arXiv:2601.00922v1 [eess.IV] 1 Jan 2026

MetaFormer驱动的鲁棒医学语义分割编码网络

Le-AnhTran、Chung NguyenTran、Nhan Cach Dang、Anh Le Van Quoc、Jordi Carrabina、David Castells-Rufas 和 Minh Son Nguyen

摘要。 语义分割对于医学图像分析至关重要，有助于精准诊断疾病并制定治疗规划。然而，许多先进模型采用复杂架构，限制了其在资源受限的临床环境中的应用。本文提出 MFEnNet，这是一种高效的医学图像分割框架，将 MetaFormer 融入 U-Net 主干网络的编码阶段。MetaFormer 是视觉变换器的一种架构抽象，通过将图像块转换为序列来进行全局上下文建模，为卷积神经网络提供了一种灵活的替代方案。为降低自注意力带来的高昂计算成本，该框架以池化变换器块取代传统变换器模块，从而以较低的复杂度有效聚合全局特征。此外，框架采用Swish激活函数，以实现更平滑的梯度和更快的收敛；同时在瓶颈层引入空间金字塔池化，以改进多尺度特征提取。在不同医学分割基准上的全面实验表明，与最先进模型相比，所提出的 MFEnNet 方法在显著降低计算成本的同时，实现了具有竞争力的准确率。本研究的源代码可在 https://github.com/tranleanh/mfennet 获取。

关键词： 医学图像分割、语义分割、U-Net、视觉Transformer、MetaFormer

1 引言

医学图像语义分割是计算机辅助诊断、治疗规划和手术导航中的基础任务。对解剖区域和病理区域进行精确 delineation，有助于支持可靠的临床决策以及疾病分类等任务。鉴于其重要性，大量研究致力于开发能够在精度与计算效率之间取得平衡的分割模型。早期的卷积神经网络（卷积神经网络），如 U-Net [1]及其变体，凭借捕获局部模式并适应各种成像模态的能力，已广泛应用于医学成像。然而，卷积神经网络难以处理长程依赖，限制了其准确分割复杂或弥漫性结构的能力。

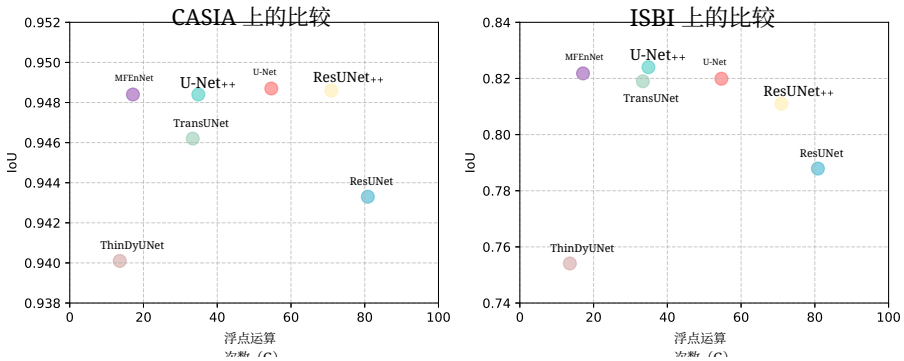


图 1：准确率（交并比）与复杂度（浮点运算次数）之间的权衡。

视觉变换器（视觉Transformer）的出现通过将图像处理为图像块序列，并利用自注意力捕获全局上下文，彻底改变了图像分析，在计算机视觉领域取得了顶尖性能。然而，其较高的计算和内存需求阻碍了其在资源受限的医学成像环境中的应用，例如便携式设备和边缘系统。为解决这一问题，研究人员开发了轻量级变换器变体和混合架构，以平衡准确性与效率。值得注意的是，MetaFormer [3]已成为一种颇具前景的架构抽象，表明变换器的核心设计可以与自注意力解耦，并通过替代性令牌混合器进行泛化，这意味着高效的令牌聚合机制可以作为医学图像分割中自注意力的可行替代方案。

本文提出了 MFEnNet，这是一种由 MetaFormer 驱动的用于医学图像分割的编码网络，其编码骨干网络采用池化令牌混合器。通过以基于池化的操作替代传统的自注意力，MFEnNet 以显著降低的计算成本实现了全局特征聚合。本文采用 Swish激活函数，以确保反向传播过程中的梯度更加平滑，从而缓解梯度消失问题，实现更稳定的训练。在瓶颈层处策略性地引入空间金字塔池化（SPP） [4] ，以改进多尺度特征提取，有效捕获不同空间分辨率下的局部细节和全局上下文。所提出的 MFEnNet 在多个医学分割基准上进行了评估，包括 CASIA [5] 和 ISBI [6], ，以考察其分割准确率和计算效率。图1展示了这两个方面之间的权衡，结果表明，MFEnNet 在计算成本显著更低的情况下，仍能达到与当前最先进模型相当的准确率。总而言之，本文的主要贡献总结如下：1）将 MetaFormer 应用于医学图像分割，证明了其作为传统卷积神经网络和基于自注意力的变换器骨干网络替代方案的有效性；2）在变换器块中以池化操作替代自注意力，在保持具有竞争力的分割准确率的同时，实现高效的特征聚合；

3）在多个医学基准数据集上对所提出的 MFEnNet 进行了评估，以显著更低的计算成本实现了当前最先进的准确率，从而提升了其对资源受限应用的适用性。

2 相关工作

医学图像分割的早期进展由 U-Net [1],推动，该模型引入了对称的编码器—解码器架构，并成为众多视觉任务的事实标准基线 [7,8]。在 U-Net 的基础上，出现了许多变体，例如 ResUNet [9],、UNet++ [10] 和 ResUNet++ [11]，它们引入残差学习、重新设计跳跃连接，并加入注意力机制，以进一步优化特征表示并提高分割准确率。

受视觉Transformer能够捕获长程依赖能力的启发，许多研究将 CNN骨干网络 与 Transformer编码器相结合，使模型能够利用全局上下文，同时保留空间精度。例如，TransUNet[12]采用卷积神经网络提取低级特征，将其令牌化后输入 Transformer编码器进行处理，随后以 U-Net 风格进行解码。
Swin-Unet [13]以分层窗口式 Transformer 替代卷积编码器，对局部与全局表示进行联合建模。TransFuse [14]显式融合并行的 CNN与Transformer流，以利用互补信息。除架构设计之外，研究者还提出了多种方法，以应对医学数据集特有的统计特征和有限数据条件，例如门控/轴向注意力 [15] 以及局部—全局并行聚合 [16]。尽管视觉Transformer中的自注意力具有强大的表征能力，但其计算复杂度给高分辨率医学图像处理和资源受限部署带来了挑战。这推动了针对更高效令牌混合策略的研究。值得注意的是，MetaFormer [3]揭示了 Transformer 的大部分有效性源于其通用架构设计，而非注意力机制本身；研究表明，在不显著损失性能的情况下，可以使用更简单的替代方案替换令牌混合器。

受 MetaFormer 启发，所提出的框架将其原理融入医学图像分割中，采用池化Transformer块实现高效的全局上下文聚合；与传统的基于自注意力的模型相比，其计算开销显著降低。

3 方法

所提出的 MFEnNet 将受 MetaFormer 启发的模块集成到 U-Net 编码器中，以进行全局上下文建模；在瓶颈层引入 SPP模块，以实现多尺度特征聚合；并采用 Swish激活函数来改善梯度流，最终以较低的计算成本实现了优异的性能。

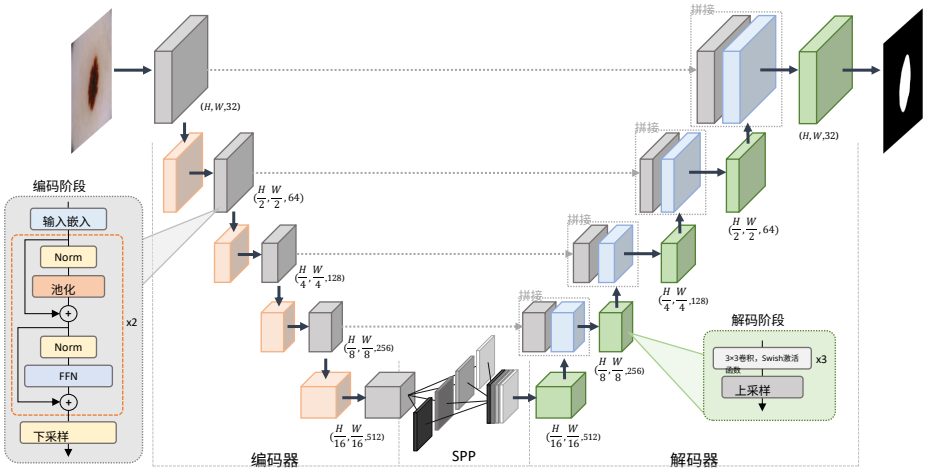


图2：所提出的网络。

3.1 MetaFormer驱动的编码器

在标准 ViT模块中，输入特征图 X 被划分为图像块，并嵌入为令牌序列：

$$Z = \text{InputEmb}(X), \tag{1}$$

其中， $Z \in \mathbb{R}^{N \times C}$ 表示具有 C 个通道的 N 令牌。随后，序列 Z 经过两个子模块处理。第一个子模块执行令牌混合：

$$Z_1 = Z + \text{TokenMixer}(\text{Norm}(Z)), \tag{2}$$

第二个子模块则应用前馈变换：

$$Z_2 = Z_1 + \text{FFN}(\text{Norm}(Z_1)). \tag{3}$$

其中，TokenMixer(.)实现令牌之间的交互（通常通过自注意力实现），Norm(.)表示归一化，用于稳定训练，而 FFN(.)是一个带有非线性激活的两层前馈网络。

在所提出的模型中，编码器遵循 MetaFormer 范式 [3]进行了重新构造。具体而言，令牌混合器被替换为池化算子，从而得到修改后的第一个子模块：

$$Z_1 = Z + \text{Pooling}(\text{Norm}(Z)), \tag{4}$$

而第二个子块保持不变。FFN(.)采用两个全连接层，扩展比为 $r = 4$ ，并使用 Swish激活函数 σ；每个子块均配备跳跃连接，以促进信息流动。对于输入嵌入，采用 3×3 卷积。所提出框架的编码阶段如图 2 所示。该设计保留了 ViT 架构的长程依赖建模特性，同时显著降低了内存使用量和计算需求。

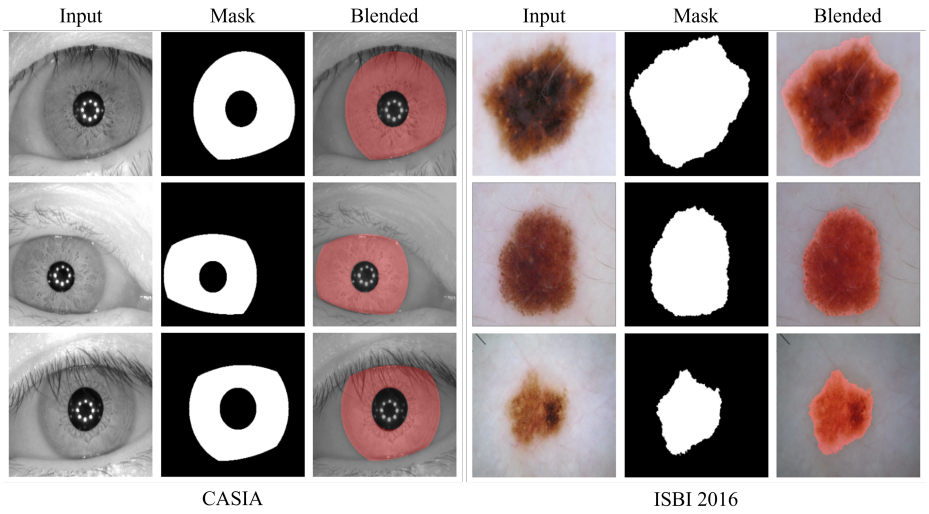


图 3：实验中使用的数据示例。

3.2 MFEnNet

所提出的 MFEnNet 构建于原始 U-Net 架构之上，由一个带有长跳跃连接的编码器–解码器结构组成，以跨尺度保留细粒度空间信息，如图 2 所示。在编码器中，每个层级都以输入嵌入层开始，随后连接一组编码块，并采用池化作为令牌混合策略。通过最大池化执行下采样，在压缩空间分辨率的同时逐步扩大感受野。解码器沿用 U-Net 的设计，通过上采样以及与对应层级编码器特征的拼接，逐步重建高分辨率特征图。然而，与标准 U-Net 不同，我们用 Swish 激活函数替换 ReLU；该函数能够提供更平滑的梯度，并已被证明可以提升优化稳定性和表征能力。受一些具有代表性的图像到图像转换工作 [17,18], 启发，我们在瓶颈层引入了 SPP模块。该模块通过对不同大小区域执行池化操作来聚合多尺度上下文信息，使网络能够同时捕获全局语义和精细的局部结构。该网络由五个阶段组成（包括输入/输出阶段以及四个下采样/上采样步骤），特征图维度分别为 $H \times W \times 32$、$\frac{H}{2} \times \frac{W}{2} \times 64$、$\frac{H}{4} \times \frac{W}{4} \times 128$、$\frac{H}{8} \times \frac{W}{8} \times 256$和 $\frac{H}{16} \times \frac{W}{16} \times 512$，其中 H 和 W 分别表示输入图像的高度和宽度。MetaFormer 块仅应用于编码器中，因为基于 ViT 的块特别擅长对输入序列中的长程依赖进行建模，这对于编码阶段的上下文理解至关重要 [12,19]。在输出层，普通的 1×1 卷积层生成最终的分割图。

表1：各种方法在准确率和复杂度方面的比较。

模型	CASIA		ISBI		复杂度	
	IoU	Dice	IoU	Dice	参数量（M）	浮点运算次数（G）
U-Net [1]	0.9487	0.9734	0.8199	0.8911	31.04	54.66
ResUNet [9]	0.9433	0.9705	0.7879	0.8675	13.04	80.83
U-Net++ [10]	0.9484	0.9730	0.8238	0.8947	9.16	34.87
TransUNet [12]	0.9462	0.9720	0.8195	0.8894	3.63	33.36
ResUNet++ [11]	0.9486	0.9733	0.8110	0.8854	14.48	70.92
ThinDyUNet [20]	0.9401	0.9688	0.7541	0.8428	0.81	13.56
MFEnNet（本文方法）	0.9484	0.9732	0.8218	0.8913	11.14	17.13

4 实验

4.1 实验设置

数据集：实验使用了两个公开数据集：CASIA虹膜间隔数据集（CASIA） [5]和ISBI 2016（ISBI） [6]。CASIA数据集包含使用同一传感器从249名受试者的395只眼睛采集的2,639张虹膜图像。该数据集被随机划分为训练集（80%）和验证集（20%）。另一方面，专为皮肤病变分割设计的 ISBI数据集包含900张训练图像和379张用于评估的测试图像。两个数据集的代表性样本如图3所示。

实验设置： 所有实验均在配备 NVIDIA Tesla T4 GPU 的 Linux 系统上进行。所提出的框架使用 PyTorch 库实现，并采用 Adam优化器训练50个周期，批量大小为16，学习率为 10^{-4}。损失函数采用二元交叉熵（BCE）。输入图像被调整为256×256 像素。为增强模型的鲁棒性，训练期间应用了包括随机翻转和随机裁剪在内的数据增强技术。定量性能使用交并比（IoU）和 Dice系数进行评估，计算效率则通过可训练参数数量（参数量，M）以及针对 256×256 输入的浮点运算次数（G）进行评估。

基线模型：我们将所提出的网络与当前先进的语义分割模型进行了比较。这些模型涵盖了广泛的方法，包括基于卷积神经网络和基于Transformer的架构，从而确保比较全面且公平。所选模型包括 U-Net [1], U-Net++ [10]、ResUNet [9], ResUNet++ [11]、TransUNet [12], 和 ThinDyUNet [20]。这些基线模型有助于全面比较各模型的性能和效率。

4.2 定量分析

表1报告了所提出的 MFEnNet 与具有代表性的卷积神经网络模型和基于Transformer的模型在 CASIA 和 ISBI 基准上的对比评估，评估指标为分割准确率和模型复杂度。

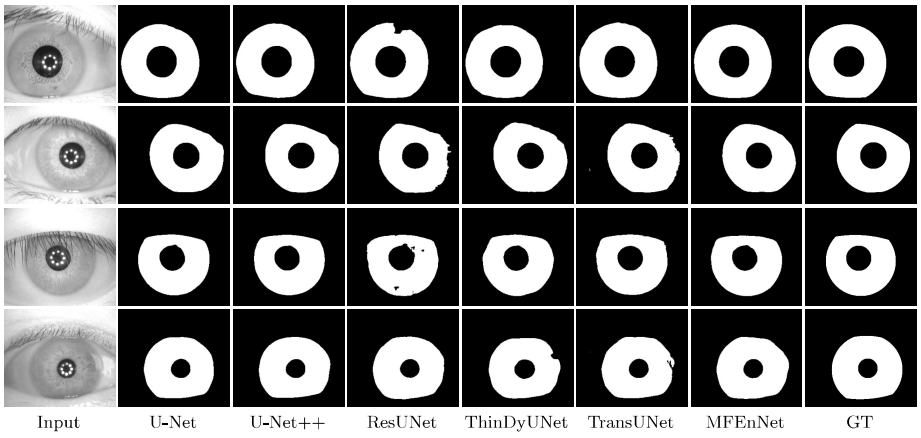


图 4：CASIA 数据上各种方法的典型视觉比较。

在这两个数据集上，U-Net 及其变体（如 ResUNet、U-Net++ 和 ResUNet++）均展现出稳定的优异分割性能，再次印证了卷积编码器—解码器架构的鲁棒性。其中，U-Net++ 在 ISBI数据集上的表现最佳，交并比为 0.8238，Dice系数为 0.8947，这凸显了密集跳跃连接在增强特征融合方面的有效性。然而，准确率的提升是以计算复杂度增加为代价的，尤其是与 ThinDyUNet 和 MFEnNet 等轻量级模型相比时。基于Transformer的方法 TransUNet 也取得了有竞争力的结果，且模型大小较小，展现了全局上下文建模的优势。尽管如此，这类模型通常仍需要显著更多的计算资源；例如，TransUNet 的参数量相对适中（3.63M），但需要 33.36 GFLOPs。类似地，ResUNet++ 的表现较好，但其参数量为 14.46M，浮点运算次数为 70.92 GFLOPs，这可能限制其在资源受限环境中的应用。相比之下，所提出的 MFEnNet 在准确率和效率之间实现了良好平衡。在 CASIA 上，其交并比为 0.9484，Dice系数为 0.9732，与表现最佳的基线模型十分接近。在 ISBI 上，其交并比为 0.8218，Dice系数为 0.8913，性能与 U-Net++相当，同时大幅减少了浮点运算次数。具体而言，MFEnNet 仅需 11.14M 参数和 17.13G 浮点运算次数；与 U-Net 相比，参数量减少了 64%，浮点运算次数减少了近 68%，同时保持了相当的分割准确率。

这些发现揭示了两个重要见解。首先，通过受 MetaFormer 启发的池化块高效实现的基于 Transformer 的令牌混合，能够保留全局上下文建模的表征优势，同时避免产生高昂的计算成本。其次，多尺度聚合（通过空间金字塔池化实现）与平滑优化动态（通过 Swish激活函数实现）的结合，进一步有助于稳定训练并取得具有竞争力的分割

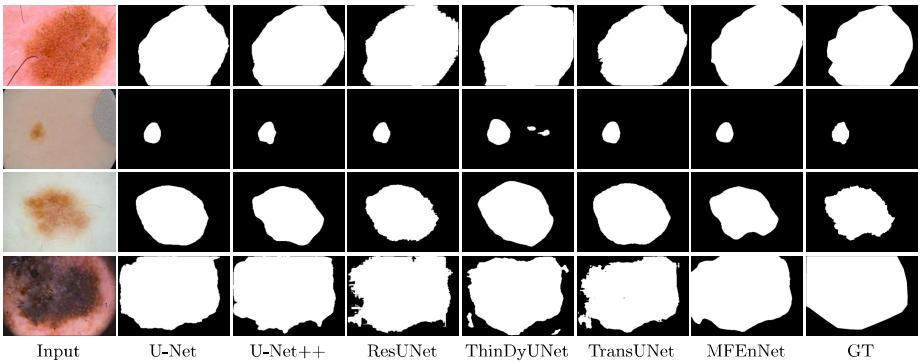


图5：各方法在 ISBI 数据上的典型视觉比较。

结果。总体而言，所提出的 MFEnNet 以显著更低的复杂度实现了最先进的分割精度，从而更适用于资源受限的医疗应用。

4.3 定性分析

图4和图5展示了最先进模型所生成分割结果的视觉比较，包括 U-Net、U-Net++、ThinDyUNet、ResUNet、TransUNet 以及本文提出的 MFEnNet。

从图4可以看出，对于 CASIA数据集中的虹膜图像，基于CNN的方法（如 U-Net 和 U-Net++）能够生成较为合理的结果，且边界清晰。ResUNet 往往会产生轻微的边界变形，并出现小块碎片化区域和噪声预测；而 ThinDyUNet 由于表征能力较弱，在多个案例中出现欠分割。TransUNet 利用自注意力改善了边界平滑度，但在变化的光照条件下仍会引入错误和局部伪影。相比之下，MFEnNet 生成的掩膜在视觉上最接近真实标注，虹膜和瞳孔区域周围的轮廓干净且连续。

另一方面，图5展示了ISBI数据集中的皮肤病变图像结果。由于存在不规则病变形状、模糊边界和不同病变大小，这些图像带来了额外挑战。U-Net和 U-Net++ 往往会产生过分割，导致掩膜超出真实病变区域。ResUNet引入了结构不一致，而ThinDyUNet遗漏了更细微的病变细节，生成不完整掩膜。TransUNet在捕捉病变范围方面表现更好，但在复杂情况下会生成不均匀轮廓，尤其是在背景纹理类似于病变模式时。相比之下，MFEnNet始终能够以很高的保真度勾画病变边界，在捕捉不规则边缘的同时避免过分割。值得注意的是，在小病变情况下，所提出的MFEnNet仍能实现精确定位，而其他模型则会出现欠分割或引入假阳性。

5 结论

在本研究中，我们提出了MFEnNet，这是一种高效的医学图像分割框架，利用受MetaFormer启发的编码器，在分割精度与计算效率之间取得平衡。通过使用池化Transformer模块替代传统的自注意力机制，MFEnNet能够在保持较低复杂度的同时，有效聚合全局上下文信息。在瓶颈层集成空间金字塔池化（SPP）模块，进一步增强了多尺度特征表示；同时，采用Swish激活函数有助于实现稳定优化并改善梯度流。在包括CASIA和ISBI在内的基准数据集上的评估表明，与最先进方法相比，MFEnNet实现了具有竞争力的分割精度，同时显著降低了计算成本。定性比较进一步证实，MFEnNet生成的分割掩膜具有更清晰的边界和更忠实的结构表示，凸显了其在虹膜边界提取、皮肤病变勾画等多种成像场景中的稳健性。凭借性能与效率之间的良好权衡，MFEnNet在资源受限的医疗应用中展现出良好的部署前景。

参考文献

1. Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-net：用于生物医学图像分割的卷积网络。在 *International Conference on Medical image computing and computer-assisted intervention* 234–241, 2015，第 – 页。2. Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xi-aohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像胜过 16x16 个词：用于大规模图像识别的变换器。在 *International Conference on Learning Representations* 2021，。3. Weihao Yu、Mi Luo、Pan Zhou、Chenyang Si、Yichen Zhou、Xinchao Wang、Jiashi Feng 和 Shuicheng Yan。MetaFormer实际上就是视觉任务所需的模型。在 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第 10819–10829 页，2022。4. Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。深度卷积网络中的空间金字塔池化，用于视觉识别。*IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 9 1904–1916, 2015，( )：–。5. 中国科学院自动化研究所。CASIA虹膜图像数据库。技术报告，中国科学院，2004。6. David Gutman、Noel CF Codella、Emre Celebi、Brian Helba、Michael Marchetti、Nabin Mishra 和 Allan Halpern。面向黑色素瘤检测的皮肤病变分析：国际生物医学成像研讨会 (ISBI) 上的一项挑战，由国际皮肤成像协作组织 (ISIC) 主办。2016, *arXiv preprint* ，。7. Le-Anh Tran 和 My-Ha Le。面向自动驾驶的稳健 U-Net 道路车道线检测。在 *2019 International Conference on System Science and Engineering (ICSSE)* 62–66 2019，第 – 页。IEEE，。

8. Le-Anh Tran、Seokyong Moon 和 Dong-Chul Park。采用引导传输图的全新编码器-解码器网络，用于单幅图像去雾。 *Procedia Computer Science* 204 682 689, 2022 ，：–。9. Zhengxin Zhang、Qingjie Liu 和 Yunhong Wang。利用深度残差 U-Net 进行道路提取。 *IEEE Geoscience and Remote Sensing Letters*，15(5)：749–753，2018。10. Zongwei Zhou、Md Mahfuzur Rahman Siddiquee、Nima Tajbakhsh 和 Jianming Liang。Unet++：用于医学图像分割的嵌套 U-Net 架构。在 *International workshop on deep learning in medical image analysis*，第 3–11 页。施普林格，2018。11. Debesh Jha、Pia H Smedsrud、Michael A Riegler、Dag Johansen、Thomas De Lange、P˚al Halvorsen 和 H˚avard D Johansen。Resunet++：一种先进的医学图像分割架构。在 *2019 IEEE international symposium on multimedia (ISM)* 225 2255 2019 ，第 – 页。IEEE，。12. Jieneng Chen、Jieru Mei、Xianhang Li、Yongyi Lu、Qihang Yu、Qingyue Wei、Xi-angde Luo、Yutong Xie、Ehsan Adeli、Yan Wang 等。Transunet：从变换器的视角重新思考医学图像分割中的 U-Net 架构设计。 *Medical Image Analysis*，97：103280，2024。13. Hu Cao、Yueyue Wang、Joy Chen、Dongsheng Jiang、Xiaopeng Zhang、Qi Tian 和 Manning Wang。Swin-unet：用于医学图像分割的类 U-Net 纯变换器。在 *European conference on computer vision*，第 205–218 页，2022。14. Yundong Zhang、Huiye Liu 和 Qiang Hu。Transfuse：融合变换器与卷积神经网络的医学图像分割方法。在 *International conference on medical image computing and computer-assisted intervention* 14 24 2021 ，第 – 页。施普林格，。15. Jeya Maria Jose Valanarasu、Poojan Oza、Ilker Hacihaliloglu 和 Vishal M Patel。医学变换器：用于医学图像分割的门控轴向注意力。在 *International conference on medical image computing and computer-assisted intervention* 36 46 2021 ，第 – 页。施普林格，。16. Wentao Liu、Tong Tian、Weijin Xu、Huihua Yang、Xipeng Pan、Songlin Yan 和 Lemeng Wang。Phtrans：并行聚合全局与局部表征的医学图像分割方法。在 *International Conference on Medical Image Computing and Computer-Assisted Intervention* 235 244 2022 ，第 – 页。施普林格，。17. Le-Anh Tran 和 Dong-Chul Park。采用引导传输图的编码器-解码器网络，实现有效图像去雾。 *The Visual Computer* 41 1 359 382, , 2025 ( )：–。18. Hengshuang Zhao、Jianping Shi、Xiaojuan Qi、Xiaogang Wang 和 Jiaya Jia。金字塔场景解析网络。在 *Proceedings of the IEEE conference on computer vision and pattern recognition* 2881 2890, 2017 ，第 – 页。19. Le-Anh Tran 和 Dong-Chul Park。用于高效真实图像去雾的蒸馏池化变换器编码器。 *Neural Computing and Applications*，37(6)：5203–5221，2025。20. Sang-Chul Kim 和 Yeong Min Jang。面向反无人机应用的语义分割数据集与实时定位模型。 *Applied Sciences* 15 13 7183, , ( )：。2025