

localswin : 基于 swin-transformer 的高效胃癌病理图像分割

毛松仁¹ 刘 杰^{2,3}

MAO Songren LIU Jie

摘 要

医学图像分割技术在疾病诊断中发挥着重要作用,但是 CNN 在处理图像时缺乏全局上下文信息,而 transformer 对于局部信息处理能力相对较弱,这都影响分割任务的准确性。针对这一问题,提出名为 localswin 的高性能分割网络,在 U-Net 架构的基础上,编码部分使用 swin transformer 模块,将反向残差块中所使用的深度卷积引入前馈网络中,增强了对局部特征的提取能力。利用所提出的策略来融合全局和局部特征,提高了模型的性能和效果。在公开的 BOT 胃切片数据集上评估了所提出的网络。实验结果表明,localswin 与其他深度分割模型相比,在分割任务上有更好的效果,对基于切片的分割的准确率达到 86.27%。

关键词

医学图像分割; swin transformer; localswin; U 型架构

doi: 10.3969/j.issn.1672-9528.2024.01.045

0 引言

医学图像分割作为医学图像处理中的重要步骤,其目的是将医学图像中器官和病灶等关键区域分割出来,为医生的诊断提供依据。部分病灶图像具有目标区域与背景区域的对比度低、病灶边界难以被区分的特点。胃癌^[1-2]是最常见的恶性肿瘤之一,2020 年居世界发病率第 5 位,死亡率第 4 位,2020 年有超过 100 万新发病例和大约 77 万死亡病例,胃癌是中国第二大高发癌症,2020 年中国约有 46.96 万宗胃癌病个案,约有 34.12 万宗死亡病例,占全球胃癌死亡病例总数的 44.4%。我国近半数患者确诊时已为晚期,即使接受根治术后仍有大约 50% 患者会复发转移,因此具有分期晚、肿瘤负荷大、异质性强及预后差的特点。病理检查是胃癌最常用的诊断方法之一,它提供了明确的疾病诊断来指导患者的治疗和管理决策^[3-4]。组织病理学图像是癌症诊断^[5]的金标准。

苏木精和伊红 (H&E) 染色的全载玻片图像 (WSIs) 是手术评估不可或缺的参考,准确的 WSI 分割可以提供肿瘤微环境^[6]的详细分析。但是,传统的病理学分析是人工通过显微镜进行检查,因此在对医学图像的诊断耗费了大量的精力的同时,也会因为长时间进行机械和重复性的工作让医生由

于过度疲劳而产生误判。所以迫切需要一种全自动、高效、可靠的 WSI 分割方法,这样可以大大减轻病理学家的工作量。

随着计算机技术和人工智能的飞速发展,出现了许多深度神经网络,如全卷积网 (fully convolutional networks, FCN)^[7]、U-Net^[8]、SegNet^[9] 和密集网络 (dense convolutional network, DenseNet)^[10] 等,这些技术的出现,极大影响着分割任务。在医学图像分割以及辅助疾病诊断方面有着广阔的应用前景。

虽然,卷积神经网络已经成为当今医学图像分割中的标准,然而由于卷积中归纳偏置的局部性和权重共享,这些网络使用的卷积操作不可避免地在建模远程依赖方面存在局限性。虽然 CNN 具有一些难以解决的问题,但是这些问题却能被 transformer 处理。transformer 是一种基于自注意力机制的序列建模方法,最初主要用于自然语言处理 (natural language processing, NLP) 领域^[11],但现在引起了计算机视觉研究人员的极大兴趣。Dosovitskiy 等人^[12]提出了纯自注意力视觉转换器 (vision transformer, ViT),并首次将其应用在计算机视觉领域,它将图像分类问题转化为序列建模问题,在大型外部数据集上进行预训练,这个模型在 ImageNet^[13]上具有较好的结果。Zheng 等人^[14]提出的分割 transformer (segmentation transformer, SETR),在传统的、基于编码器解码器的网络中用 transformer 替换编码器,从而成功地在自然图像分割任务上获得最先进的结果。随后,swin transformer^[15]通过引入基于滑动窗口的自注意力机制,结合了局部感受野,提高了计算效率和准确率。该模型在多种计算机视觉任务上取得了显著的性能提升。swin transformer V2^[16]进一步优化了

1. 太原师范学院计算机科学与技术学院 山西晋中 030600

2. 太原工业学院计算机工程系 山西太原 030000

3. 中北大学信息探测与处理山西省重点实验室 山西太原 030051

[基金项目] 信息探测与处理山西省重点实验室开放基金资助 (2022-001)

原始 swin transformer 的结构,提高了模型性能和训练稳定性。Han 等人在他们的研究中详细介绍了 transformer 在计算机视觉领域的最新研究进展。

为解决上述问题,本文提出 localswin,即在 U 型架构的基础上,编码部分使用 swin transformer 模块,将反向残差块中所使用的深度卷积引入前馈网络中,从而为其引入局部性。解码部分 swin transformer 模块输出的细化特征逐级上采样,分别与编码部分各级分辨率的特征图进行跳跃连接。最终实验结果得知,该方法增强了模型对图像中不同区域的关联性建模能力,从而获得了更好的分割效果。

1 基础知识

本节简要介绍本文中使用的相關基础知识。首先,第 1.1 节介绍整个 swin transformer 体系结构;然后,第 1.2 节介绍 swin transformer block;最后,第 1.3 节介绍移动窗口自注意力 SW-MSA。

1.1 swin transformer 框架

整个 swin transformer 架构由 4 个 stage 构建,如图 1 所示,每个 stage 中都是类似的重复单元,和 ViT 类似。通过 patch partition 将输入图片 $H \times W \times 3$ 划分为不重合的 patch 集合,其中每个 patch 尺寸为 4×4 ,那么每个 patch 的特征维度为 $4 \times 4 \times 3 = 48$,patch 块的数量为 $H/4 \times W/4$ 。stage1 部分,首先通过一个 linear embedding 将划分后的 patch 特征维度变成 C 。其次送入 swin transformer Block。stage2-stage4 操作相同,先通过一个 patch merging,将输入按照 2×2 的相邻 patches 合并,这样子 patch 块的数量就变成了 $H/8 \times W/8$,特征维度就变成了 $4C$ 。然后和 stage1 一样使用 linear embedding 将 $4C$ 压缩成 $2C$ 。最后送入 swin transformer Block,如图 1 所示。

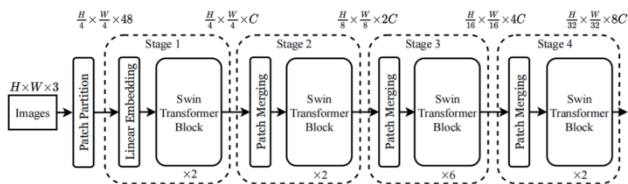


图 1 swin transformer 体系结构

1.2 swin transformer Block

图像在输入到 swin transformer 模块之前通过 patch merging 层进行下采样,生成一张高深度的特征图,经过分块和展平后作为向量输入到 swin transformer 模块中。输入向量首先经过一个 LN (layer norm) 层,再输入到窗口化多头自注意力 (Windows multi-head self-attention, W-MSA) 层中。不同于普通的 MSA (multi-head self-attention)^[17] 层对于特征图中每个像素 (token) 在 Self-Attention 计算过程中需要和所有

的像素去计算,引入 W-MSA 模块是为了减少计算量。首先将特征图以的大小划分成一个个窗口,然后对每个窗口内部单独进行 Self-Attention 计算,可大大降低模型计算量并加快训练速率,最后通过残差结构得到输出向量,再进入下一个 LN 层。

如图 2 所示,两个连续的 swin transformer Block。其中一个 swin transformer Block 由一个带两层 MLP 的 shifted window based MSA 组成,另一个 swin transformer Block 由一个带两层 MLP 的 window based MSA 组成。在每个 MSA 模块和每个 MLP 之前使用 LayerNorm(LN) 层,并在每个 MSA 和 MLP 之后使用残差连接。

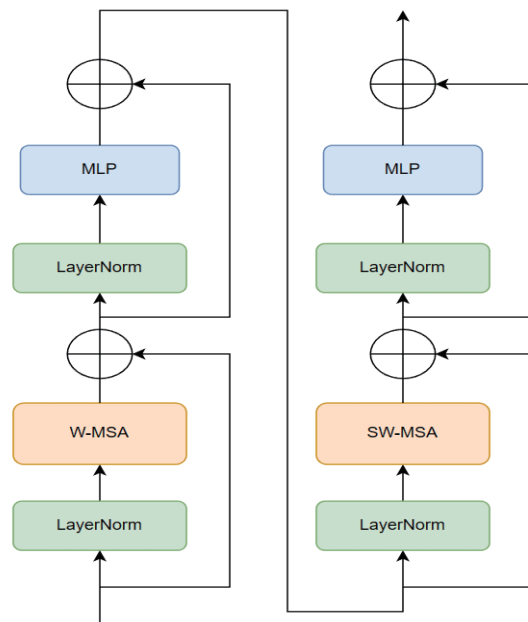


图 2 swin transformer Block

1.3 移动窗口自注意力 SW-MSA

transformer 初衷是理解上下文,是一种信息的传递交互,但采用 W-MSA 模块时,只会在每个窗口内进行自注意力计算,所以窗口与窗口之间是无法进行信息交流。为此引入了 shifted Windows multi-head self-Attention (SW-MSA) 模块来解决这个问题,即进行偏移的 W-MSA。如图 3 所示,左右两幅图对比能够发现窗口发生了偏移,即窗口从左上角分别向右侧和下方各偏移了 $M/2$ 个 patch。

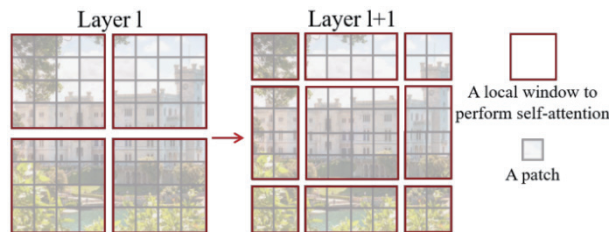


图 3 W-MSA 和 SW-MSA 对图像的窗口化计算

在 L1 层使用的是 W-MSA, L1+1 层使用的是 SW-MSA, 在 L1 时只能同一个窗口里的 patch 相互学习, 而到了 L1+1 层时, 由于窗口的移动, 导致一些 patch 进入新的窗口。这些带有上一层窗口信息的 patch 可以和其他带有上一层前窗口信息的 patch 相互学习, 所以 L1+1 层中心的 4×4 窗口的学习就是 L1 层四个窗口的绝大部分相邻信息融合。因此, SW-MSA 可以使得窗口之间可以进行信息交流, 增强了空间特征信息的提取。

2 localswin 模型结构

为了提高模型分割的准确性, 图像中的语义和空间上下文信息是必不可少的。CNN 受限于卷积核的固定大小, 无法对全局语义信息进行建模。transformer 可以通过自注意力计算获得全局语义信息, 但自注意力计算需要将 patch 拉伸为一维向量, 并且会丢失 patch 内部的空间信息。为了解决这些问题, 在 U-Net 架构的基础上, 编码部分使用 swin transformer 模块, 将反向残差块中所使用的深度卷积引入前馈网络中, 为其引入局部特征。如图 4 所示为 localswin 的网络结构, 其提高了特征表示能力, 在医学图像分割上表现出良好的性能。

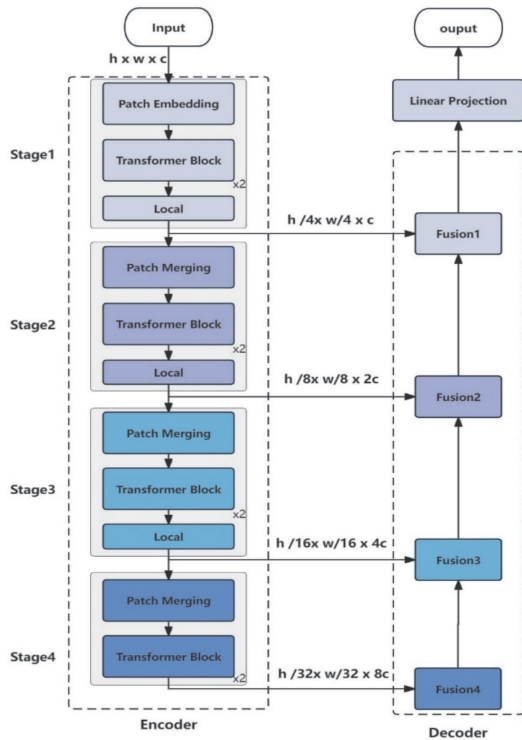
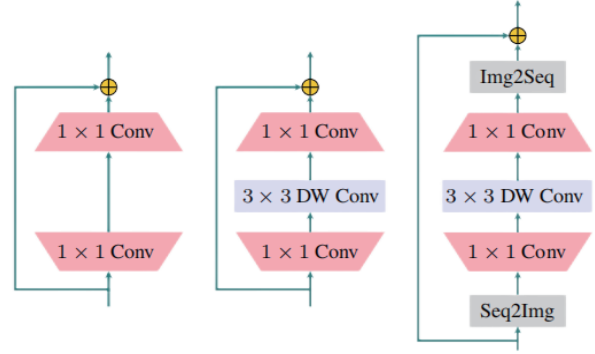


图 4 localswin 的网络框架图

2.1 前馈网络中引入深度卷积

前馈网络和反向残差都通过 1×1 的卷积展开和压缩隐藏维数, 唯一的区别是在反向残差块中存在深度卷积。深度

卷积在每个通道上应用一个 $k \times k$ ($k > 1$) 卷积核。图 5 (a) 所示为视觉变压器中前馈网络的卷积版本, 由于仅对特征图应用了 1×1 卷积, 因此相邻像素之间缺乏信息。此外, transformer 的自注意力部分仅捕获所有令牌之间的全局依赖关系。因此, 在图 5 (c) 中为了应对卷积操作, 通过“Seq2Img”和“Img2Seq”添加了序列和图像特征图之间的转换, 通过这种机制来建模附近像素之间的局部依赖关系。



(a) 为视觉变压器 (b) 为反向残差块 (c) 为将局部性机制引入变压器的网络

图 5 视觉变压器中前馈网络的卷积版本、反向残差块和引入局部性机制的变压器网络之间的比较

考虑到这一点, 在变压器的前馈网络中重新引入了深度卷积。而计算结果可以表示为:

$$Y^r = f(f(Z^r \otimes W_1^r) \otimes W_d^r) \otimes W_2^r \quad (1)$$

式中: $W_d \in R^{d \times 1 \times k \times k}$ 是深度卷积的核心。

2.2 损失函数

整个网络利用加权骰子损失和交叉熵损失 $L = L_{dice}^w + L_e^w$ 进行端到端训练。同时, 还采用了深度监督的方法, 通过额外监督 Fusion2 和 Fusion4 与辅助分割头来解决梯度消失和慢收敛问题。因此, 将总训练损失函数为:

$$L_{total} = \alpha L(G, head(f_0)) + \beta L(G, head(f_2)) + \gamma L(G, head(f_4)) \quad (2)$$

式中: α 、 β 、 γ 、为可调超参数, 实现分割头生成像素级预测, G 为地面真相。 f_i 为 $Fusion_i$ 模块的输出。由于 $Fusion_4$ 是网络的第一个上采样块, 并且没有来自上一个模块的输入, 因此 $Fusion_4$ 模块可以计算为:

$$f_4 = \text{concat}[\text{transpose}_t(SC(x_i))_{t=1}^M] \quad (3)$$

$$f_4 \in R^{\frac{h}{8} \times \frac{w}{8} \times 4C}$$

$Fusion_{2,3}$ 模块可以计算为:

$$f_i = \text{concat}[\text{transpose}_t(SC(x_i, f_{i+1}))_{t=1}^M] \quad (4)$$

$$f_i \in R^{\frac{h}{2^i} \times \frac{w}{2^i} \times 2^{i-2} \times C}, i = 2, 3$$

而 $Fusion_1$ 模块可以计算为:

$$\begin{aligned} f_1 &= \text{concat}[\text{transpose}_t(SC(x_1, f_2))_{t=1}^M] \\ f_1 &\in R^{h \times w \times c} \end{aligned} \tag{5}$$

式中： x_i 为编码器中的特性， M 为转置卷积的总数， SC 为跳跃连接过程。在 $Fusion_1$ 中设置 M 为 4，在 $Fusion_{2,3,4}$ 中设置 M 为 2。

通过对 $Fusion_1$ 模块的输出使用线性投影（LP）层来获得图像 x 的像素级分割。

$$f_0 = \text{Segmentation}(x) = LP(f_1) \tag{6}$$

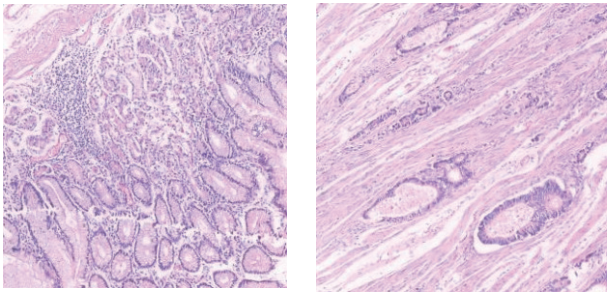
3 实验结果与分析

3.1 实验环境

本文的实验是在 Python 3.8 和 PyTorch 1.11.0 上进行的。在实验开始阶段对所有的数据样本进行了水平翻转、垂直翻转和随机 $[-90^\circ, 90^\circ]$ 旋转，来增强数据的多样性。这些数据均被随机分为 0.8:0.2 的训练集和测试集。模型训练 30 个 epoch，batch size 为 16，训练批次大小为 24。 α 、 β 、 γ 的值分别设为 0.5、0.3、0.2，采用学习率为 $1e-4$ 的 adam 优化器。本文采用的实验显卡为 NVIDIA GeForce RTX3090，内存容量 64 GB，Windows10 专业版操作系统，开发工具为 JetBrains PyCharm 2021.2 专业版。

3.2 数据集与预处理

本研究的数据集来自 BOT 胃癌病理切片识别 AI 挑战赛（网址为：<http://www.datadreams.org/>）。数据集包含 560 个胃癌切片和 140 个正常切片。切片用苏木精-伊红(H&E)染色，放大倍数为 20 倍。胃切片的分辨率为 2048×2048 ，肿瘤区域部分由数据提供者标注。由于深度学习网络的规模太大，无法直接处理。从胃图像中裁剪大小为 224×224 的补丁，形成训练集。对于正常的胃切片，在整个图像上滑动该窗口来生成训练补丁，而该窗口只位于胃癌切片的癌症区域之上。因为数据匮乏，为扩大训练集，对所有的数据样本进行了水平翻转、垂直翻转和随机 $[-90^\circ, 90^\circ]$ 旋转。在实验中，随机选择 80% 的胃切片（正常和癌症）进行网络训练，而剩下的 20% 的切片用于测试。



(a) 胃癌（阳性） (b) 胃正常（阴性）

图 6 胃的病理样本图

3.3 评价指标

为了定量评价方法的性能，采用 Dice 相似系数（dice similarity coefficient, DSC）、准确率（accuracy, Acc）作为评价指标。

Dice 为相似系数：相似系数表示预测目标区域与实际目标区域的相似性。对测试集中所有测试结果的相似系数总和，记作 Dice。Dice 的公式表示为：

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{7}$$

Acc 为准确率：准确率表示正确分类的像素个数和总像素个数之间的比值。准确率越高，分类性能越好。Acc 的计算公式为：

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

式中： TP 、 TN 、 FP 、 FN 的真阳性数，分别代表了特征中属于真阳性、真阴性、假阳性和假阴性的图片像素点数目。

3.4 实验结果

评估了提出的深度学习和几个著名的网络的性能，对比 U-Net、Res-U-Net、FCN、DeeplabV3、ConvNeXt、swin transformer 等先进模型分割精度的对比结果，如表 1 所示。

表 1 BOT 胃切片数据集的分割性能

Model	Dice/%	Acc/%
U-Net	83.67	78.63
Res-U-Net	86.81	78.92
FCN	84.92	76.95
DeeplabV3	87.67	80.10
ConvNeXt	88.05	81.50
swin transformer	90.13	85.54
localswin	91.79	86.27

从表 1 中可以得出，和以往的模型相比，localswin 在 Dice、Acc 指标上效果较好，各指标分别为 91.79% 和 86.27%，比 U-Net 分别高出 8.12% 和 7.64%，比 Res-U-Net 分别高出 4.98% 和 7.35%，比 FCN 分别高出 6.87% 和 9.32%，比 DeeplabV3 分别高出 4.12% 和 6.17%，比 ConvNeXt 分别高出 3.74% 和 4.77%，比 swin transformer 分别高出 1.66% 和 0.73%。由此可见，localswin 在 BOT 胃切片数据集上分割的效果更好。

4 结论与展望

在本文中，提出了一个深度学习框架分割胃癌图像，即 localswin。在 U-Net 架构的基础上，编码部分使用 swin trans-

former 模块, 将反向残差块中所使用的深度卷积引入前馈网络中, 从而为其引入局部性。利用 swin transformer 模块能有效地将特征图像分块, 加快了模型训练的速率。同时将特征图各分块之间的特征信息有效融合用于 WSI 分割, 在公开的 BOT 胃切片数据集上进行了评估。实验结果表明, localswin 的分割准确率达到了 86.27%。

参考文献:

- [1] SHEIKH I A, MIRZA Z, ALI A, et al. A proteomics based approach for the identification of gastric cancer related markers[J]. Current pharmaceutical design, 2016, 22(7): 804-811.
- [2] LIU J, CHENG Z, ZHANG J, et al. Diagnosis of gastric cancer based on hybrid genes selection approach[J]. Biotechnology and genetic engineering reviews, 2023(20): 1-20.
- [3] GHAZNAVI F, EVANS A, MADABHUSHI A, et al. Digital imaging in pathology: whole-slide imaging and beyond[J]. Annual review of pathology, 2013(8): 8:331.
- [4] GURCAN M N, BOUCHERON L, CAN A, et al. Histopathological image analysis: a review[J]. IEEE reviews in biomedical engineering, 2009(2): 147.
- [5] ABELS E, PANTANOWITZ L, AEFFNER F, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association[J]. National center for biotechnology information, 2019, 249 (3): 286-294.
- [6] WANG X Y, FANG Y Q, YANG S, et al. A hybrid network for automatic hepatocellular carcinoma segmentation in H&E-stained whole slide images[J]. National center for biotechnology information, 2021, 68: 101-115.
- [7] LONG J, SHELHMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 3431-3440.
- [8] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI). Piscataway: IEEE, 2015: 234-241.
- [9] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence (PAMI), 2017, 39(12): 2481-2495.
- [10] HUANG G, LIU Z, VAN D M L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 4700-4708.
- [11] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM computing surveys, 2023, 55(9): 1-35.
- [12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J/OL]. arXiv preprint arXiv:2010.11929, [2023-05-14]. <https://doi.org/10.48550/arXiv.2010.11929>.
- [13] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.
- [14] ZHENG S, LU J, ZHAO H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 6877-6886.
- [15] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 10012-10022.
- [16] LIU Z, HU H, LIN Y, et al. Swin transformer v2: Scaling up capacity and resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway: IEEE, 2022: 12009-12019.
- [17] VOITA E, TALBOT D, MOISEEV F, et al. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned[EB/OL]. (2019-05-23) [2023-07-01]. <https://arxiv.org/abs/1905.09418>.

【作者简介】

毛松仁(1997—), 男, 江苏盐城人, 研究方向: 深度学习、图像处理。

刘杰(1980—), 男, 山西武乡人, 博士, 副教授, 研究方向: 图形图像处理、数据挖掘、深度学习。

(收稿日期: 2023-09-12)