Computer Science and Application 计算机科学与应用, 2024, 14(1), 134-146
Published Online January 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.141015

Hans汉斯

融合Transformer和CNN的U型神经网络遥感影像道路提取算法

高　琳，陈　晨，张咏琪

沈阳理工大学，信息科学与工程学院，辽宁 沈阳

收稿日期：2023年12月26日；录用日期：2024年1月24日；发布日期：2024年1月31日

摘　要

U型网络作为一种经典的编码－解码结构网络，不只在医学影像领域内发挥出色，在图像分割领域也有广泛的影响。以U型网络为基础其它衍生网络层出不穷。U型网络最经典的思想是编码和解码，再加上编解码之间的跳跃连接。由于道路遥感图像和医学影像有众多相似的地方，如今U型网络又被用于从遥感图像中提取道路。U型网络使用跳跃连接的方式将下采样低维特征拼接到上采样的高维特征中，以保留更多的空间位置信息和语义信息。因此U型网络更能处理一些特征信息明显的图像数据。但浅层的UNet无法准确提取道路丰富多维的细节信息，在高分辨率卫星遥感图像上无法奏效。所以本文提出一种融合蛇形动态卷积和Swin-Transformer的U型网络用于提高道路提取任务的分割精确度。

关键词

U型网络，遥感图像，蛇形动态卷积，Swin-Transformer，道路提取

Remote Sensing Image Road Extraction Algorithm Based on U-Type Neural Network and Transformer Combined with CNN

Lin Gao, Chen Chen, Yongqi Zhang

College of Information Science and Engineering, Shenyang Ligong University, Shenyang Liaoning

Received: Dec. 26th, 2023; accepted: Jan. 24th, 2024; published: Jan. 31st, 2024

Abstract

As a classical coding-decoding structure network, U-shaped network not only plays an excellent role in the field of medical imaging, but also has a wide impact in the field of image segmentation. Based on U-shaped network, other derivative networks emerge in an endless stream. The most classic idea of U-shaped networks is encoding and decoding, plus jumping connections between coding and decoding.

文章引用: 高琳, 陈晨, 张咏琪. 融合 Transformer 和 CNN 的 U 型神经网络遥感影像道路提取算法[J]. 计算机科学与应用, 2024, 14(1): 134-146. DOI: 10.12677/csa.2024.141015

Due to the similarities between road remote sensing images and medical images, U-shaped networks are now used to extract roads from remote sensing images. The U-shaped network spliced the low-dimensional features from the down-sampled to the high-dimensional features from the up-sampled by means of jump connection, so as to retain more spatial location information and semantic information. Therefore, U-shaped network is more capable of processing some image data with obvious feature information. However, shallow UNet can not accurately extract the rich multi-dimensional detailed information of the road, and can not be effective in high-resolution satellite remote sensing images. Therefore, this paper proposes a U-shaped network combining serpentine dynamic convolution and Swin-Transformer to improve the segmentation accuracy of road extraction tasks.

Keywords

U-Shaped Network, Remote Sensing Images, Serpentine Dynamic Convolution, Swin-Transformer, Road Extraction

1. 网络结构

由于 Transformer 将整张图片作为序列作为网络输入时，会在所有阶段内只专注于全局上下文信息的建模，因此会导致缺乏详细的低分辨率特征。若直接上采样到原分辨率会无法有效地恢复该信息，从而

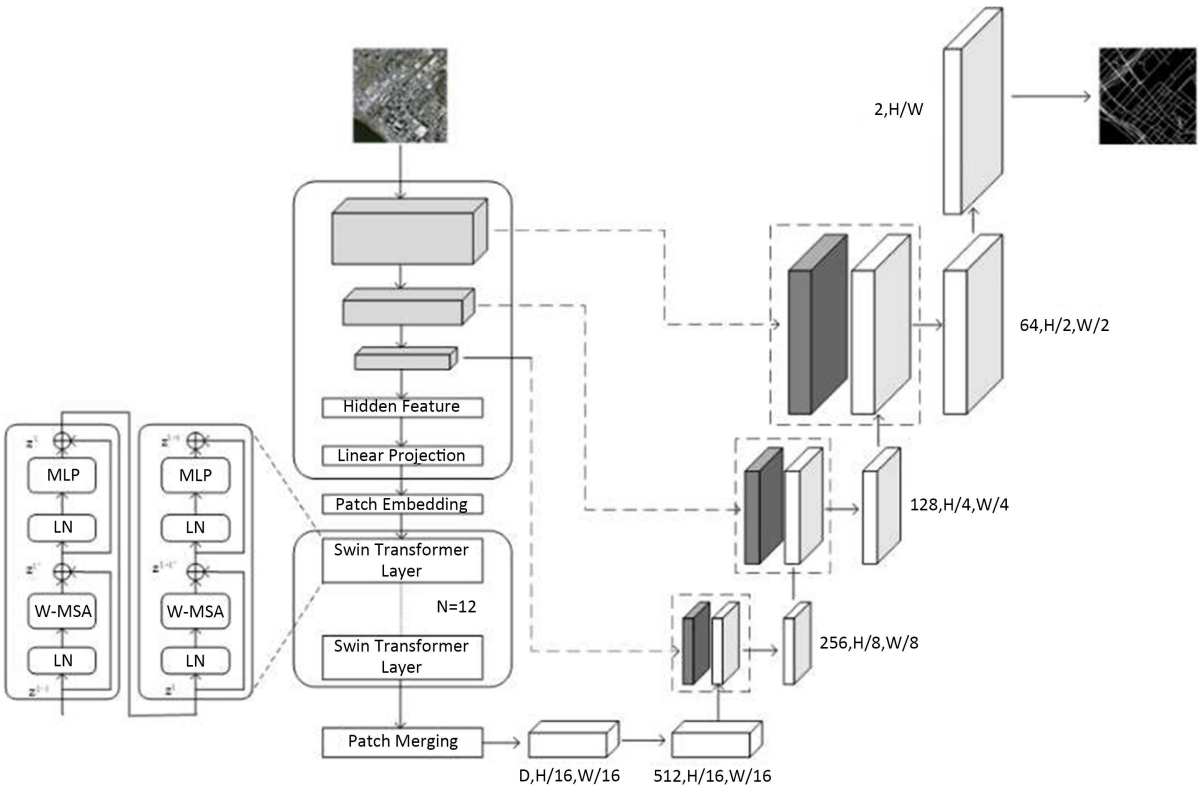


Figure 1. The structure schematic diagram of STransUNet
图 1. STransUNet 的结构示意图

导致分割结果不够精确。因此，本文采用一种融合 CNN-SwinTransformer 的网络结构 STransUNet，该网络基于经典的 U 型编码－解码结构实现，如图 1 所示。

在编码器阶段，不同于常规的对称编码－解码结构，STransUNet 在编码部分中首先将图像输入到 3 个 CNN 卷积块当中，每个卷积块由 2 个卷积层、BN 层、ReLU 激活函数和最大池化层组成，并在首个卷积块将常规卷积层替换为蛇形动态卷积模块[1]，目的是为了更加有效地提取有管状特征的道路。

输入图像经过 CNN 部分后会输出一张特征图，在经过 Patch Embedding 层后将特征图序列化并且加上位置编码，输入到堆叠的 Swin-Transformer 块进行深层编码，本文中在这一层中设置了数量为 12 的 Swin-Transformer 块。

在解码器阶段，首先将对特征图输出进行 Patch Merging，再经过卷积核大小为 1 的卷积层进行降维得到深层特征图。之后进行级联上采样，同时与编码器的 CNN 部分进行跳跃连接，将经过上采样之后的特征图与编码器部分对应大小的特征图进行拼接操作，最后逐层恢复到原图的输入大小，输出最终预测结果。

1.1. 蛇形动态卷积模块

对于道路这种管状类结构的特征精确提取目前仍然具有挑战性：

细长且脆弱的局部结构。如图 2 所示，细长的结构仅占整个图像的一小部分，像素的组成有限。此外，这些结构容易受到复杂背景的干扰，因此模型很难精确分辨目标的细微变化，从而导致分割出现破碎与断裂。

Figure 2. Slender tubular structure
图 2. 细长管状结构

复杂且多变的全局形态。图 2 显示了细长管状结构复杂多变的形态，即使在同一张图像中也是如此。位于不同区域的目标的形态变化取决于分支的数量、分叉的位置，路径长度以及其在图像中的位置。因此当数据表现出未曾见过的形态特征时，模型倾向于过拟合到已见过的特征，无法识别未见过的特征形态，从而导致泛化性较弱。

为了获得更好的性能，在计算机视觉领域里已经提出了各种方法，根据管状结构的形态设计特定的网络体系结构和模块。(以著名的膨胀卷积[2]和可变形卷积[3]为代表，提出了基于卷积核设计的方法来处理 CNN 中固有的有限几何变换，在复杂的检测和分割任务中表现出了优异的性能。这些方法[4] [5] [6] [7]也被设计成动态地感知对象的几何特征，以适应具有可改变形态的结构。例如，文献[6]中提出的 Dunet 算法将可变形卷积算法集成到 U 形结构中，并根据血管的大小和形状自适应地调整接收野。因此提出了基于网络结构设计的方法来学习管状结构的特殊几何拓扑特征。PointScatter [8]提出了用点集来表示管状结构，替代了管状结构提取任务中的分段模型。文献[9]提出了一种树形结构的卷积门控复发单元来显式模拟冠状动脉的拓扑结构。与上述允许模型完全自由学习几何变化的思想不同，考虑到过度随机性带来

的收敛困难的局限性以及模型可能会聚焦于目标的意外区域。蛇形动态卷积模块结合了管状结构形态的领域知识，在特征提取过程中稳定地提高了管状结构的百分率。

首先本文的目标是在遥感影像中提取竖直的管状道路，蛇形动态卷积模块从可变形卷积的概念中获得灵感，使模型能够在学习特征的同时动态适应其卷积核的形状。这种方法使模型能够集中于管状地层的基本结构属性。然而，在最初的实验中，由于管状结构的比例相对较小，模型往往会失去对这些特定结构的感知。因此，卷积核显著偏离其预期焦点。为了解决这个问题，蛇形动态卷积模块提出了一种专门针对管状结构特点的网络结构的设计。这种专门的结构作为一个指导框架，确保该模型有效地确定关键功能的优先顺序。

蛇形动态卷积的目标是允许卷积核自由地适应结构以进行有效的特征学习，同时确保它在定义的约束内不会偏离目标结构太远。这一观察使蛇形动态卷积将其与动物的特征进行了类比：蛇。其设想了卷积核像蛇一样动态地扭曲和扭曲以符合目标结构，从而能够更精确地提取特征。

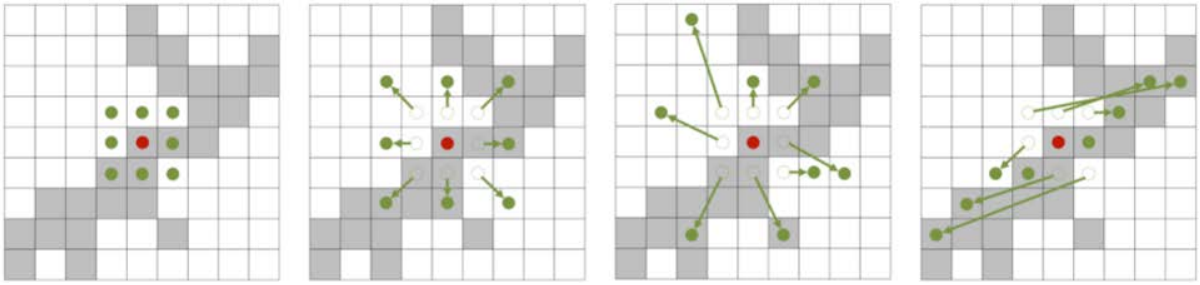


Figure 3. Snake dynamic convolution diagram
图 3. 蛇形动态卷积示意图

结果完全自由的成型有时会导致精细结构细节的丢失，这在分割微妙的管状结构的背景下构成了一个巨大的挑战。如图 3 所示，动态蛇形卷积从蛇的运动中获得灵感，蛇的头部一节接一节地引导身体不断前进，创造了波浪式的运动。

为了解决这个问题，蛇形动态卷积在卷积核的设计中引入了连续性约束。在每个卷积位置，前一位置用作自由选择卷积的移动方向的参考点。这种方法既保证了自适应结构的自由度，又保证了特征感知的连续性。

为了使卷积核更灵活地聚焦于目标的复杂几何特征，动态蛇形卷积引入了变形偏移Δ。然而，如果让模型自由学习变形偏移量，感知场往往会偏离目标，特别是在薄管结构的情况下。因此，动态蛇形卷积使用了 ITER-ative 策略，为每个要处理的目标依次选择要观察的以下位置，从而确保注意力的连续性，并且不会因为大的变形偏移量而将感觉领域扩展得太远。

$$K=\{(x-1,y-1),(x-1,y),\cdots,(x+1,y+1)\} \tag{1-1}$$

在 DSConv 中，其将标准卷积算子在 x 轴和 y 轴方向上都拉直。DSConv 考虑大小为 9 的卷积核，并以 x 轴方向为例，每个网格在 K 中的具体位置表示为：$k_i \pm c=(x_i \pm c, y_i \pm c)$，其中 $c=\{0,1,2,3,4\}$ 表示到中心网格的水平距离。卷积核 K 中每个网格位置 $k_i \pm c$ 的选择是一个累加过程。从中心位置 K_i 开始，远离中心网格的位置取决于前一网格的位置：与 K_i 相比，用偏移量 $\Delta=\{\in[-1,1]\}$ 来增加 $\delta|\delta+1$。因此，偏移量需要为 Σ，从而确保卷积核符合线性形态结构。公式 1-1 在 x 轴方向变为：

$$K_{i\pm c}=\begin{cases}(x_{i+c},y_{i+c})=\left(x_i+c,y_i+\sum_i^{i+c}\Delta y\right),\\(x_{i-c},y_{i-c})=\left(x_i-c,y_i+\sum_{i-c}^{i}\Delta y\right),\end{cases} \tag{1-2}$$

并且 y 轴方向上的公式 1-2 变成：

$$K_{j\pm c}=\begin{cases}\left(x_{j+c},y_{j+c}\right)=\left(x_j+\sum_j^{j+c}\Delta x,y_i+c\right),\\ \left(x_{j-c},y_{j-c}\right)=\left(x_j+\sum_{j-c}^{j}\Delta x,y_i-c\right),\end{cases} \tag{1-3}$$

其中 K 表示公式 1-2 和公式 1-3 的分数位置，K'枚举所有整数空间位置，B 是双线性内插核，它被分成两个一维核，如下公式 1-4 所示：

$$B\left(K,K'\right)=b\left(K_x,K'_x\right)\cdot b\left(K_y,K'_y\right) \tag{1-4}$$

如图 3 所示，由于二维(x 轴、y 轴)的变化 DSConv 在变形过程中覆盖了 9 × 9 的范围。DSConv 的设计是为了更好地适应细长筒体结构的动态结构，从而更好地感知关键特征。

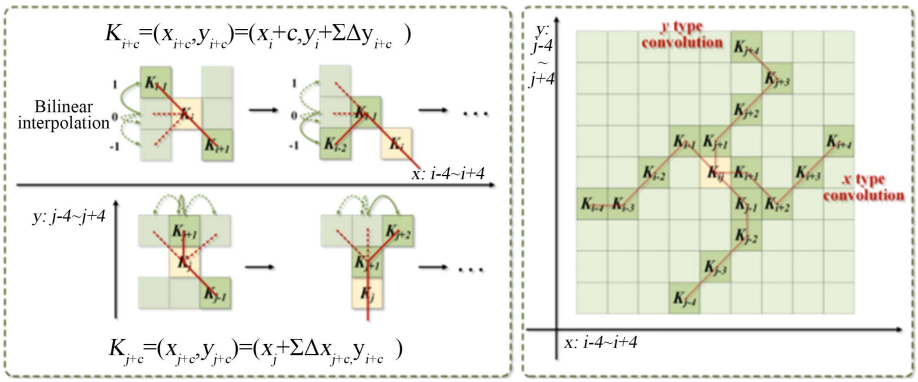


Figure 4. The working principle of snake dynamic convolution
图 4. 蛇形动态卷积的工作原理

1.2. Swin Transformer 模块

传统的 Transformer 模型在自然语言处理领域表现出色，但在计算机视觉任务中，特别是在处理高分辨率图像时，存在着一些挑战。原始 Transformer 架构不适合处理大规模图像，因为它需要在全局上计算自注意力，这导致了高昂的计算成本和内存需求。Swin-Transformer [10]的提出正是为了应对这一问题，使得 Transformer 可以更好地处理大尺寸图像数据。Swin Transformer 使用了类似卷积神经网络中的层次化构建方法(Hierarchical feature maps)，比如特征图尺寸中有对图像下采样 4 倍的，8 倍的以及 16 倍的，这样的 backbone 有助于在此基础上构建目标检测，实例分割等任务。而在之前的 Vision Transformer 中是一开始就直接下采样 16 倍，后面的特征图也是维持这个下采样率不变。

在 Swin Transformer 中使用了 Windows Multi-Head Self-Attention (W-MSA)的概念，比如在下图的 4 倍下采样和 8 倍下采样中，将特征图划分成了多个不相交的区域 0，并且 Multi-Head Self-Attention 只在每个窗口内进行。相对于 Vision Transformer 中直接对整个特征图进行 Multi-Head Self-Attention，这样做的目的是能够减少计算量的，尤其是在浅层特征图很大的时候。这样做虽然减少了计算量但也会隔绝不同窗口之间的信息传递，所以在论文中作者提出了 Shifted Windows Multi-Head Self-Attention (SW-MSA)的概念，通过此方法能够让信息在相邻的窗口中进行传递。

通过图 5 可以看出整个框架的基本流程如下：

首先将图片输入到 Patch Partition 模块中进行分块，即每 4 × 4 相邻的像素为一个 Patch，然后在 channel 方向展平。假设输入的是 RGB 三通道图片，那么每个 patch 就有 4 × 4 = 16 个像素，然后每个像素有 R、

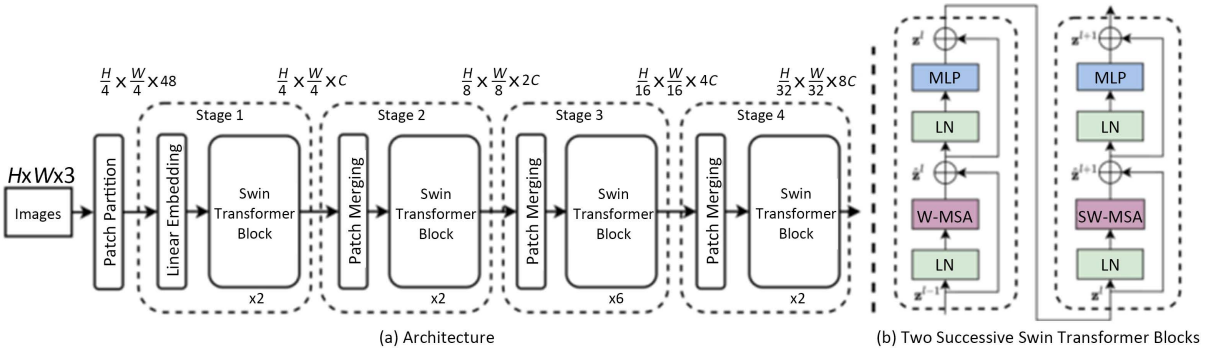


Figure 5. Swin transformer structure diagram
图 5. Swin Transformer 结构示意图

G、B 三个值所以展平后是 16 × 3 = 48，所以通过 Patch Partition 后图像 shape 由[H, W, 3]变成了[H/4, W/4, 48]。然后在通过 Linear Embeding 层对每个像素的 channel 数据做线性变换，由 48 变成 C，即图像 shape 再由[H/4, W/4, 48]变成了[H/4, W/4, C]。其实在源码中 Patch Partition 和 Linear Embeding 就是直接通过一个卷积层实现的，和之前 Vision Transformer 中的 Embedding 层结构一模一样。

然后就是通过四个 Stage 构建不同大小的特征图，除了 Stage1 中先通过一个 Linear Embeding 层外，剩下三个 stage 都是先通过一个 Patch Merging 层进行下采样。然后都是重复堆叠 Swin Transformer Block 注意这里的 Block 其实有两种结构，如图 6 中所示，这两种结构的不同之处仅在于一个使用了 W-MSA 结构，一个使用了 SW-MSA 结构。而且这两个结构是成对使用的，先使用一个 W-MSA 结构再使用一个 SW-MSA 结构。

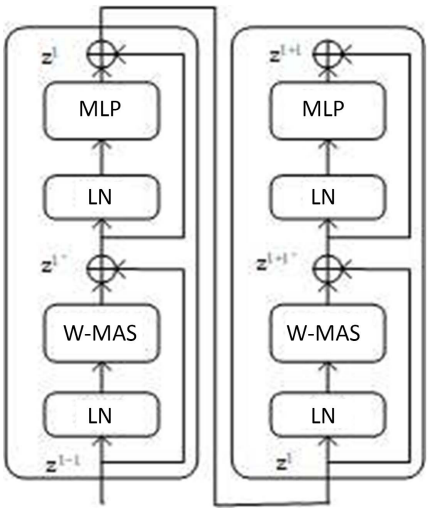


Figure 6. Paired Swin Transformer modules
图 6. 成对的 Swin Transformer 模块

2. 实验设置

实验的硬件配置为 NVIDIA GeForceGTX3060 (12GB)GPU、操作系统是 win10，并使用 Pytorch 作为本文的深度学习框架。在训练网络时，为合理利用计算资源，将 batchsize 设置为 16，学习率设置为 1e-4，并在每 20 轮训练轮次结束后将其调整为原学习率的 20%。此外，训练时采用 Adam 算法对模型进行优化。

在本文中，使用的数据集为美国马萨诸塞州地区的遥感影像数据集和高分二号遥感影像数据集作为

实验数据，图像尺寸为 1500 × 1500，由于该数据集中包含背景内容丰富，包括乡村、城市和山地等不同的地理环境，且面积大、范围广，能够满足实验需求。首先 1500 × 1500 像素的数据集裁剪成模型输入需要的 256 × 256 × 3 大小的数据集，并按照 6:2:2 的比例分割成训练集、测试集和验证集，所有数据集内的图像大小均为 256 × 256 × 3。接下来我们对准备完毕的道路数据集进行预处理，其目的是为了通过对数据集进行翻转、裁剪等操作，可以生成额外的训练样本，从而增加训练数据的多样性。这有助于提高模型的泛化能力，使其更好地适应不同场景和变化。

3. 道路提取结果分析

3.1. 消融实验

本文为了展示出 U 型结构下融合蛇形动态卷积和 Swin Transformer 模块带来的道路提取任务精度提升，在 Massachusetts 道路数据集上进行了消融实验。首先使用未使用 Swin Transformer 与蛇形动态卷积模块的网络结构，使用纯 ViT 模块的 U 型网络结构作为基准，设计添加 Swin Transformer 与蛇形动态卷积模块的网络模型，并且在此基础上使用 12 个 Swin Transformer 模块作为对比实验，测试并且对比这几组模型之间的道路提取性能。本文所使用的模型使用了 12 个 Swin Transformer 模块。

选取其中一张道路提取结果进行评估分析，道路提取结果如图 7 所示，提取精度如表 1 所示。

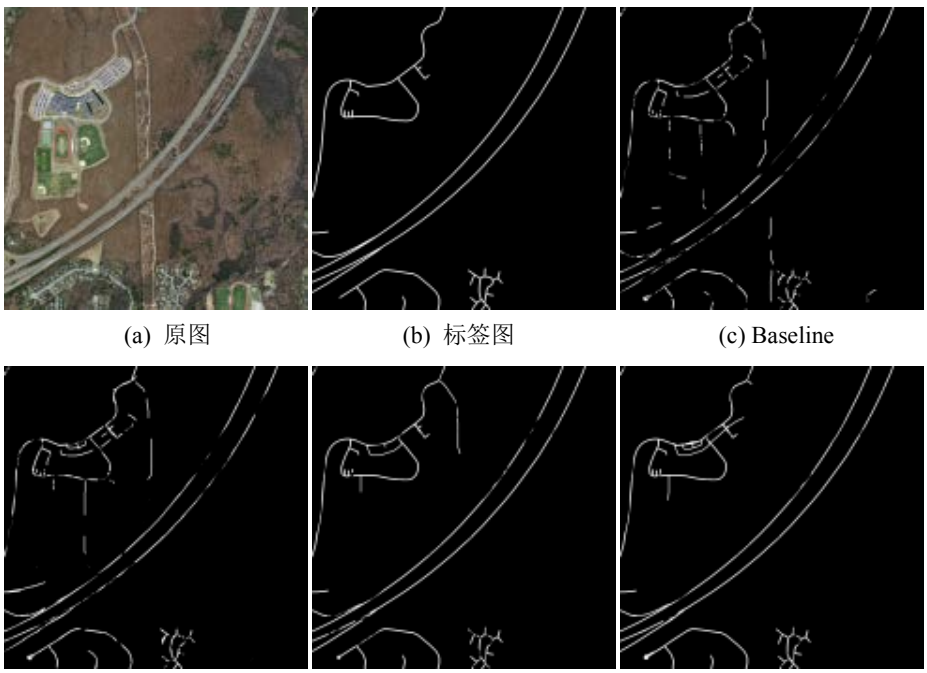


Figure 7. Road extraction results under Ablation experiment
图 7. 消融实验下的道路提取结果

在图 7 中可以看出，未进行改进的纯 ViT 模型的提取结果 c 有诸多误提取结果，且道路提取结果不连贯，也受到遮挡物的干扰，导致其精度不高。图 7d 中是将模型前两层替换成 DSConv 进行提取的结果，可以看出误提取的道路像素减少，并增加了其结果的连贯性。图 7e 为在 Baseline 的基础上将 ViT 模块替换成了 Swin Transformer 模块，道路提取精度明显提高，且断裂减少，也明显规避了许多误提取的道路像素。图 7g 为本文使用的模型，在添加了 DSConv 模块与 12 个 Swin Transformer 模块后，道路提取精

度为四个模型中最高，道路连续性有很大提升，且受到干扰减少，误提取的道路像素在图 7e 的基础上也有所下降。

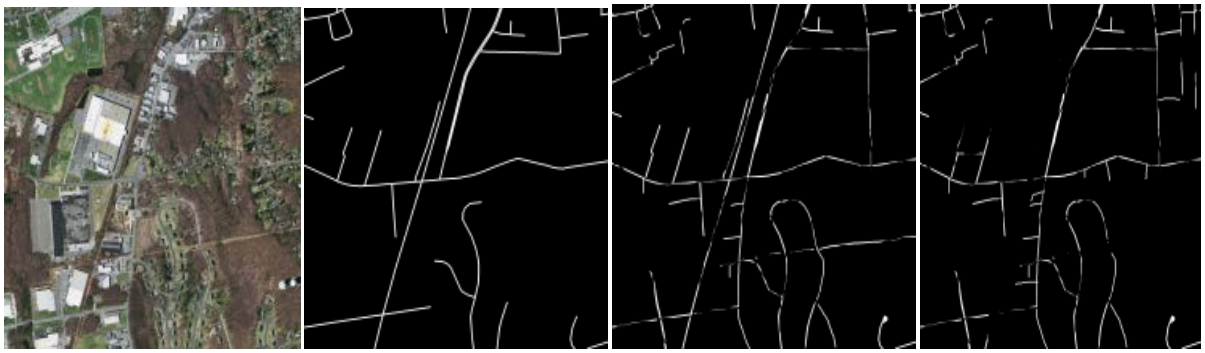
Table 1. Various compared results in the Massachusetts Road dataset Figure 7
表 1. 在 Massachusetts 数据集图 7 下多模型对比实验

Methods	OA (%)	P (%)	R (%)	F1 (%)	IoU (%)
Baseline	94.85	95.03	99.45	96.94	94.52
Baseline + DSConv	95.16	95.68	99.79	97.69	95.03
Baseline + 12 ST	96.79	98.32	97.64	97.97	96.66
STransUNet	97.07	98.81	98.06	98.43	96.97

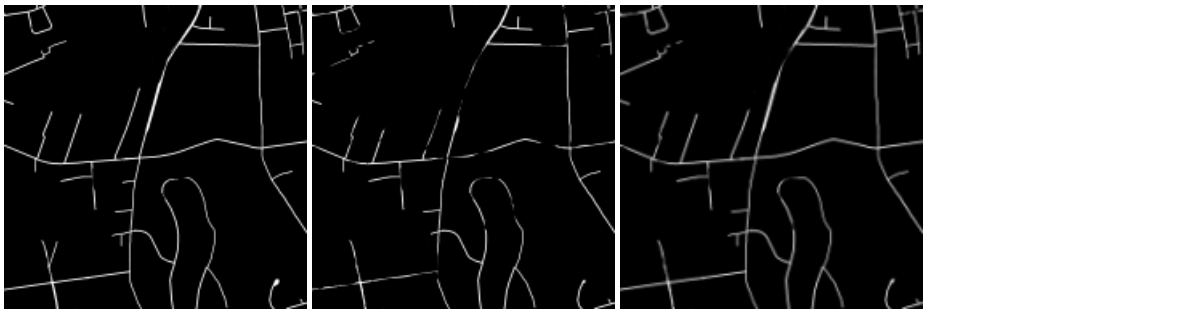
表 1 为消融实验在图 7a 数据集上的评估结果，可以看到未添加 DSConv 与 Swin Transformer 模块的 Baseline 模型 OA 最低，为 94.85%，IoU 也在四个模型中最低，为 95.03%。在添加了 DSConv 模块后，OA 有所提升，提升了 0.31 个百分点，达到了 95.16%，IoU 也提升了 0.51 个百分点，达到了 95.03%，可以得出添加了 DSConv 模块的网络模型在道路提取任务中是有效的，但是与替换 ViT 为 Swin Transformer 的模块相比，提升有限。替换 Swin Transformer 模块的网络模型的 OA 相较于原模型，提升了 1.94 个百分点，IoU 相较于原模型，提升了 2.14 个百分点，精度提升显著，其原因可能是由于 ViT 模型在对遥感影像进行推演处理时的效果相较于 Swin Transformer 缺少了滑动窗口的处理，在推演过程中对遥感影像的全局特征提取未有 Swin Transformer 有效。在表 1 的最后一行为本文提出的 STransUNet 模型，OA 达到了最高的 97.07%，IoU 也为最高值，达到了 96.97%，相较于只替换 Swin Transformer 的 Baseline 提升为 0.3 个百分点左右，说明 DSConv 模块在道路提取任务中也起到了一个正面的作用，帮助模型进行更好的管状特征提取。

3.2. Massachusetts 道路数据集

图 8 为各个模型在 Massachusetts 道路测试集上的对比实验，表可以看到图 8c 和图 8d 中 CNN 架构的网络模型在对于尺度较大的道路提取上有断裂的问题存在，同时对于细微的道路像素也有着不错的提取精度。图 8e 为纯 Transformer 架构的 Swin-UNet 提取结果，由于其对于全局特征的提取优势，在对于大视野下的道路提取结果有着非常高的提取精度。图 8f 为本文的 Baseline 模型，图 8g 为本文提出的网络模型 STransUNet，可以看到图 8g 在图 8f 的基础上，提取精度大大提高，同时路网的断裂、误提取像素也大大减少，可以说明该模型在 Massachusetts 道路数据集上有着不错的性能，同时在 Baseline 的性能基础上有进一步的提高。

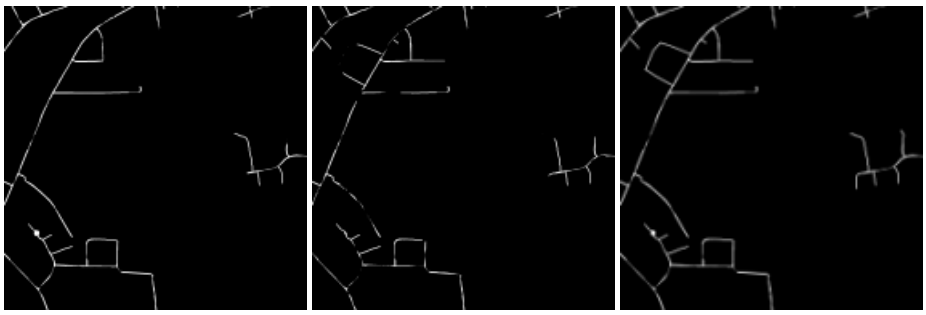


(a) 原图　(b) 标签图　(c) UNet　(d) Deeplabv3+



(e) Swin-UNet (f) TransUNet (g) StransUNet

(a) 原图 (b) 标签图 (c) UNet (d) Deeplabv3+



(e) Swin-UNet (f) TransUNet (g) StransUNet

Figure 8. Different results outputting of five methods from Massachusetts dataset
图 8. 从 Massachusetts 的数据集输出五种方法不同结果

Table 2. Quantitative evaluation results in Figure 8
表 2. 图 8 的精度评价指标

Methods	Image 1				
	OA (%)	P (%)	R (%)	F1 (%)	IoU (%)
U-Net	95.47	95.45	99.20	95.32	95.31
DeepLabv3+	95.06	94.89	99.75	97.25	94.97
Swin-UNet	96.76	98.04	96.81	97.42	96.81
TransUNet	95.97	97.56	99.01	97.27	96.04
STransUNet	96.74	98.21	97.23	97.71	96.88
Methods	**Image 2**				
	OA (%)	**P (%)**	**R (%)**	**F1 (%)**	**IoU (%)**
U-Net	96.24	96.17	98.97	98.11	96.01

续表

DeepLabv3+	96.03	95.89	99.02	97.42	96.36
Swin-UNet	97.25	99.26	98.02	98.63	97.30
TransUNet	97.01	98.21	99.24	98.22	96.89
STransUNet	97.38	99.31	98.25	98.76	97.55

表 2 为图 8 中五种方法在 Massachusetts 道路数据集上的精度评估，其中本文提出的 STransUNet 在两张图片中都取得了较好的表现。其中两张图片的平均 OA 达到了 97.06%，为五种模型中的最高值，平均 IoU 也有较好的表现，而且与 Baseline 模型 TransUNet 的性能进行对比，道路提取的精度都有明显提高。相较于传统的 CNN 网络 UNet 和 DeepLabv3+两个模型，OA 平均提升了 1 个百分比以上，IoU 也平均提升了 1.5 个百分比以上。

Table 3. Various compared results in the Massachusetts Road test dataset
表 3. 在 Massachusetts 数据集下多模型的对比实验

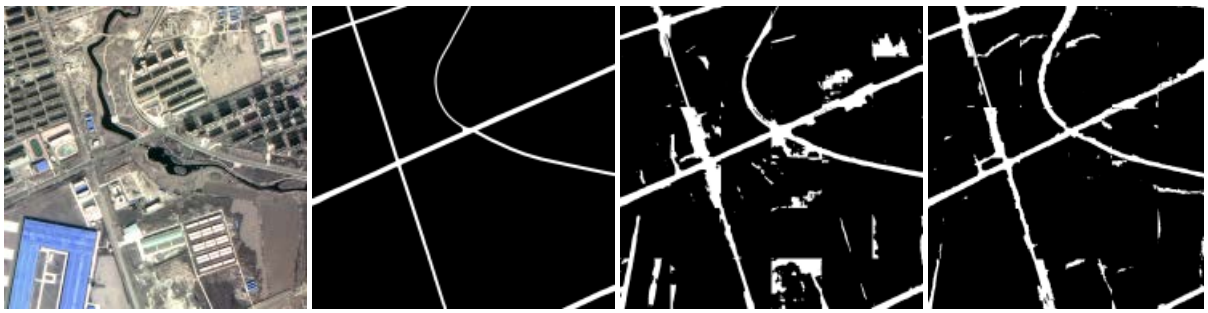
Dataset	Methods	OA (%)	P (%)	R (%)	F1 (%)	IoU (%)
Massachusetts	U-Net	89.95	65.26	87.78	74.16	60.73
	DeepLabv3+	90.46	66.40	81.30	72.82	61.65
	Swin-UNet	93.07	81.67	71.58	76.29	61.67
	TransUNet	91.36	72.05	82.63	76.97	60.88
	STransUNet	92.98	82.79	73.60	77.92	61.94

表 3 为各个模型在 Massachusetts 道路验证集上的对比实验，CNN 架构的 U-Net 与 DeepLabv3+模型在道路提取任务上的平均 OA 在 90.25%左右，平均 IoU 在 61.2%左右，为对比实验里所有模型的最低值。

作为 Baseline 的 TransUNet 在 Massachusetts 数据集下的 OA 与 IoU 与传统 CNN 架构的网络相比，略微存在一些优势。本文提出的 STransUNet 在 TransUNet 的基础上 OA 提高了 1.6 个百分点，IoU 提高了 1 个百分点，与 Swin-UNet 网络的性能相差无几，IoU 相较于 Swin-UNet 略微高出 0.3 个百分点。

3.3. GF-2 道路数据集

高分二号数据集中的路段多为城市、乡镇路段，道路特征相对明显，但由于城市与乡镇中车辆、建筑物数量多，导致阴影遮挡的区域相对于 Massachusetts 数据集更大，因此对于道路提取结果连通性的测试有较的高研究意义。



(a) 原图 (b) 标签图 (c) UNet (d) Deeplabv3+

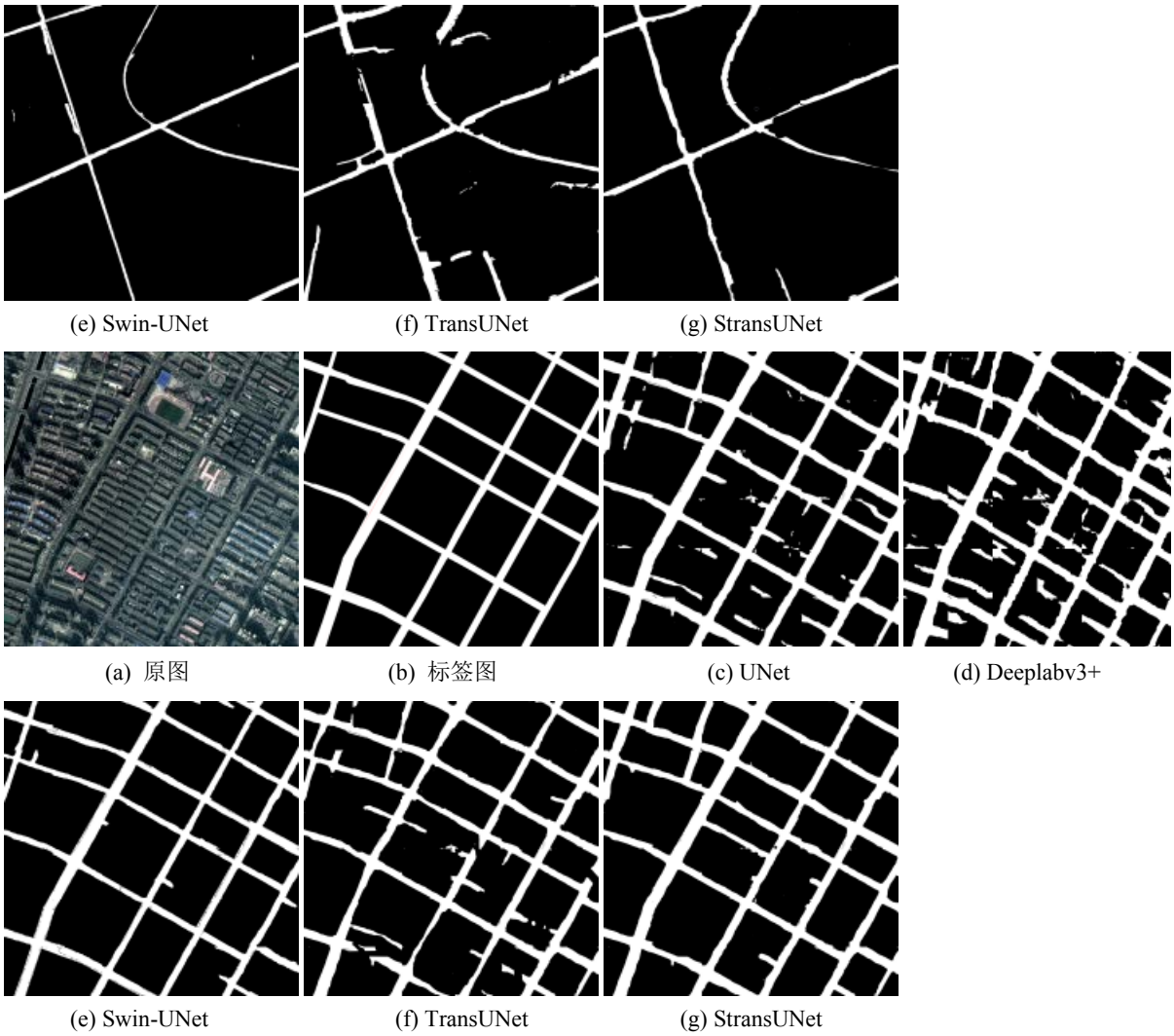


Figure 9. Different results outputting of five methods from GF-2 dataset
图 9. 从 GF-2 的数据集输出五种方法的不同结果

从图 9 中可以看到，CNN 架构下的预测图像 c 和 d 表现出较多的道路误提取，而 Transformer 架构下的 e、f 和 g 预测结果，并未有过多的误提取表现，且道路主干提取结果较为完整，但是 g 作为 Baseline 的 TransUNet 偶尔有道路断裂的表现，对于这种现象的出现，是由于建筑物导致的阴影过多。g 预测图像对于 f 中出现的断裂现象有很大改进，道路连通性也较为良好。

Table 4. Various compared results in the GF-2 Road test dataset
表 4. 在 GF-2 数据集下多模型的对比实验

Dataset	Methods	OA (%)	P (%)	R (%)	F1 (%)	IoU (%)
GF-2	U-Net	90.52	56.52	75.11	64.50	56.24
	DeepLabv3+	90.69	54.87	78.38	64.55	58.16
	Swin-UNet	93.87	72.30	63.23	67.46	72.46
	TransUNet	92.03	63.79	71.24	63.20	67.52
	STransUNet	93.79	75.68	63.01	68.76	72.84

表 4 显示了在 GF-2 验证集下五种网络模型的性能对比，可以看到 STransUNet 在六种模型中的 IoU 为最优值，达到了 72.84%，相较于传统的 CNN 架构网络 U-Net 与 DeepLabv3+平均提高了 15 个百分点，OA 与 Swin-UNet 相差无几，分别达到了 93.87%与 93.79%。相比于经典的卷积神经网络 U-Net 和 DeepLabv3+有明显的提升，同时与同是 Transformer 架构的 TransUNet 相比，也有性能上的提升，OA 和 IoU 平均提升了 1.7%和 5%，F1 为五个模型中的最高值 68.76%。

Table 5. Quantitative evaluation results in Figure 9
表 5. 图 9 的精度评价指标

Methods	Image 1				
	OA (%)	P (%)	R (%)	F1 (%)	IoU (%)
U-Net	93.08	93.11	99.66	96.28	92.83
DeepLabv3+	95.77	96.01	99.57	97.76	95.62
Swin-UNet	98.64	99.52	99.22	99.36	98.61
TransUNet	97.38	97.25	99.62	98.42	97.65
STransUNet	98.44	99.41	99.13	99.07	98.53
Methods	**Image 2**				
	OA (%)	**P (%)**	**R (%)**	**F1 (%)**	**IoU (%)**
U-Net	94.32	93.31	99.67	96.39	93.03
DeepLabv3+	89.45	87.09	99.82	93.02	86.96
Swin-UNet	96.71	98.48	97.65	98.05	96.11
TransUNet	95.26	92.16	99.38	95.63	94.95
STransUNet	96.57	98.54	98.06	98.19	96.03

表 5 中的数据为五个模型在图 9 中 GF-2 数据集中两个提取结果的精度评价指标，其中 STransUNet 的性能与 Swin-UNet 相当，OA 与 IoU 分别在两张图中达到了最高值，平均 OA 分别为 97.67%与 97.51%，平均 IoU 分别为 97.36%与 97.28%，道路完整性与道路连通性均为五个提取结果中的最优，且道路误提取像素少。相比于 TransUNet 模型，STransUNet 在两张提取结果中的平均 OA 提升了 1.1 个百分点，平均 IoU 提升了 1 个百分点，道路连通性有明显提高。且在表 5 与图 9 可以明显看出，融合了 Transformer 架构的网络模型在道路提取任务中有较大优势。

4. 结论

本文提出了一种融合 Transformer 和 CNN 的 U 型神经网络遥感影像道路提取算法，其中主要提出了一种融合了蛇形动态卷积和 Swin Transformer 作为网络的编码器端，采用上采样与跳跃连接的方式来对特征提取结果进行还原的网络结构。通过消融实验和多个模型在 Massachusetts 道路数据集和高分二号数据集上的对比实验证明，蛇形动态卷积在对道路这类管状结构的提取起到积极的作用，以及 Swin Transformer 在对大尺度的长距上下文关系的提取相对于 ViT 模块更加有效，同时也能够得出 STransUNet 在对山地、农村等道路进行提取时有较好的性能表现，提取精度明显优于传统的 CNN 网络和纯 Transformer 网络。

参考文献

[1] Qi, Y., He, Y., Qi, X., *et al.* (2023) Dynamic Snake Convolution based on Topological Geometric Constraints for Tu-

bular Structure Segmentation. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, Francem, 1-6 October 2023, 6070-6079. https://doi.org/10.1109/ICCV51070.2023.00558

[2] Yu, F., Koltun, V. and Funkhouser, T. (2017) Dilated Residual Networks. *Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 21-26 July 2017. https://doi.org/10.1109/CVPR.2017.75

[3] Dai, J.F., Qi, H.Z., Xiong, Y.W., *et al*. (2017) Deformable Convolutional Networks. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, Italy, 22-29 October 2017. https://doi.org/10.1109/ICCV.2017.89

[4] Dong, S.J., Pan, Z.X., Fu, Y., *et al*. (2022) Deu-net 2.0: Enhanced deformable u-net for 3d cardiac cine mri segmentation. *Medical Image Analysis*, **78**, 102389. https://doi.org/10.1016/j.media.2022.102389

[5] Zhao, C.H., Zhu, W.X. and Feng, S. (2022) Superpixel Guided Deformable Convolution Network for Hyperspectral Image Classification. *IEEE Transactions on Image Processing*, **31**, 3838-3851. https://doi.org/10.1109/TIP.2022.3176537

[6] Jin, Q.G., Meng, Z.P., Pham, T.D., *et al*. (2019) Dunet: A Deformable Network for Retinal Vessel Segmentation. *Knowledge-Based Systems*, **178**, 149-162. https://doi.org/10.1016/j.knosys.2019.04.025

[7] Yang, X., Li, Z.q., Guo, Y.q., *et al*. (2022) DCU-net: A Deformable Convolutional Neural Network Based on Cascade U-Net for Retinal Vessel Segmentation. *Multimedia Tools and Applications*, **81**, 15593-15607. https://doi.org/10.1007/s11042-022-12418-w

[8] Wang, D., Zhang, Z., Zhao, Z.W., *et al*. (2022) Pointscatter: Point Set Representation for Tubular Structure Extraction. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., eds., *Computer Vision-ECCV* 2022, Springer, Cham. https://doi.org/10.1007/978-3-031-19803-8_22

[9] Kong, B., Wang, X., Bai, J.J., *et al*. (2020) Learning Tree-Structured Representation for 3d Coronary Artery Segmentation. *Computerized Medical Imaging and Graphics*, **80**, 101688. https://doi.org/10.1016/j.compmedimag.2019.101688

[10] Liu, Z., Lin, Y., Cao, Y., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 10-17 October 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986