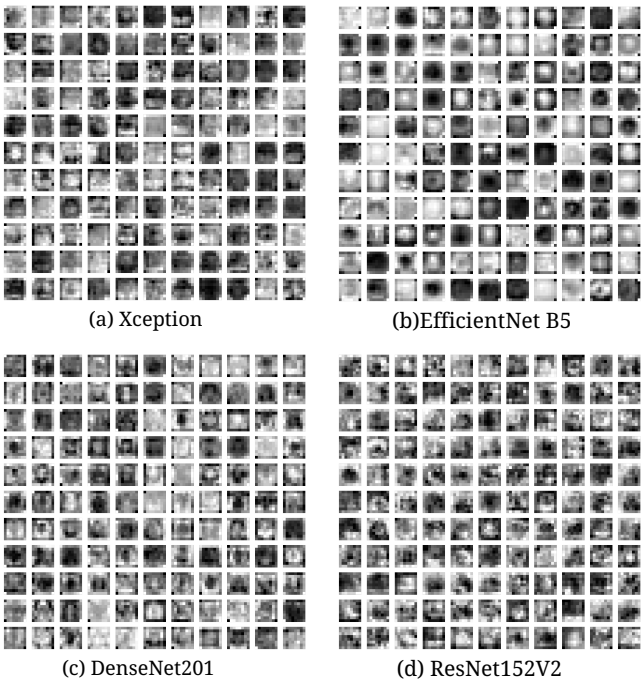
	准确率	Loss	可训练参-数
Scopeformer（L型）/8	94.69%	0.1197	102M
Deep Scopeformer（L型）/8	96.03%	0.1088	102M
Deep Scopeformer TR (L)/8	95.40%	0.1176	**6M**
高效 Scopeformer	95.77%	0.1160	**6M**
高效 Scopeformer (P)	**96.94%**	**0.0833**	**5M**

1) 全局和 ViT 特征图

图 6、7 和 8 展示了三种 Scopeformer 架构为一个硬膜外病例生成的卷积特征：Scopeformer (L)/8、Deep Scopeformer (L)/8 和 Efficient Scopeformer。我们观察到，每种 CNN 架构生成的特征都具有高度多样性。此外，我们还发现，对于所有 Scopeformer 变体，不同 CNN 生成的特征之间没有明显的相似性。因此，生成的全局特征图冗余度较低，特征丰富度较高。然而，在这些模型中，我们注意到 DenseNet 模型在所观察到的特征中表现出最高的特征冗余度。因此，我们对 Deep Scopeformer TR 进行了消融研究，结果是从 Efficient Scopeformer 模型骨干网络中移除了 DenseNet201 模型。

2) 注意力模式可视化

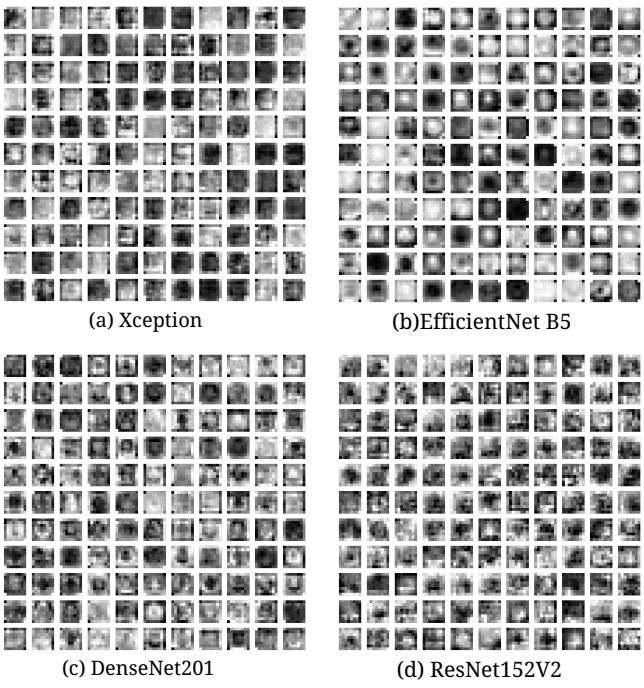
图 9 展示了第一个和最后一个 ViT 编码器中 16 个 MHRA 注意力头的注意力模式可视化结果。在第一个 ViT 层中，我们观察到，模型从每种 CNN 架构中提取的特征之间具有较高的相关性。这一观察表明，



(a) Xception (b)EfficientNet B5

(c) DenseNet201 (d) ResNet152V2

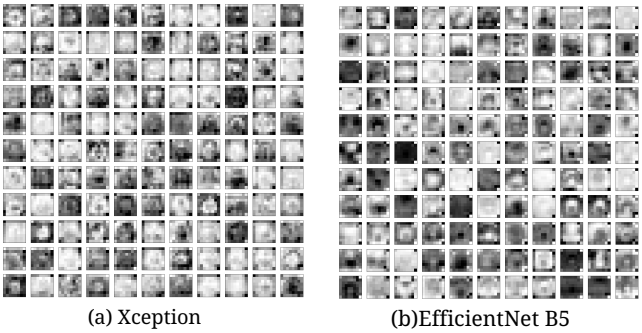
图 6：硬膜外出血示例的特征图可视化。Scopeformer(L)/8



(a) Xception (b)EfficientNet B5

(c) DenseNet201 (d) ResNet152V2

图7：硬膜外出血示例的特征图可视化。Deep Scopeformer(L)/8



(a) Xception (b)EfficientNet B5

(c) ResNet152V2

图 8：硬膜外出血示例的特征图可视化。EfficientScopeformer

每种 CNN 模型的输入特征之间具有较高的相似性。每个注意力头都会学习特征集合中不同的相关模式。然而，随着模型层次加深，我们观察到模型学会提取所有 CNN 特征之间的全局相关模式。生成的特征集合补充了每个特征相对于其他特征的相关性信息，这有助于提升模型性能。

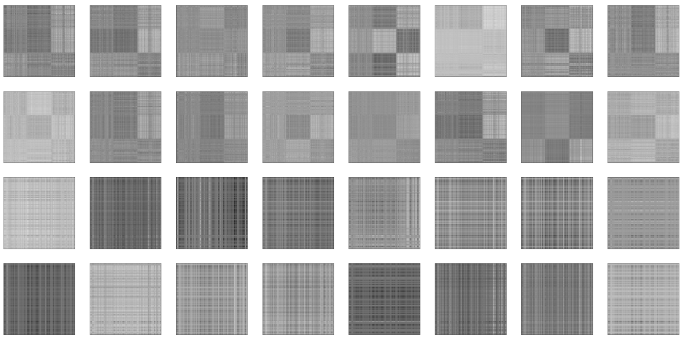


图 9：Efficient Scopeformer模型的注意力模式可视化。第一行和第二行展示第一个编码器层的 16 个注意力头。第三行和第四行展示最后一个编码器层的 16 个注意力头。每个注意力图的维度为 384×384。

V. 结论

我们提出了一组基于卷积的视觉变换器模型，称为 Scopeformer，用于解决脑部 CT 扫描中出血类型分类这一颇具挑战性的问题。我们为卷积神经网络和视觉 Transformer 定义了一系列模型架构。我们探究了使用多个现成卷积神经网络模型对架构全局特征丰富度的影响，并研究了一种特征投影方法，以将大型冗余特征空间缩减为更低维、更高效的空间。我们开展了参数优化研究，以评估模型大小对性能和效率的影响。我们实现了三种 ViT 配置，以评估 Scopeformer 模型中的再注意力模块，以及全局特征图的按通道与按特征提取图块方法。结果表明，不同的 CNN 架构提升了特征丰富度。再注意力模块增加了 ViT 特征之间的差异，从而提升了性能，并支持构建更深的模型。采用我们提出的按特征提取图块方法后，模型大小缩减至原来的 1/17，同时性能相当。我们提出的高效 Transformer 模块改善了全局特征图之间的相关性，并有助于提升性能。此外，我们发现，在目标数据集上对卷积块进行预训练，并在训练期间冻结整个卷积块，其效果优于端到端训练；前者仅需使用 Efficient Scopeformer 卷积块中 30% 的可训练参数。

致谢

本研究部分得到美国国家科学基金会项目 ECCS-1903466 和 OAC-2008690 的资助。

参考文献

[1] Elliott, J. and Smith, M.，《脑内出血的急性期管理：临床综述》。《麻醉与镇痛》**2010年**，10，1419-1427。[2] Wardlaw, J.M.，《Cochrane溶栓荟萃分析概述》。《神经病学》**2001年**，10，pp.S69-S76。[3] Gong, T., Liu, R., Tan, C.L., Farzad, N., Lee, C.K., Pang, B.C., Tian, Q., Tang, S. and Zhang, Z.，《头部创伤脑部CT图像分类》。《Springer》**2007年**，10，pp. 401-408。[4] Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P. and Warier, P.，《用于检测头部CT扫描中危急征象的深度学习算法：一项回顾性研究》。《柳叶刀》**2018年**，10，pp.2388-2396。[5] Patel, A., Van De Leemput, S.C., Prokop, M., Van Ginneken, B. and Manniesing, R.，《图像级训练与预测：三维非增强CT中的颅内出血识别》。《IEEE Access》**2019年**，10，pp.92355-92364。[6] Flanders, A.E., Prevedello, L.M., Shih, G., Halabi, S.S., Kalpathy-Cramer, J., Ball, R., Mongan, J.T., Stein, A., Kitamura, F.C., Lungren, M.P. and Choudhary, G.，《通过协作构建机器学习数据集：RSNA 2019脑部CT出血挑战赛》。《放射学：人工智能》**2020年**，10，2(3)。[7] Burduja, M., Ionescu, R.T. and Verga, N.，《利用卷积神经网络和长短期记忆神经网络，在三维CT扫描中准确、高效地检测颅内出血并进行亚型分类》。《传感器》**2020年**，10，5611页。。[8] DelRocini, M., Angelini, C. and Rasool, G.，《头部计算机断层扫描异常识别》。《IEEE SPMB》**2020年**，10，pp. 1-4。[9] Carannante G., Dera D., Bouaynaya N. C., Ghulam R., Fathallah-Shaykh H.，《利用不确定性估计实现可信的医学图像分割》，arXiv预印本 arXiv:2111.05978，2021年。[10] Dera D., Ghulam R., Bouaynaya N. C.，《用于在卷积神经网络中传播不确定性的扩展变分推断》。《2019 IEEE第29届信号处理机器学习国际研讨会（MLSP）》，2019年，p.1-6。

[11] Waqas A., Dera D., Ghulam R., Bouaynaya N. C., Fathallah-Shaykh H.，《利用深度人工神经网络进行脑肿瘤分割与监测》。《生物医学数据分析的深度学习》，2021年，p.311-350。[12] Dera D., Bouaynaya N. C., Ghulam R.，《PremiUm-CNN：将不确定性传播至鲁棒卷积神经网络》，《IEEE信号处理汇刊》，2021年，p.4669-4684。[13] Ahmed S., Nielsen I. E., Tripathi A., Siddiqui S., Rasool G., Ramachandran, R. P.，《Transformer在时间序列分析中的应用：教程》。arXiv预印本 arXiv:2205.01138，2022年。[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I.，《注意力就是你所需要的一切》。《神经信息处理系统进展》**2017年**，10，30。[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J.，《一幅图像胜过16×16个词：面向大规模图像识别的Transformer》。《arXiv预印本》**2021年**，10，arXiv:2010.11929。[16] Dai, Y., Gao, Y. and Liu, F.，《TransMed：Transformer推动多模态医学图像分类》。《Diagnositics》**2021年**，10，11(8)，p.1384。[17] Barhoumi, Y. and Ghulam, R.，《Scopeformer：用于颅内出血分类的n-CNN-ViT混合模型》。《arXiv预印本》**2021年**，10，arXiv:2107.04575。[18] Chollet, F.，《Xception：采用深度可分离卷积的深度学习》。《IEEE计算机视觉与模式识别会议》**2017年**，10，1251 1258页。 - 。[19] Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q. and Feng, J.，《Deepvit：迈向更深层的视觉 Transformer》。《arXiv预印本》**2021年**，10，arXiv:2103.11886。[20] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C.，《ImageNet大规模视觉识别挑战赛》。《国际计算机视觉杂志》**2015年**，10，115(3)，pp.211-252。[21] Broder, J. and Preston, R.，《头部与脑部成像》。《急诊医师诊断成像》。《WB Saunders》
2011, 10, 1 45 pp. -
[22] https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-intracranial-hemorrhage-detection-challenge-2019。[23] Yamashita, R., Nishio, M., Do, R.K.G. et al.《卷积神经网络：概述及其在放射学中的应用》。《Insights Imaging》9，611 629 2018 10 1007 13244 018 0639 9 – ( )。https://doi.org/ 。 /s - - - 。[24] Zhang, Z. and Zhang, W.《用于医学图像分割的金字塔医学Transformer》。arXiv，2021年。https://doi.org/10.48550/arxiv.2104.14702。[25] Muschelli J.《头部CT数据处理建议》。《Front Neuroinform.》2019年9月4日；13:61。doi: 10.3389/fninf.2019.00061。PMID: 31551745；PMCID: PMC6738271。[26] Aubry, M. and Russell, B., C.，《利用计算机生成图像理解深度特征》。《IEEE国际计算机视觉会议（ICCV）论文集》**2015年**，10。[27] Jiuxiang G., Zhenhua W., Jason K., Lianyang M., Amir S., Bing S., Ting L., Xingxing W., Gang W., Jianfei C., Tsuhan C.，《卷积神经网络的最新进展》，《模式识别》，第77卷，2018年，页354-377。[28] Chen W., Liu Y., Wang W., Bakker E. M., Georgiou T., Fieguth P., Liu L., Lew M. S.，《实例检索的深度学习：综述》，《IEEE模式分析与机器智能汇刊》，2022年，p.1-20。[29] Szegedy C., Liu W,. Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A.，《Xception：采用深度可分离卷积的深度学习》，**2017年**。[30] Kaiming H., Xiangyu Z., Shaoqing R., and Jian S.，《用于图像识别的深度残差学习》。《IEEE计算机视觉与模式识别会议论文集》，页770–778，**2016年**。[31] Mingxing T., Quoc L.，《EfficientNet：重新思考卷积神经网络的模型缩放》。《国际机器学习会议》，页6105–6114。PMLR，**2019年**。[32] Huang G., Liu Z., Van der Maaten L., Weinberger K. Q.，《密集连接卷积网络》。《计算机视觉与模式识别（cs.CV）》、《机器学习（cs.LG）》、《FOS：计算机与信息科学》、《FOS：计算机与信息科学》[33] **2016** Simonyan K., Zisserman A.，《用于大规模图像识别的超深卷积网络》。《计算机视觉与模式识别（cs.CV）》、《FOS：计算机与信息科学》、《FOS：计算机与信息科学》**2014年**。[34] Haehn D., Tompkin J., Pfister H.，《使用卷积神经网络评估“图形感知”》。《IEEE可视化与计算机图形学汇刊》，25(1)，641-650，（2018年）。[35] Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexan- dre Sablayrolles 和 Hervé Jégou。通过注意力机制训练数据高效的图像 Transformer，并进行蒸馏。载于《国际机器学习会议》，第 10347–10357 页。PMLR，2021年。[36] Ze L.、Yutong L.、Yue C.、Han H.、Yixuan W.、Zheng Z.、Stephen L.、Bain- ing G.，Swin变换器：使用移位窗口的层次视觉变换器。载于《IEEE/CVF国际计算机视觉会议论文集》，第 10012–10022 页，2021年。[37] Alex K.、Ilya S.、Geoffrey E H.，采用深度卷积神经网络进行 ImageNet分类。载于《神经信息处理系统进展》，第 1097–1105 页，2012年。[38] Gatys L. A.、Ecker A. S.、Bethge, M.，《神经风格迁移算法》。计算机视觉与模式识别（cs.CV）、神经与进化计算（cs.NE）、神经元与认知（q-bio.NC）、FOS：计算机与信息科学、FOS：计算机与信息科学、FOS：生物科学、FOS：生物科学**2015年**。[39] Kaiming H.、Xiangyu Z.、Shaoqing R.、Jian S.，深度残差网络中的恒等映射。载于《欧洲计算机视觉会议》，第 630–645 页。Springer，2016年。[40] Chollet, F.等，2015年。Keras。网址：https://github.com/fchollet/keras。[41] Geirhos R.、Rubisch P.、Michaelis C.、Bethge M.、Wichmann F. A.、Bren- del W.，经ImageNet训练的卷积神经网络存在纹理偏差；增强形状偏差可提高准确率和鲁棒性。计算机视觉与模式识别（cs.CV）、人工智能（cs.AI）、机器学习（cs.LG）、神经元与认知（q-bio.NC）、机器学习（统计学）、FOS：计算机与信息科学、FOS：计算机与信息科学、FOS：生物科学、FOS：生物科学**2018年**。