光电技术及应用

DOI：10.16818/j.issn1001－5868.2023111003

结合视觉 Transformer 和 CNN 的道路裂缝检测方法

代少升[1]，刘科生[1]，余自安[2*]

（1. 重庆邮电大学 通信与信息工程学院，重庆 400065；2. 昆明云内动力股份有限公司，昆明 650200）

摘　要： 提出了一种结合视觉 Transformer 和 CNN 的道路裂缝检测方法。利用 CNN 来捕获局部的细节信息，同时利用视觉 Transformer 来捕获全局特征。通过设计的 Fusion 特征融合模块将两者提取的特征有机地结合在一起，从而解决了单独使用 CNN 或视觉 Transformer 方法存在的局限。最终将结果传递至交互式解码器，生成道路裂缝的检测结果。实验结果表明，无论是在公开的数据集上还是在自建的数据集上，相较于单独使用 CNN 或视觉 Transformer 的方法，所提出的方法在道路裂缝检测任务中有更好的效果。

关键词： 道路裂缝检测；视觉 Transformer 和 CNN；动态加权交叉特征融合

中图分类号： TP183　**文章编号：** 1001－5868(2024)02－0252－09

Road Crack Detection Method Combining A Visual Transformer and CNN

DAI Shaosheng[1]，LIU Kesheng[1]，YU Zian[2]

（1. School of Communication and Information Engineering，Chongqing University of Posts and Telecommunications，Chongqing 400065，CHN；2. Kunming Yunnei Power Co.，Ltd.，Kunming 650200，CHN）

Abstract： This study introduces an integrated approach for road crack detection that harnesses the strengths of both visual transformers and CNN. A CNN is employed to capture fine-grained details，and a visual transformer is fully utilized to capture global characteristics. A feature fusion module is then designed to seamlessly merge the extracted features from both methods，thereby addressing the limitations of using CNN or visual transformer methods separately. Finally，the results are fed into an interactive decoder to produce accurate road crack detection results. Experimental results demonstrate that，whether on a publicly available or self-constructed dataset，the proposed method demonstrates an improvement in performance as compared with using CNN or visual transformer methods separately for road crack detection tasks.

Key words： road crack detection；visual transformer and CNN；dynamic weighted cross feature fusion

0　引言

路面裂缝是道路养护中的常见问题，它可以指示路面潜在的损坏和退化情况。裂缝是各种路面病害的最初表现形式，及时发现并修复这些裂缝对避免其继续恶化、降低维护成本至关重要。传统的人工检测方法需要工作人员现场寻找裂缝并进行判断，不仅耗时费力，而且可能存在安全隐患。随着计算机视觉技术的不断发展，自动化的道路裂缝检测方法得到了广泛应用，有效地提高了检测效率和安全性。这些方法主要分为基于传统图像处理的方法和基于深度学习的方法两类。传统的图像处理方法包括基于边缘检测的方法[1]、基于区域生长的方法[2]以及基于阈值分割的方法[3]等。这些方法在特

收稿日期：2023－11－10.

* **通信作者：**余自安

定情境下可能表现出色，然而它们的鲁棒性和性能会受到多种因素的影响，如噪声、光照不均和阴影等。由于这些干扰因素的存在，这些方法通常采用相对简单的特征，难以提取出道路裂缝丰富的信息。

近年来，深度学习技术的飞速发展为道路裂缝检测带来了全新的思路和方法。深度学习方法具备自动学习特征的能力，对于噪声、光照不均和阴影等干扰因素具有更强的适应性。基于深度学习的道路裂缝检测方法已成为研究热点之一，吸引了越来越多的学者关注。相较于传统的图像处理方法，基于深度学习的道路裂缝检测方法更加出色。它们能够从图像中提取更多的特征和上下文信息，同时具备更强的泛化能力。基于深度学习的道路裂缝检测方法主要分为三类，即基于图像分类的方法[4-5]（对整个图像进行分类来判断是否存在裂缝）、基于目标检测的方法[6-7]（检测和定位裂缝的位置）以及基于像素级分割的方法[8-10]（像素级别地分割裂缝区域）。具体来说，基于图像分类的方法将整个图像作为输入，然后利用分类器对道路裂缝和非道路裂缝进行分类。而基于目标检测的方法可以检测并标记出每个裂缝的位置和边界框，但是无法捕获裂缝的分布路径、拓扑结构和形状延展等特征。相对于图像分类和目标检测的方法，基于像素级分割的方法能够对图像进行更为精细的像素级分割，可以提供更为准确的道路裂缝信息，如裂缝长度、平均宽度、最大宽度、面积等量化信息。此外，基于像素级分割的方法可以提供更加直观和清晰的检测结果，在实际应用中具有较高的实用性和可靠性。

虽然像素级分割的方法在提供道路裂缝信息方面更为细致和精确，但其仍然面临着一些挑战。首先，在道路裂缝图像中，裂缝区域所占的像素数量相对于背景像素数量较少。这种情况可能导致模型对裂缝的检测能力不足，容易产生漏检和误检等问题。其次，道路裂缝通常呈现出一定的方向性排列，为了实现对裂缝的精确检测，需要在像素级别上有效地捕捉裂缝图像中像素点之间的全局联系。然而，基于卷积神经网络（CNN）的方法通常使用固定大小的卷积核去提取输入图像的特征，固定大小的卷积核会导致感受野受限，无法捕捉图像中像素点之间的全局联系。因此，容易在裂缝检测中出现不连续分割结果的情况，影响了模型对于裂缝完整轮廓的捕获能力。此外，在进行多尺度特征融合时，现有的方法大多采用简单的方法将浅层特征和深层特征直接进行融合，忽略了不同特征之间的交互。这样的融合方式无法充分利用不同尺度特征之间的互补信息，可能导致信息的丢失或冗余，影响道路裂缝检测的精度和鲁棒性。

针对以上存在的问题，本文提出了一种结合视觉 Transformer 和 CNN 的道路裂缝检测方法。利用 Pyramid Vision Transformer（PVT）[11] 和改进的 ResNet50[12] 作为编码器模块分别提取输入图像的全局多尺度特征和局部多尺度特征。同时，设计了特征融合模块和解码器模块，以实现特征的融合和图像的恢复。这样能更全面地捕获道路裂缝图像的信息，从而提升道路裂缝分割的准确性。

1　道路裂缝数据采集及网络选取

1.1　道路裂缝数据采集及数据增强

通过道路裂缝采集车在实际道路上进行拍摄以采集道路裂缝图像。道路裂缝采集车总共收集了 1 606 张高质量的道路裂缝图像，并通过标注软件 labelme 进行了精准的手动标注，形成了完整的道路裂缝数据集。为了进一步丰富数据多样性，本文对数据集中的图像以及其对应的标签进行了随机旋转、水平翻转、垂直翻转等，用于数据集扩充的操作。

1.2　道路裂缝分割网络选取

目前，基于 CNN 的方法已成为道路裂缝图像分割的主流方法。然而，由于卷积操作局限于局部信息和权重共享，这使得网络在建模远程依赖时存在一定的局限性。为了克服这个问题，一些最新的方法[13-15] 引入了视觉 Transformer 网络来取代卷积神经网络。虽然它们成功地弥补了卷积网络在远程建模上的缺陷，但在处理需要获取图像细节信息的道路裂缝检测任务时，Transformer 网络仍然存在一定不足。

因此，本文综合考虑了视觉 Transformer 和卷积神经网络各自的优势，充分利用它们的互补特性以提高道路裂缝图像分割的准确性。

2　网络模型

图像中的语义信息和空间上下文信息对于准确检测和分割道路裂缝至关重要。传统的 CNN 受限于局部感受野和权重共享，因而在捕获全局信息方面存在一定局限性。相比之下，视觉 Transformer 虽然能通过自注意力机制获取全局信息，但其在处

理细节和局部特征方面表现相对较弱。为了解决这些问题,本文将 PVT 与改进的 ResNet50 相结合,充分发挥它们在全局和局部特征提取方面的优势。通过融合 PVT 分支和改进的 ResNet50 分支提取的特征,实现了全局和局部信息的有机结合,从而提升了模型的性能。如图 1 所示,模型包含两个并行的编码分支,分别是基于 PVT 编码结构和改进的 ResNet50 编码结构。这两个编码分支提取的特征将传递至 Fusion 融合模块进行特征融合。随后,融合后的特征将传送至解码器,执行图像的恢复以得到最终的道路裂缝分割结果。

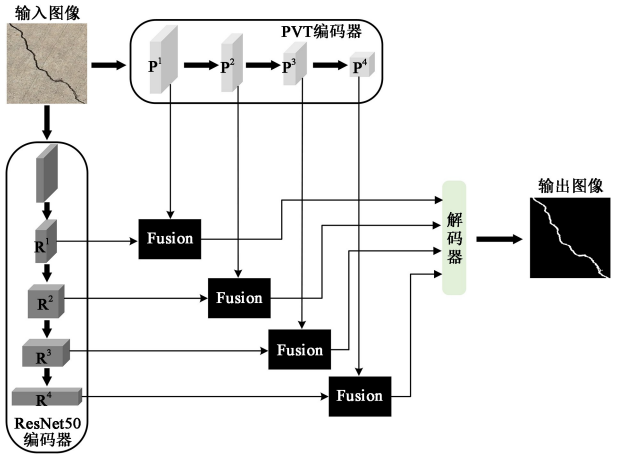


图 1 本文方法的总体结构框图

2.1 视觉 Transformer 编码分支

视觉 Transformer 编码分支选择 PVT 作为特征提取网络。PVT 在视觉 Transformer 中嵌入了金字塔结构并采用纯自注意力机制去提取多尺度特征。PVT 编码器能够对整幅图像进行全局感知,捕获道路裂缝的整体特征和分布情况,有效地解决了 CNN 编码器感受野受限的问题。这有助于提高模型对道路裂缝的识别和分割能力,特别是在处理大范围裂缝或复杂场景时表现优异。由图 1 所示,PVT 编码器由四个串联的特征提取阶段组成,每个阶段通过 Transformer 块去提取不同尺度的特征。具体地,给定一张道路裂缝图像 $I=\mathbb{R}^{3\times H\times W}$,其中 H 和 W 分别表示输入图像的高度和宽度,3 代表裂缝图像的通道数。PVT 通过 Transformer 块从道路裂缝图像中提取多尺度特征 P,可表示为 $P=\{P^i\in\mathbb{R}^{C_i\times H_i\times W_i}\mid i=1,2,3,4\}$,其中 P^i 表示第 i 个阶段提取的特征,C_i 表示第 i 个阶段所提取的特征通道数,H_i 和 W_i 分别表示第 i 个阶段所提取特征的高度和宽度。通过 PVT 编码器成功地从道路裂缝图像中提取了全局的多尺度特征,这为后续的裂缝分割过程提供了有效的特征支持,使得模型能够更准确地识别裂缝并建立全局联系。

2.2 改进的 ResNet50 编码分支

ResNet50 在道路裂缝检测中同样扮演着重要的角色。作为卷积神经网络中的一种经典架构,其深度残差结构使得模型能够更好地捕获图像中的抽象特征,从而提高了道路裂缝检测的准确性和鲁棒性。然而,考虑到道路裂缝具有明显的方向性特征,传统的卷积操作可能无法有效捕获这些方向性信息,从而导致裂缝的分割性能下降,出现分割结果不连续或者不完整的情况。因此,本文在 ResNet50 的每个阶段的 Bottleneck 中加入了轴向注意力机制[16],即 Axial-Attention 模块。轴向注意力机制可以简单地理解为首先在竖直方向进行 Self-Attention,然后再在水平方向进行 Self-Attention,通过这样的方式不仅可以提升模型对道路裂缝图像中特定方向信息的关注,还能降低计算复杂度。

2.3 Fusion 特征融合模块

本文采用了双分支的编码结构,因此一个关键问题是如何有效地融合两个编码分支提取的多尺度特征。基于此,设计了一个特征融合模块,如图 2 所示,对两个编码分支获取的特征并不是进行简单地叠加处理。假设从编码器得到的两个特征分别用 $R\in F^{H\times W\times C_1}$,$P\in F^{H\times W\times C_2}$ 表示,其中 R 表示从改进的 ResNet50 中提取的特征图,P 表示从 PVT 编码器中提取的特征图。首先,将特征图 R 和 P 分别进行卷积和批归一化操作,得到具有相同分辨率和通道数的特征图 $R'\in F^{H\times W\times C}$ 和 $P'\in F^{H\times W\times C}$。紧接着对其进行逐元素相加操作,然后对相加后得到的特征分别进行全局注意力和局部注意力计算后分别得到了全局特征 $G\in F^{1\times 1\times C}$ 和局部特征 $L\in F^{H\times W\times C}$。全局注意力机制通过将特征图中的每个位置的特征值与整个特征图的统计信息进行比较,以确定每个位置的重要性。因此,将全局注意力机制作用于改进的 ResNet50 编码器提取的特征有助于更好的关注全局信息。而局部注意力机制则通过在特定区域内调整特征值的权重来突出区域的特征信息,将局部注意力机制作用于 PVT 编码器提取的特征,有助于更加精细地突出局部区域的特征。然后将经过 Sigmoid 归一化的特征 G 和 L 分别与 R' 和 P' 逐元素相乘以获得加权后的特征图。最后将加权后的特征图按逐元素相加的方式得到最终的

融合结果 $PR \in F^{H\times W\times C}$。

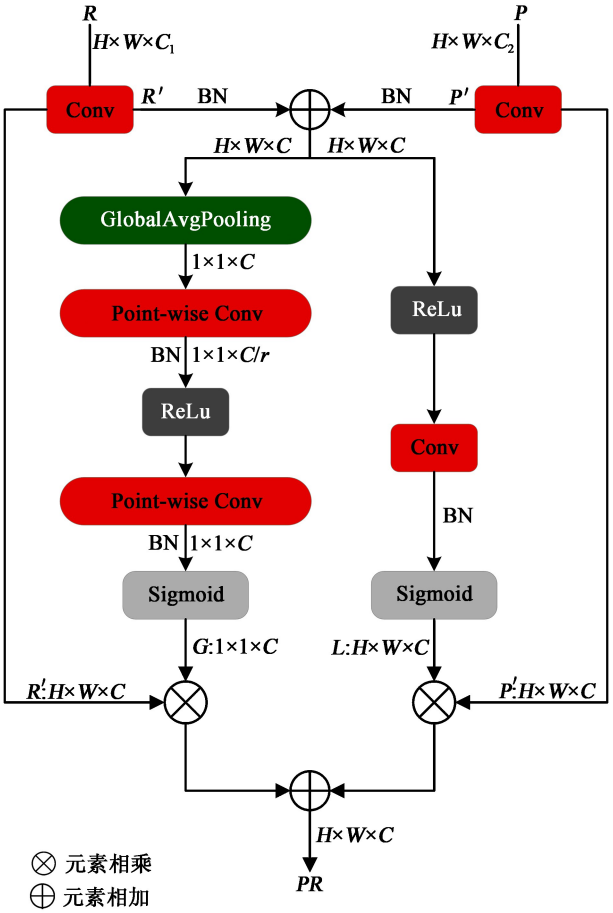


图 2　Fusion 特征融合模块结构图

2.4　交互式解码器

交互式解码器的结构如图 3 所示。首先，对从 Fusion 特征融合模块中得到的多尺度特征 $PR \in \{F^i \in \mathbb{R}^{C_i\times H_i\times W_i} \mid i=1,2,3,4\}$ 进行相应的调整，以便更有效地进行后续的解码操作。即通过 1×1 卷积操作对每个特征进行维度变换，使它们具有相同的维度 C。卷积操作后得到了新的多尺度特征 $T=\{T^i \in \mathbb{R}^{C\times H_i\times W_i} \mid i=1,2,3,4\}$。交互式解码器中采用了一种动态的特征融合方式，即动态加权交叉融合方式。这种融合方式考虑到了不同尺度特征的重要性，使得解码器能够高效地利用这些特征。动态加权交叉融合过程涉及以下几个步骤：首先，对于相邻的特征，进行上采样或下采样操作，使它们具有相同的分辨率。这是为了确保不同尺度的特征能够对齐，方便后续的融合操作。接下来，为每个特征分配一个可学习的权重因子，用于控制不同特征之间的融合强度。这些权重因子在模型训练过程中会自适应地更新，以最大限度地利用不同特征之间的互补信息，并突出每个特征的重要性。最后，通过加权求和的方式将特征进行融合，得到融合后的特征表示。通过以上步骤，加权交叉融合过程能够根据每个特征的重要性和互补性，自适应地融合不同尺度的特征。加权交叉融合的公式如下所示：

$$D=\{D^i=\sum_{j=1}^{N} w_{i,j}T^i \mid i=1,2,3,4;j=1,2,3,4\} \tag{1}$$

式中，N 的取值与 i 有关，当 $i=1,4$ 时，N 的取值为 2；而当 $i=2,3$ 时，N 的值为 3。$w_{i,j}$ 为赋给每个特征的权重值，i 表示第 i 个特征，j 表示第 j 个权重值。

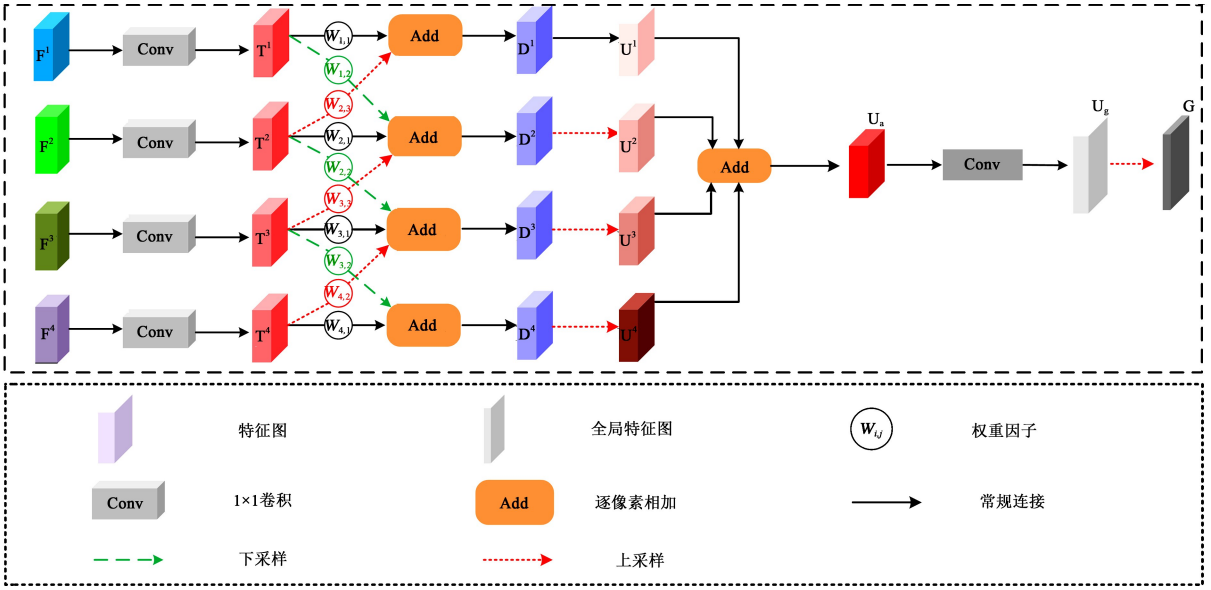


图 3　交互式解码器结构图

通过加权交叉融合过程后得到了充分交互后的多尺度特征 $D=\{D^i \in \mathbb{R}^{32\times H_i\times W_i} \mid i=1,2,3,4\}$，其具有相同的维度，即具有相同的通道数。为了获得最后的分割结果，首先，通过上采样得到了分辨率一致的多尺度特征 $U=\{U^i \in \mathbb{R}^{C\times H\times W} \mid i=1,2,3,4\}$。然后对多尺度特征 U 使用逐元素相加的方式获得了叠加后的特征 U_a。紧接着使用 1×1 卷积对特征 U_a 降维，得到了全局特征 U_g。最后，将全局特征 U_g 上采样以匹配输入图像的分辨率，从而得到最终预测结果 G。具体的细节如下公式所示：

$$G=\mathrm{Up}(\mathrm{Conv}(\mathrm{Add}(\mathrm{Up}(D^1),\mathrm{Up}(D^2),\mathrm{Up}(D^3),\mathrm{Up}(D^4)))) \in \mathbb{R}^{1\times H\times W} \quad (2)$$

式中，Up(·)表示上采样操作，Add(·)表示逐元素相加，Conv(·)表示 1×1 卷积操作。

2.5 损失函数

损失函数在深度学习模型中是用来衡量模型预测值与实际标签之间的差异，并通过反向传播算法来更新模型的参数，从而使得模型逐渐收敛于最优状态。在本文的方法中，使用了多任务损失函数，包括像素级别的二元交叉熵损失(The Binary Cross-Entropy Loss，BCE)和全局特征的对比损失。

(1)像素级别的二元交叉熵损失

该损失用于衡量模型在每个像素点上的预测与真实标签之间的差异。它能够有效地指导模型学习正确的像素分类。其定义如下：

$$L_{\mathrm{BCE}}=-y\log\hat{y}-(1-y)\log(1-\hat{y}) \quad (3)$$

式中，y 表示真实的标签值，$\hat{y}$ 表示预测值。

(2)全局特征的对比损失

为了保证模型能够建立全局联系，设计了全局特征的对比损失，用于衡量模型在提取全局特征方面的性能。这个损失可以促使模型更好地理解整体图像的语义信息。全局对比损失由两个损失函数组合而来，即 Dice Loss 和 Focal loss。Dice Loss 通常用于分割任务，特别是在需要准确地定位边界的情况下，它将模型的预测结果与真实标签的边界进行比较，促使模型产生更加准确的分割结果。而 Focal loss 则能够用于缓解前景像素和背景像素不平衡的问题，使模型在训练的过程中能够更多地去关注难分类的样本，从而提高道路裂缝检测的准确性。其定义分别如下：

$$L_{\mathrm{Dice}}=1-\frac{2\sum_{i=1}^{N}y_i\hat{y}_i}{\sum_{i=1}^{N}y_i+\sum_{i=1}^{N}\hat{y}_i} \quad (4)$$

式中，N 表示像素点总个数，y_i 表示真实的标签值，$\hat{y}_i$ 表示预测值。

$$L_{\mathrm{Focal}}=-\alpha(1-\hat{y})^{\gamma}y\log\hat{y}-(1-\alpha)\hat{y}^{\gamma}\cdot(1-y)\log(1-\hat{y}) \quad (5)$$

式中，α 表示平衡因子，γ 表示调制因子，y 表示真实的标签值，$\hat{y}$ 表示预测值。

(3)最终的损失函数

本文模型训练过程中的损失函数，其定义如下：

$$L_{\mathrm{loss}}=L_{\mathrm{BCE}}+L_{\mathrm{Dice}}+L_{\mathrm{Focal}} \quad (6)$$

3 实验与结果分析

3.1 实验平台

实验平台软硬件配置如表 1 所示。

表 1 实验软硬件配置

名称	实验配置
编程语言	Python 3.8
深度学习框架	PyTorch 1.10.0
CPU	12th Gen Intel(R) Core (TM) i5-12400
GPU	NVIDIA RTX 1080Ti
Cuda	Cuda11.2
平台	PyCharm 2022.2.23

3.2 网络训练

本文在 PyTorch 框架上实现了一种采用编码器-解码器结构的方法。在每个卷积层后使用批量归一化来加快训练期间的收敛速度，并通过 PReLU 添加网络层间的非线性关系，以克服过拟合和梯度消失问题。选用了 AdamW 优化器来进行模型训练，参数设置如下：输入图像的分辨率大小为 352×352，初始学习率为 0.000 1，每 25 个 epoch 学习率降低 10 倍，权重衰减率设定为 0.000 2。模型训练共进行 50 个 epoch。

3.3 数据集

(1)自建的数据集：道路裂缝采集车共获取了 1 606 张分辨率为 1 923×1 000 像素的道路裂缝图像，利用标注软件 labelme 对这些图像进行了标签标注。随后，将它们裁剪至 600×600 像素的分辨率，并仅保留了包含大于 1 000 像素的裂缝图像以及其对应的标签。最终，将获得的 1 757 张道路裂缝图像按照一定比例分为训练数据集和测试数据集两部分，其中训练数据集包含 1 500 张图像，测试数据集包含 257 张图像。

(2)DeepCrack 数据集：Liu 等[17]建立了公开道路裂缝数据集 DeepCrack，它包含 537 张分辨率为 554×384 像素的道路裂缝图像。此裂缝数据集包

含具有多种纹理以及场景的裂缝图像，这使得裂缝分割任务更具有挑战性。同样将数据集按照一定的比例分为了训练集和测试集，其中训练集包含 380 张图像，测试数据集包含 157 张图像。

此外，为了避免过拟合，采用了数据增强的方式对训练数据集进行扩充，包括随机翻转、裁剪和旋转等操作。而为了方便模型训练，在图像输入模型训练前将图像的分辨率统一调整为 352×352 像素。

3.4　评价指标

本文从以下两个角度来综合评价模型的性能：

(1)语义分割评价指标

本文选用的语义分割评价指标主要包括准确率(Precision)、召回率(Recall)、平均交并比(mIoU)、F_1 分数以及像素准确率(PA)等，它们的具体定义如下：

$$Precision=\frac{TP}{TP+FP} \tag{7}$$

$$Recall=\frac{TP}{TP+FN} \tag{8}$$

$$mIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP+FP+FN} \tag{9}$$

$$F_1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{10}$$

$$PA=\frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

式中，TP 表示正确分类为裂缝的像素数量，FP 表示被错误的分类为裂缝的背景像素数量，TN 表示被正确分类为背景的像素的数量，FN 表示被错误的分类为背景的裂缝像素数量，k 表示类别数量，在道路裂缝检测任务中，类别的数量为 1，因此 k 的值为 1。

(2)浮点运算量(FLOPs)和参数量(Params)

模型的复杂度也是衡量一个模型性能的重要评价指标。模型的复杂度可以通过其浮点运算量(FLOPs)和参数量(Params)来衡量。它们的具体定义如下：

$$FLOPs=2HW(C_{in}K^2+1)C_{out} \tag{12}$$

$$Params=(C_{in}K^2+1)C_{out} \tag{13}$$

式中，H 和 W 分别表示特征图的高度和宽度，C_{in} 和 C_{out} 分别表示特征的输入通道数和输出通道数，K 表示卷积核的大小。

3.5　不同数据集上模型对比实验结果与分析

3.5.1　自建的数据集

为了验证提出的道路裂缝图像分割方法的有效性，将其与目前比较先进的方法进行了比较，包括 FCN[18]，Unet[19]，SegNet[20]，DeepLab v3 +[21]，DeepCrack[22]，CrackFormer[23] 和 Swin-Unet[24]。表 2 中呈现了在自建的数据集上各个方法的性能比较结果。

表 2　本文方法和其他方法在自建的数据集上的性能比较

方法	Precision/%	Recall/%	mIoU/%	F_1/%	PA/%
FCN	68.58	71.04	53.60	69.79	90.19
Unet	65.96	68.44	50.58	67.18	88.22
SegNet	69.34	60.34	47.63	64.53	87.38
Deeplab v3+	74.72	45.90	39.73	56.87	83.12
DeepCrack	62.25	77.98	52.94	69.23	89.78
CrackFormer	64.77	78.69	55.10	71.05	90.46
Swin-UNet	61.03	68.07	47.44	64.35	87.10
本文方法	73.03	74.75	57.50	73.88	92.49

从实验结果中可以看出，和其他方法相比，本文的方法在 mIoU，F_1，PA 指标上效果好。在各个指标上分别达到了 57.50%，73.88%和 92.45%，比表现次优的方法 CrackFormer 分别高出了 2.40%，2.83%和 2.03%，而与这些方法中表现最差的 DeepLab v3+相比，分别高出了 17.77%，17.01%和 9.37%。此外，为了能够更加直观地去比较这些方法，图 4 为各个方法道路裂缝图像检测结果的可视化对比图。可以看出，本文的方法能够更精确地分割出道路裂缝图像的轮廓，同时能够更好地保留裂缝的细节信息，并有效区分道路标志线与道路裂缝。相比之下，其他常见的对比方法，如 DeepCrack，FCN 和 SegNet 等，出现将标志线错误地分割为道路裂缝的问题，即误检的情况。

3.5.2　DeepCrack 数据集

DeepCrack 数据集是一个广泛使用的公开道路裂缝数据集，由于其包含了丰富的道路裂缝图像样本，成为了道路裂缝检测任务中的重要基准。本文选用了 DeepCrack 数据集作为对比实验的数据集，以验证所提出方法的有效性以及性能表现。同样地，选用了 FCN，Unet，SegNet，DeepLab v3 +，DeepCrack，CrackFormer 和 Swin-Unet 与本文的方法进行了对比实验，得到的各个性能指标如表 3 所示。

从表 3 中可以看出，本文的方法在 mIoU，F_1，PA 指标上效果仍然是最好的，分别为 75.62%，86.12%，98.57%。同样地，为了能够更加直观地去比较这些方法，图 5 为各个方法的道路裂缝检测结果可视化对比图。可以看到，本文的方法对于裂缝的分割效果更优。例如，对于一些图像中裂缝像素

与背景的区别不明显的情况，基于卷积神经网络的方法如 DeepLab v3+，Unet 以及 SegNet 等方法会在裂缝的分割上遇到困难，难以准确分割得到裂缝，出现了漏检的情况。此外，基于视觉 Transformer 的方法 Swin-Unet 也同样存在无法准确分割出道路裂缝的情况。与之不同，本文提出的结合视觉 Transformer 和 CNN 的方法通过充分利用视觉 Transformer 的全局关联性和 CNN 的局部特征提取能力，能够更精准地分割出道路裂缝。

表 3 本文方法和其他方法在公开数据集 DeepCrack 上的性能比较

方法	Precision/%	Recall/%	mIoU/%	F_1/%	PA/%
FCN	85.06	80.92	70.85	82.94	94.03
Unet	81.55	85.07	71.34	83.27	95.17
SegNet	88.77	81.02	73.49	84.72	96.44
Deeplab v3+	83.93	77.82	67.72	80.76	92.03
DeepCrack	87.67	81.79	73.35	84.63	96.32
CrackFormer	84.88	86.39	74.87	85.63	97.45
Swin-UNet	83.69	82.40	71.00	83.04	94.20
本文方法	89.04	83.38	75.62	86.12	98.57

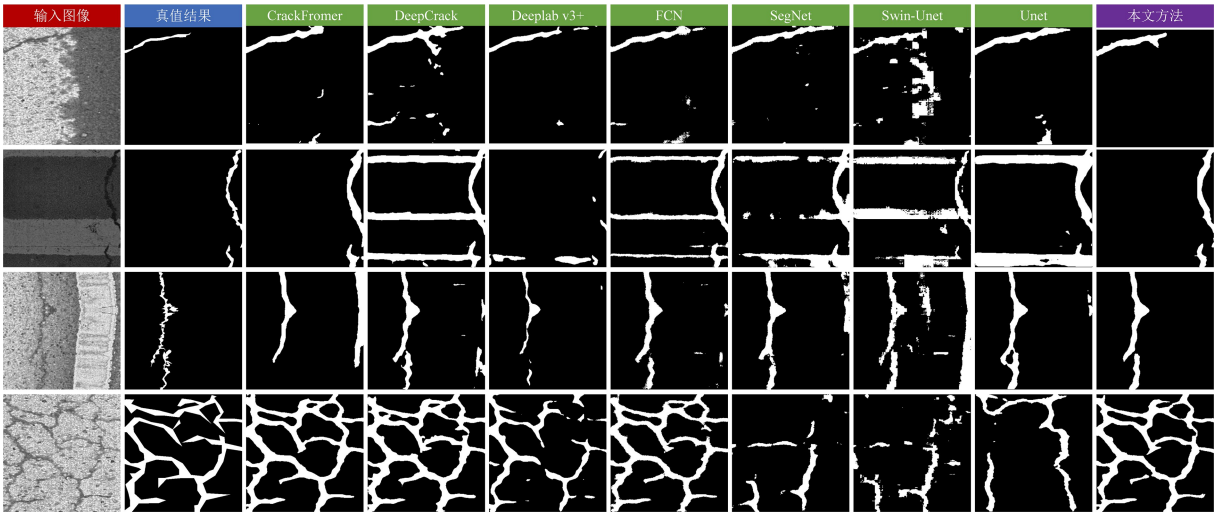


图 4 自建的数据集中道路裂缝检测结果的可视化对比图

3.6 模型复杂度比较

当输入的道路裂缝图像的分辨率为 352×352 像素时，计算得到模型的参数量(Params)和浮点运算量(FLOPs)如表 4 所示。

表 4 不同方法的浮点运算量和参数量比较

方法	FLOPs/G	Params/M
FCN	48.29	18.64
Unet	75.79	17.26
SegNet	75.87	29.44
Deeplab v3+	64.29	54.51
DeepCrack	258.87	30.91
CrackFormer	39.27	4.96
Swin-UNet	14.19	27.20
本文方法	20.41	38.94

根据表 4 可知，CrackFormer 拥有最小的参数量，仅有 4.96 M，这使得它相对容易训练，但其性能不如本文提出的方法。另一方面，Swin-Unet 网络在浮点运算量方面表现最优，然而在分割道路裂缝方面的效果并不理想。具体来看，从图 4 和图 5 中可见其在道路裂缝检测中存在误检和漏检的情况。相比之下，虽然本文的方法在浮点运算量和参数量两个指标上没有达到最优，但仍然是可接受的。因为本文采用了两个模型相结合的策略，充分利用了它们的优势，并取得了比其他对比方法性能更好的效果。

3.7 消融实验

为了探讨不同因素对模型性能的影响，进行了一系列的消融实验。通过逐步调整特定因素并保持其他条件不变以清晰地判断每个因素对模型性能的影响。共包括五组实验，分别验证了编码器分支和 Fusion 特征融合模块的性能。从表 5 的结果可知，单独使用 PVT 编码器或 ResNet50 编码器的效果均不如将两者结合使用时的性能好，这表明了双编码分支结构的有效性。其次，改进的 ResNet50 相较于 ResNet50 在道路裂缝检测任务中表现更好，在 mIoU，F_1 以及 PA 指标上，改进的 ResNet50 比 ResNet50 分别提高了 7.41%，6.92% 和 8.21%。因为本文的方法在 ResNet50 编码分支中加入轴向注意力机制去捕捉道路裂缝的方向性信息，这样的作法能够提升模型的性能。此外，从表 5 中的最后两行结果可以看出，引入 Fusion 融合模块去有效地融合两个编码分支提取的特征可以提升模型的性

能。在 Fusion 融合模块中，模型能够充分利用不同编码分支提取的特征，并有效地融合全局特征和局部特征，从而提升了道路裂缝的检测性能。因此，通过消融实验可知，本文模型采用的双编码分支以及设计的 Fusion 特征融合模块能够提升道路裂缝检测性能。

表 5　在自建的数据集上的消融实验的性能比较

方法	Precision/%	Recall/%	mIoU/%	F_1/%	PA/%
PVT	59.08	68.13	46.25	63.29	79.11
ResNet50	55.12	64.16	42.14	59.30	77.07
PVT＋ResNet50	66.18	63.91	48.18	65.03	81.34
改进的 ResNet50	63.40	69.40	49.55	66.26	85.28
PVT＋改进的 ResNet50	70.26	72.03	54.17	71.13	89.65
PVT＋改进的 ResNet50＋Fusion	73.03	74.75	57.50	73.88	92.49

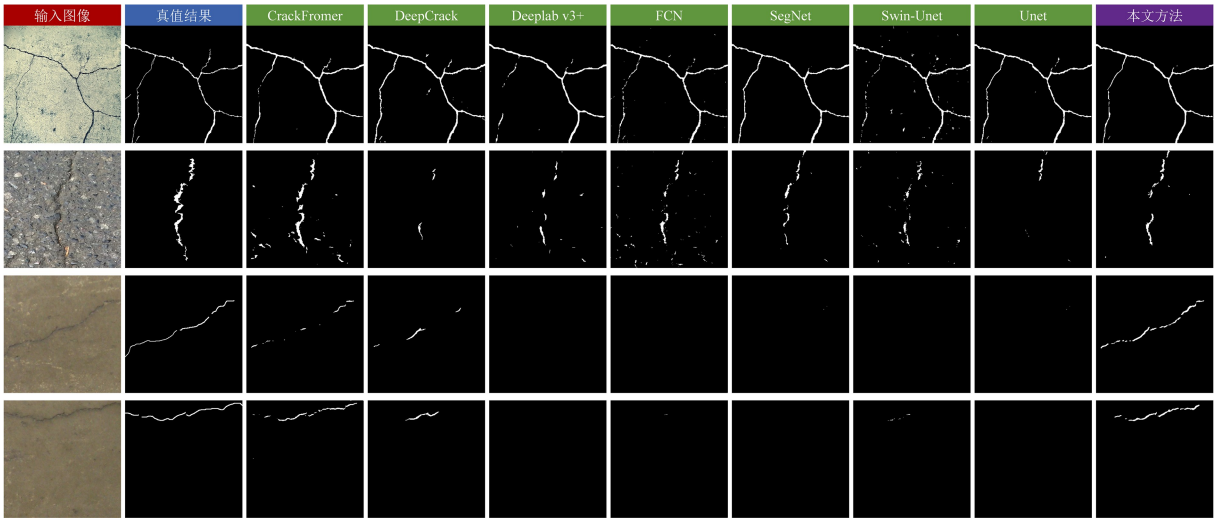


图 5　DeepCrack 数据集中道路裂缝检测结果的可视化对比图

3.8　可视化结果

在实际应用中，道路裂缝检测结果可视化对于工程实践至关重要。通过直观展示裂缝的分割效果，可以为道路维护和修复提供清晰的参考。此外，通过可视化结果，还能够及时发现可能存在的分割误差或不足之处，为模型的进一步优化提供指导。因此，在模型性能评估的过程中，利用可视化结果进行分析的方法也是十分必要的。

4　总结

本文提出了一种结合视觉 Transformer 和 CNN 的道路裂缝检测方法，以解决当前道路裂缝检测中长距离依赖建模以及细节信息捕获不全的问题。采用双编码分支有效地提取了道路裂缝图像中的关键特征。PVT 编码分支通过纯自注意力机制去捕获道路裂缝图像像素间的全局关联，为获取全局特征提供了有力支持。改进的 ResNet50 编码分支则考虑到了道路裂缝图像的方向性特点，从而能够捕获更多与裂缝相关的细节信息。然后，通过设计的 Fusion 特征融合模块将两个编码分支提取的特征进行有效地融合。最后，通过交互式编码器以高效的方式去融合多尺度特征，确保获得最佳的道路裂缝图像分割结果。实验结果表明，本文的方法在公开数据集 DeepCrack 以及自建的道路裂缝数据集上均实现了性能提升。相对于单独使用基于卷积神经网络和视觉 Transformer 网络的方法，提出的方法能够更精确地捕捉道路裂缝的细节特征。同时，此方法在处理长距离依赖性方面也具有明显的优势，相比单一的卷积神经网络模型，具有更强的全局信息感知能力。这对于准确识别道路裂缝的位置和形状，为道路养护工作提供了强有力的支持。综上所述，本文的研究为道路裂缝检测提供了一种可靠高效的解决方案。

参考文献：

[1] 张　恒，蒙　华，刘唐志. 沥青路面裂缝图像边缘检测算法研究[J]. 北方交通，2015(3)：56-59.
Zhang Heng, Meng Hua, Liu Tangzhi. Analysis on the image edge detection algorithm of asphalt pavement crack [J]. Northern Communications, 2015(3): 56-59.

[2] 瞿　中，林丽丹，郭　阳. 形态学与区域延伸相结合的图像裂缝检测算法研究[J]. 计算机科学，2014，41(11)：297-300.
Qu Zhong, Lin Lidan, Guo Yang. Algorithm of image crack detection based on morphology and region extends [J]. Computer Science, 2014, 41(11): 297-300.

[3] Wang S, Tang W. Pavement crack segmentation algorithm

based on local optimal threshold of cracks density distribution [C]// Advanced Intelligent Computing: 7th International Conference, ICIC 2011, 2011: 298-302.

[4] 杨济瑞，张晓燕，杜小甫. 一种基于深度学习的图像裂缝分类方法[J]. 电子测试，2021(4)：44-45，104.
Yang Jirui, Zhang Xiaoyan, Du Xiaofu. A crack detection and classification method based on deep learning[J]. Electronic Test, 2021(4): 44-45, 104.

[5] Feng H, Xu G, Guo Y. Multi-scale classification network for road crack detection[J]. IET Intel. Transport Syst., 2019, 13(2): 398-405.

[6] 任安虎，姜子渊，马晨浩. 基于改进 YOLOv5s 的道路裂缝检测算法[J/OL]. 激光杂志：1-7[2023-11-10]. http://kns.cnki.net/kcms/detail/50.1085.TN.20230626.1756.012.html.
Ren Anhu, Jiang Ziyuan, Ma Chenhao. Road crack detection algorithm based on improved YOLOv5s[J/OL]. Laser J.: 1-7 [2023-11-10]. http://kns.cnki.net/kcms/detail/50.1085.TN.20230626.1756.012.html.

[7] Zhang L, Yang F, Zhang Daniel Y, et al. Road crack detection using deep convolutional neural network[C]// 2016 IEEE Inter. Conf. on Image Processing (ICIP), 2016: 3708-3712.

[8] Xu Z, Sun Z, Huyan J, et al. Pixel-level pavement crack detection using enhanced high-resolution semantic network [J]. Int. J. Pavement Eng., 2022, 23(14): 4943-4957.

[9] Kheradmandi N, Mehranfar V. A critical review and comparative study on image segmentation-based techniques for pavement crack detection[J]. Constr. Build. Mater., 2022, 321: 126162.

[10] Tang Y, Zhang A A, Luo L, et al. Pixel-level pavement crack segmentation with encoder-decoder network [J]. Measurement, 2021, 184: 109914.

[11] Wang W, Xie E, Li X, et al. PVT v2: Improved baselines with pyramid vision transformer[J]. Comput. Visual Media, 2022(3): 415-424.

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]// Proc. of 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.

[13] 许正森，雷相达，管海燕. 多尺度局部特征增强 Transformer 道路裂缝检测模型[J]. 中国图象图形学报，2023，28(4)：1019-1028.
Xu Zhengsen, Lei Xiangda, Guan Haiyan. Multi-scale local feature enhanced transformer network for pavement crack detection[J]. J. of Image and Graphics, 2023, 28(4): 1019-1028.

[14] Liu H J, Miao X Y, Mertz C, et al. Crackformer: Transformer network for fine-grained crack detection[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 3763-3772.

[15] Ju X, Zhao X, Qian S. TransMF: Transformer-based multi-scale fusion model for crack detection[J]. Mathematics, 2022, 10(13): 2354.

[16] Ho J, Kalchbrenner N, Weissenborn D, et al. Axial attention in multidimensional transformers[J/OL]. ArXiv, 2019, abs/1912: 12180. https://api.semanticscholar.org./Corpus ID: 209323787.

[17] Liu Y, Yao J, Lu X, et al. DeepCrack: A deep hierarchical feature learning architecture for crack segmentation [J]. Neurocomputing, 2019, 338: 139-153.

[18] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 39(4): 640-651.

[19] Olaf R, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]// Bildverarbeitung für die Medizin, 2017, 2017: 1-3.

[20] Vijay B, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 39(12): 2481-2495.

[21] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]// Proc. of the European Conf. on Computer Vision (ECCV), 2018: 833-851.

[22] Zou Q, Zhang Z, Li Q, et al. DeepCrack: Learning hierarchical convolutional features for crack detection[J]. IEEE Trans. Image Process., 2019, 28(3): 1498-1512.

[23] Liu H, Yang J, Miao X, et al. CrackFormer network for pavement crack segmentation [J]. IEEE Trans. Intell. Transp. Syst., 2023, 24(9): 9240-9252.

[24] Cao H, Wang Y, Chen J, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation[C]// European Conf. on Computer Vision, 2022: 205-218.

作者简介：

代少升(1974－)，男，河南省潢川县人，博士，教授，主要从事红外成像系统及 SOPC 嵌入式系统的设计与开发；

余自安(1974－)，男，河南省光山县人，高级工程师，主要从事机械设计及理论研究。