Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Abstract

基础模型目前为深度学习领域大多数令人兴奋的应用提供了动力，这些模型几乎都基于 Transformer 架构及其核心注意力模块。为了解决 Transformer 在长序列上的计算效率低下问题，人们开发了许多亚二次时间架构，如线性注意力、门控卷积和递归模型以及**结构化状态空间模型（SSM）**，但它们在语言等重要模态上的表现不如注意力。我们发现，此类模型的一个关键弱点是无法进行基于内容的推理，因此做出了几项改进。首先，只需让 SSM 参数成为输入的函数，就能解决它们在离散模态方面的弱点，使模型能够根据当前标记，有选择地沿序列长度维度传播或遗忘信息。其次，尽管这种变化阻碍了高效卷积的使用，我们还是设计了一种硬件感知的并行递归模式算法。我们将这些选择性 SSM 集成到一个简化的端到端神经网络架构中，该架构没有注意力，甚至没有 MLP 块（Mamba）。Mamba 具有快速推理（吞吐量比 Transformers 高 5 倍）和序列长度线性伸缩的特点，其性能在实际数据中可提高到百万长度序列。作为通用序列模型的骨干，Mamba 在语言、音频和基因组学等多种模式中都达到了最先进的性能。在语言建模方面，无论是预训练还是下游评估，我们的 Mamba-3B 模型都优于同等规模的 Transformers，并能与两倍于其规模的 Transformers 相媲美。

结构

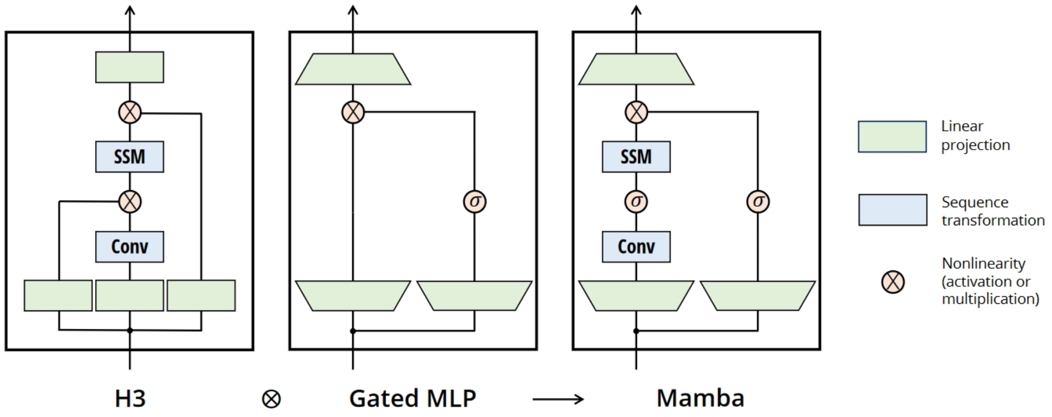

Figure 3: (**Architecture**.) Our simplified block design combines the H3 block, which is the basis of most SSM architectures, with the ubiquitous MLP block of modern neural networks. Instead of interleaving these two blocks, we simply repeat the Mamba block homogenously. Compared to the H3 block, Mamba replaces the first multiplicative gate with an activation function. Compared to the MLP block, Mamba adds an SSM to the main branch. For σ we use the SiLU / Swish activation (Hendrycks and Gimpel 2016; Ramachandran, Zoph, and Quoc V Le 2017).

U-shaped Vision Mamba for Single Image Dehazing

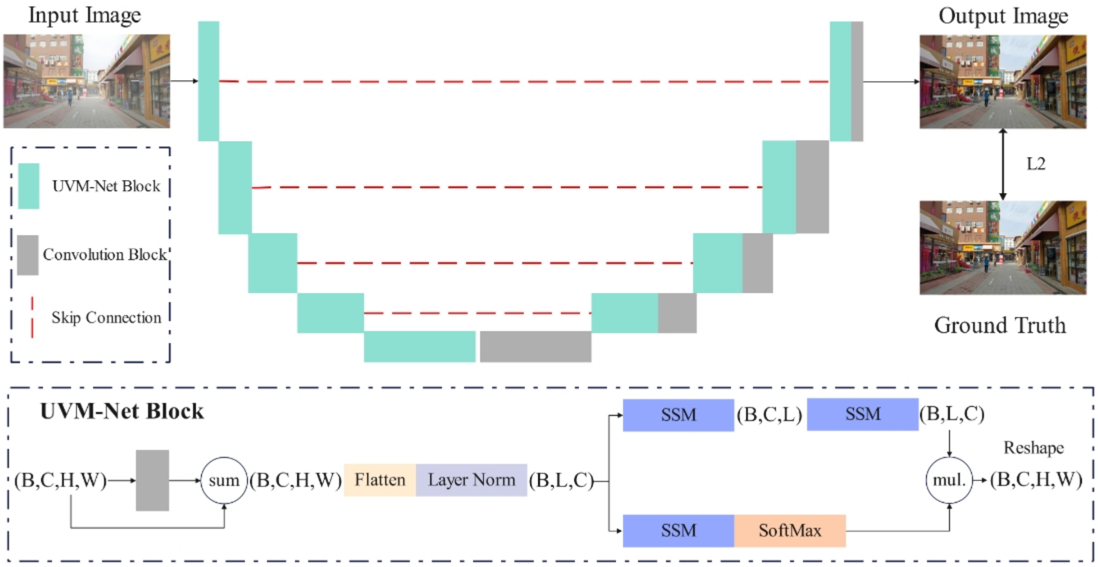

Fig. 1. Overview of the UVM-Net architecture. UVM-Net employs the encoder-decoder framework with UVM-Net blocks in the encoder and convolution blocks in the decoder, together with skip connections. In UVM-Net block, our feature maps are first applied to a convolution operation, then the unfolded pixels are modeled over SSM, and the size of the final feature is reshaped to the size of the input information.

U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation

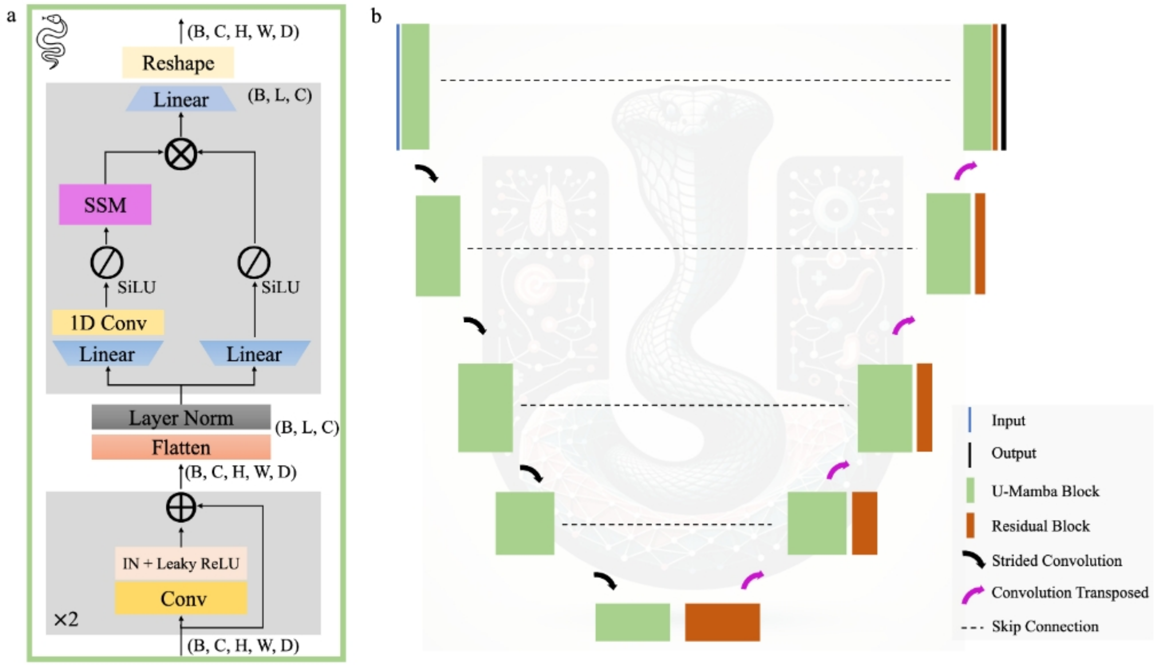

Fig. 1. Overview of the U-Mamba (Enc) architecture. **a,** U-Mamba building block contains two successive Residual blocks followed by the SSM-based Mamba block for enhanced long-range dependency modeling. **b,** U-Mamba employs the encoder-decoder framework with U-Mamba blocks in the endocer, Residual blocks in the decoder, together with skip connections. Note: This illustration serves as a conceptual representation. U-Mamba inherits the self-configuring feature from nnU-Net and the number of network blocks is automatically determined across datasets. The detailed network configurations are presented in Table 2.

Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining

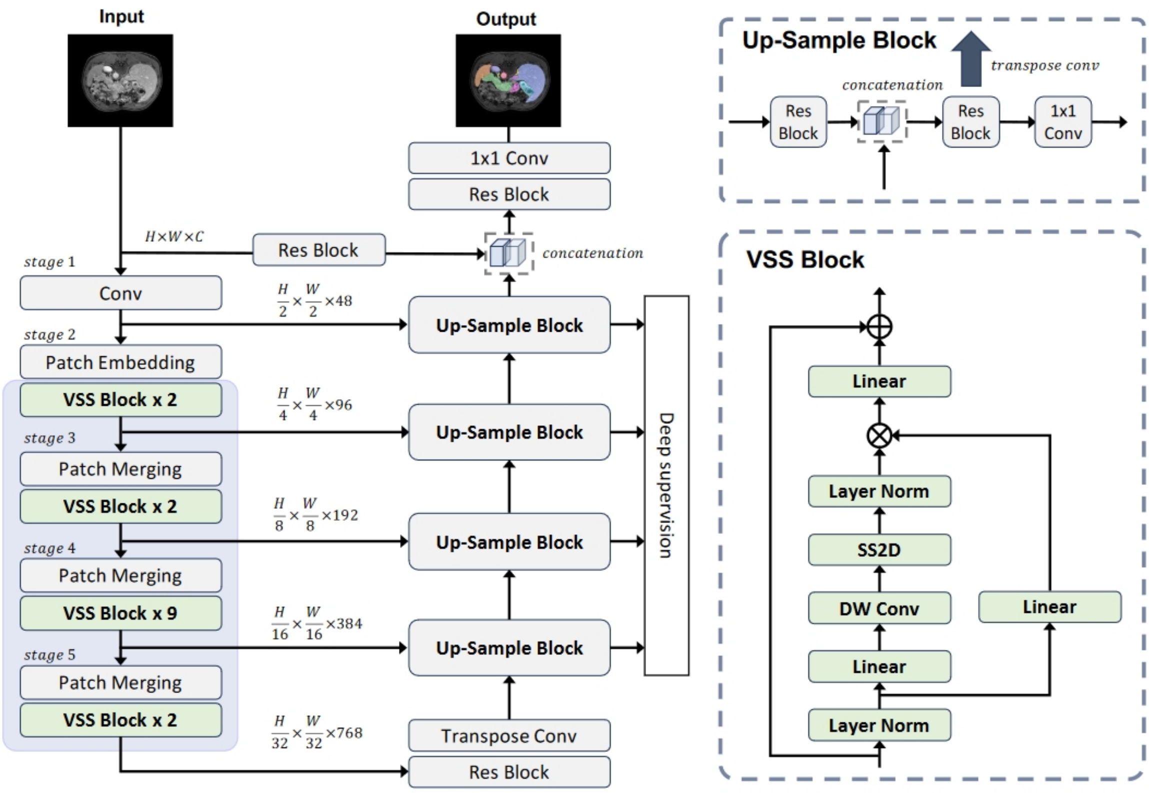

Fig. 1. The overall architecture of Swin-UMamba. Swin-UMamba can leverage the power of vision foundation models by loading the weights of pretrained models. Each block within the blue box was initialized with the ImageNet pretrained weights.

Semi-Mamba-UNet: Pixel-Level Contrastive Cross-Supervised Visual Mamba-based UNet for Semi-Supervised Medical Image Segmentation

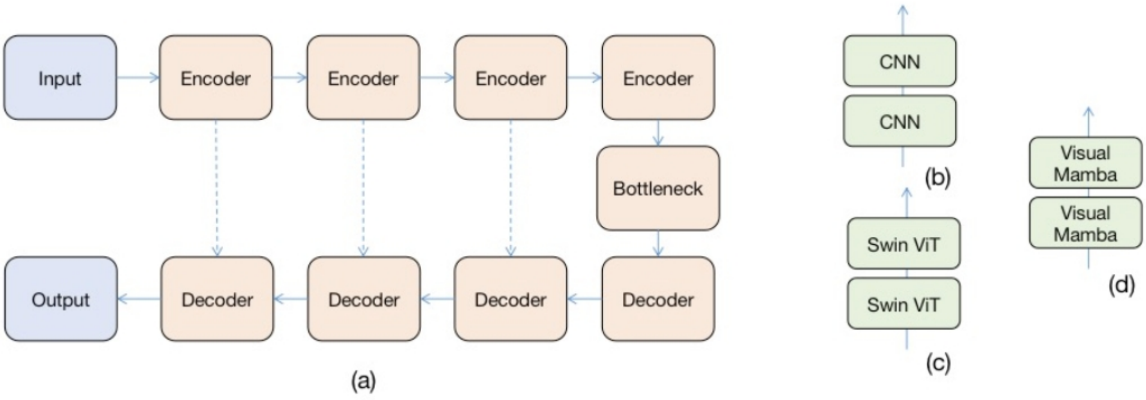

Fig. 3. The Segmentation Backbone Network in This Study. (a) Encoder-Decoder Style Segmentation Network. (b) The 2-Layer CNN-based Network Block of UNet. (c) The 2-Layer Swin ViT-based Network Block of Swin-UNet. (d) The 2-Layer Visual Mamba-based Network Block of Mamba-UNet.

SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation

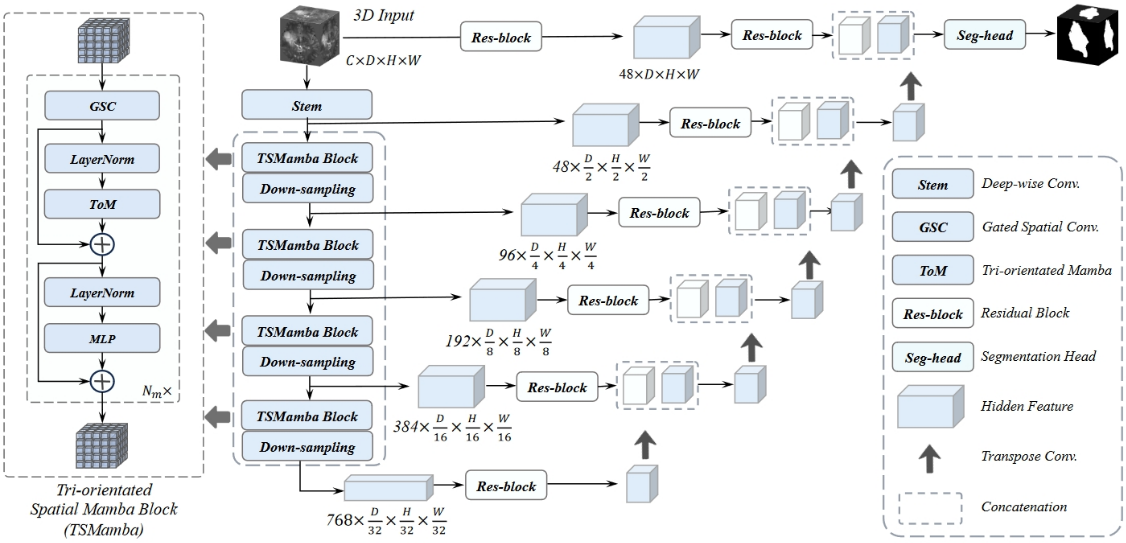
3D Input
C×D×H×W
Res-block
48×D×H×W
Seg-head
Stem
TSMamba Block
Down-sampling
$48\times\frac{D}{2}\times\frac{H}{2}\times\frac{W}{2}$
$96\times\frac{D}{4}\times\frac{H}{4}\times\frac{W}{4}$
$192\times\frac{D}{8}\times\frac{H}{8}\times\frac{W}{8}$
$384\times\frac{D}{16}\times\frac{H}{16}\times\frac{W}{16}$
$768\times\frac{D}{32}\times\frac{H}{32}\times\frac{W}{32}$
GSC
LayerNorm
ToM
MLP
$N_m\times$
Tri-orientated Spatial Mamba Block (TSMamba)
Stem Deep-wise Conv.
GSC Gated Spatial Conv.
ToM Tri-orientated Mamba
Res-block Residual Block
Seg-head Segmentation Head
Hidden Feature
Transpose Conv.
Concatenation

Fig. 2. The overview of the proposed SegMamba.

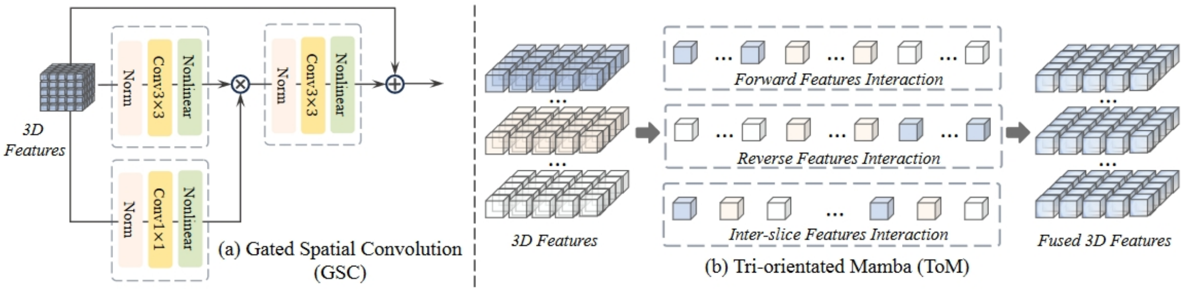
3D Features
Norm
Conv3×3
Nonlinear
Conv1×1
(a) Gated Spatial Convolution (GSC)
Forward Features Interaction
Reverse Features Interaction
Inter-slice Features Interaction
3D Features
Fused 3D Features
(b) Tri-orientated Mamba (ToM)

Fig. 3. (a) The gated spatial convolution. (b) The tri-orientated Mamba.

nnMamba: 3D Biomedical Image Segmentation, Classification and Landmark Detection with State Space Model

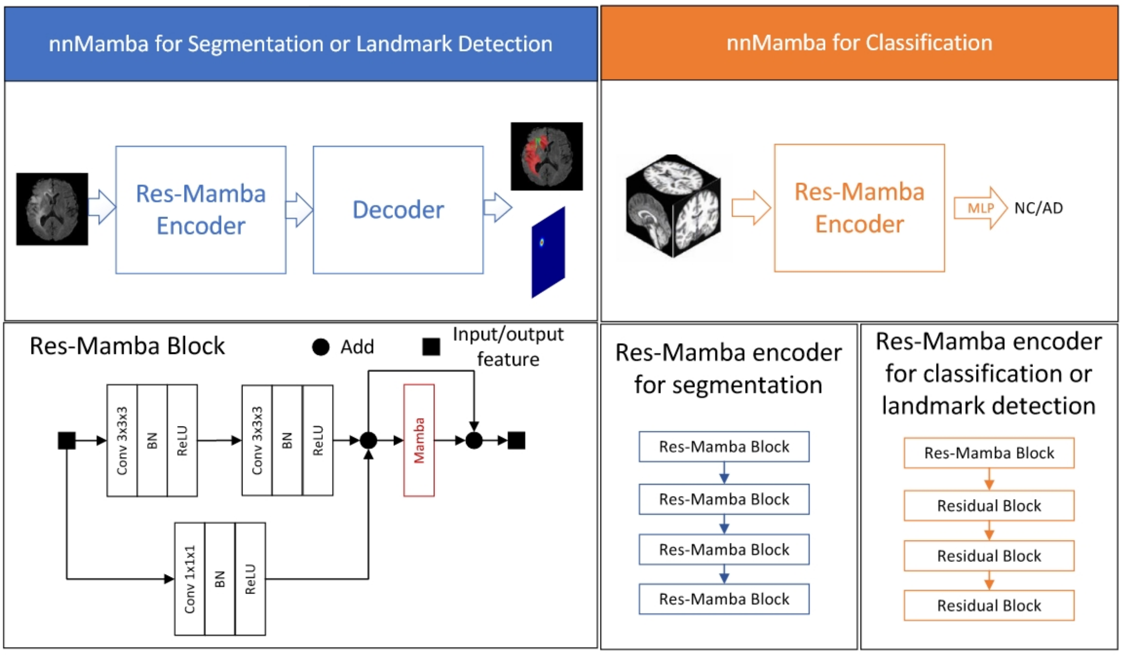

Figure 1: Our proposed nnMamba framework for 3D biomedical segmentation, classification and landmark detection. “BN” indicates the batch normliaztion operation.

Mamba-UNet: UNet-Like Pure Visual Mamba for Medical Image Segmentation

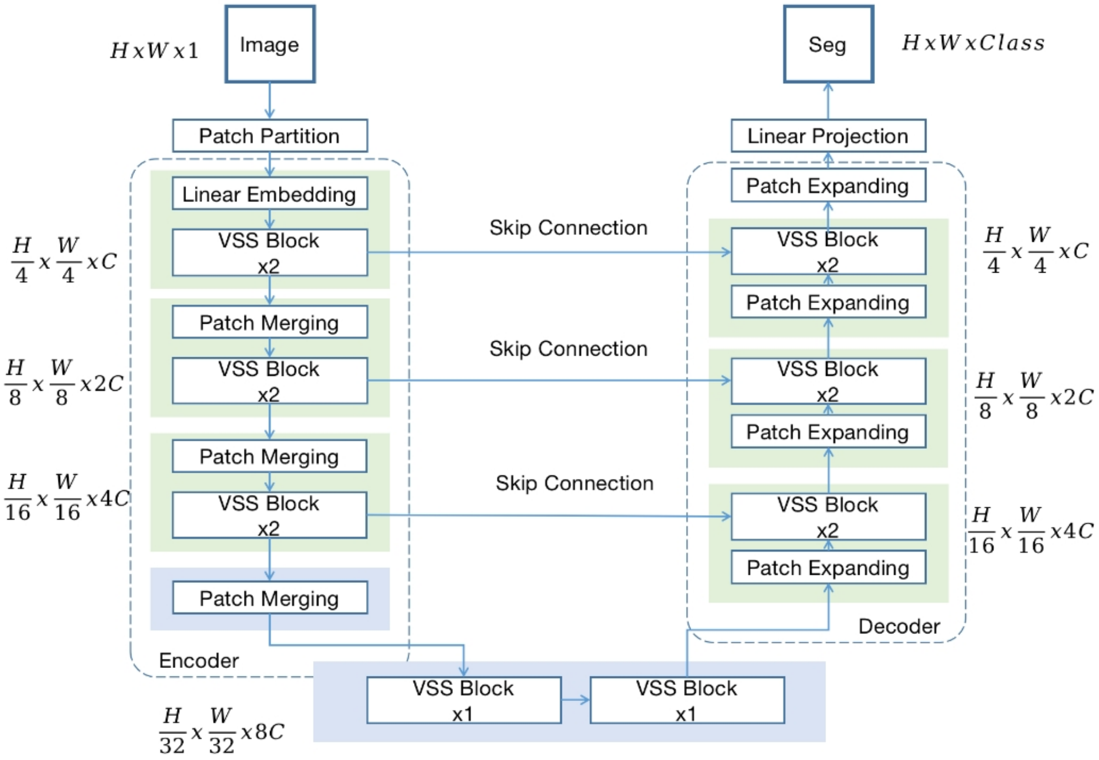

Fig. 2. The architecture of Mamba-UNet, which is composed of encoder, bottleneck, decoder and skip connections. The encoder, bottleneck and decoder are all constructed based on Visual Mamba block.

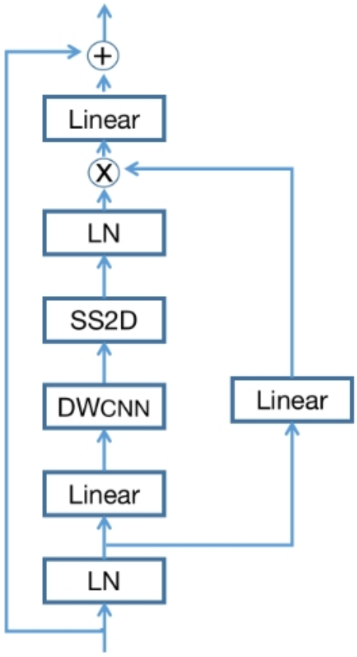

Fig. 3. The detailed structure of the Visual State Space (VSS) Block.

Weak-Mamba-UNet: Visual Mamba Makes CNN and ViT Work Better for Scribble-based Medical Image Segmentation

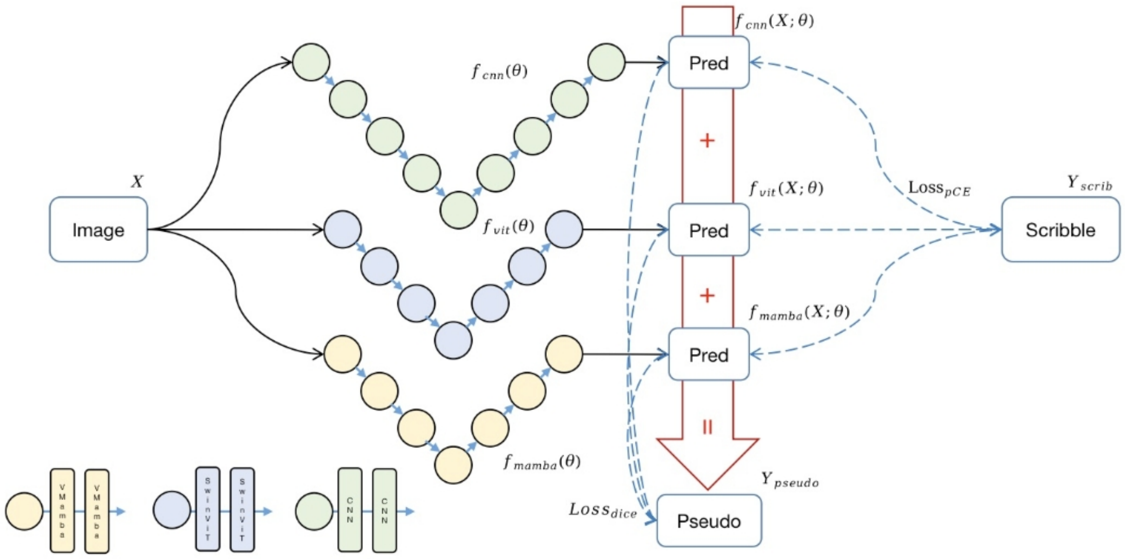

Fig. 2. Semi-Mamba-UNet: The Framework of Contrastive Cross-Supervised Visual Mamba-based UNet for Semi-Supervised Medical Image Segmentation.

VM-UNet: Vision Mamba UNet for Medical Image Segmentation

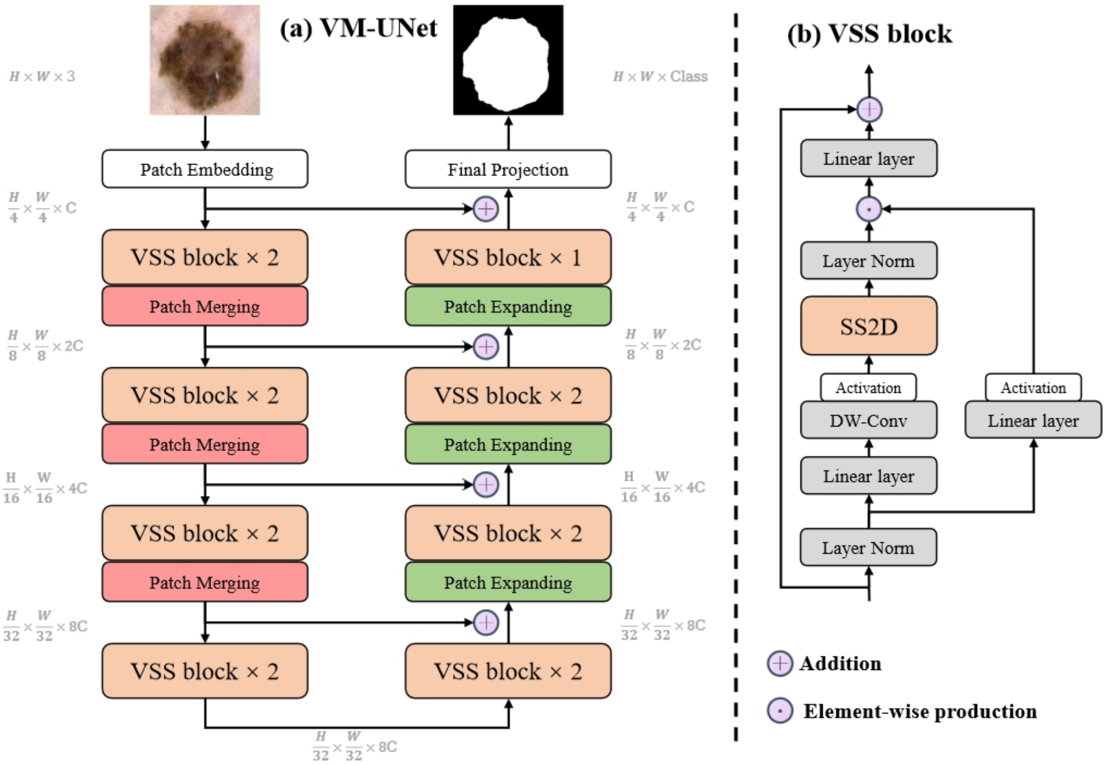

Fig. 1. (a) The overall architecture of VM-UNet. (b) VSS block is the main construction block of VM-UNet, and SS2D is the core operation in VSS block.