MambaIR: A Simple Baseline for Image Restoration with State-Space Model

摘要

近年来，图像复原技术取得了重大进展，这在很大程度上归功于现代深度神经网络（如卷积神经网络和 Transformer）的发展。然而，现有的复原网络架构往往面临着在全局感受野和高效计算之间难以抉择的困境，这阻碍了它们在实际中的应用。最近，选择性结构化状态空间模型，特别是其改进版本 Mamba，在以线性复杂度进行长距离依赖建模方面展现出了巨大潜力，这为解决上述困境提供了一种思路。然而，标准的 Mamba 在低级视觉（如局部像素丢失和通道冗余）方面仍面临一些挑战。在这项工作中，我们引入了一个简单但有效的基准模型，名为 MambaIR，它同时引入了局部增强和通道注意力机制来改进原始的 Mamba。通过这种方式，我们的 MambaIR 利用了局部像素相似性并减少了通道冗余。大量实验证明了我们方法的优越性，例如，在图像超分辨率（SR）方面，MambaIR 在计算成本相近且拥有全局感受野的情况下，比 SwinIR 的性能高出多达 0.45dB。

关键词： Mamba；图像恢复

1 引言

图像复原旨在从给定的低质量输入中重建出高质量图像，这是计算机视觉领域中一个长期存在的问题，并且它还包含超分辨率、图像去噪等众多子问题。随着卷积神经网络（[[4], [16]）和 Transformer（ [2] , [11]）等现代深度学习模型的引入，在过去几年中，最先进的性能不断被刷新。

在某种程度上，深度复原模型性能的提升很大程度上源于网络感受野的增大。首先，大的感受野使网络能够从更广泛的区域捕获信息，使其能够参考更多像素来辅助锚点像素的重建。其次，凭借更大的感受野，复原网络能够提取图像中更高级别的模式和结构，这对于一些诸如图像去噪之类的结构保持任务至关重要。最后，基于 Transformer 且具有更大感受野的复原方法在实验中表现优于基于卷积神经网络的方法，并且最近的研究也指出，激活更多像素通常会带来更好的复原结果。

尽管具有许多吸引人的特性，但对于当前的图像复原网络架构而言，在全局感受野和高效计算之间似乎存在着一种内在的选择困境。对于基于卷积神经网络的复原网络 [16]，尽管其有效感受野有限（如图 1(a) 所示），但由于卷积并行操作的高效性，它适合在资源受限的设备上进行部署。相比之下，基于 Transformer 的图像复原方法通常将标记数量设置为图像分辨率 [2] , [11]，因此，尽管具有全局感受野，但直接使用标准的 Transformer 会带来无法接受的

二次计算复杂度。此外，在图像复原中采用一些高效的注意力技术，如移位窗口注意力，通常是以牺牲全局有效感受野为代价的（如图 1(b) 所示），并且本质上并没有摆脱全局感受野和高效计算之间的权衡。

最近，结构化状态空间序列模型（S4），特别是其改进版本 Mamba，已经成为构建深度网络的一种高效且有效的网络架构 [4], [5], [7], [13], [14]。这一进展暗示了一种在图像复原中平衡全局感受野和计算效率的潜在解决方案。具体而言，Mamba 中的离散化状态空间方程可以形式化为递归形式，并且在配备专门设计的结构化重参数化 [6] 时能够对非常长距离的依赖关系进行建模。这意味着基于 Mamba 的复原网络能够自然地激活更多像素，从而提高重建质量。此外，并行扫描算法 [5] 使 Mamba 能够以并行方式处理每个标记，便于在诸如 GPU 等现代硬件上进行高效训练。上述这些有前景的特性促使我们去探索 Mamba 在实现图像复原网络的高效长距离建模方面的潜力。

然而，为自然语言处理中的一维序列数据设计的标准 Mamba [22] 并不自然适用于图像复原场景。首先，由于 Mamba 以递归方式处理展平的一维图像序列，这可能会导致在展平序列中空间上接近的像素出现在相距很远的位置，从而产生局部像素遗忘问题。其次，由于需要记忆长序列依赖关系，状态空间方程中的隐藏状态数量通常很大，这可能会导致通道冗余，从而阻碍关键通道表示的学习。

为应对上述挑战，我们引入了 MambaIR，这是一个简单但非常有效的基准模型，以使 Mamba 适应图像复原。MambaIR 由三个主要阶段构成。具体来说，1) 浅层特征提取阶段采用一个简单的卷积层来提取浅层特征。然后，2) 深层特征提取阶段由几个堆叠的残差状态空间块（RSSB）来执行。作为我们 MambaIR 的核心组件，RSSB 设计有局部卷积，以减轻将原始 Mamba 应用于二维图像时的局部像素遗忘问题，并且它还配备了通道注意力机制以减少由过多隐藏状态数量导致的通道冗余。我们还采用了可学习因子来控制每个 RSSB 内的跳跃连接。最后，3) 高质量图像重建阶段将浅层和深层特征聚合起来以生成高质量的输出图像。由于同时具备全局有效感受野和线性计算复杂度，我们的 MambaIR 可作为图像复原网络架构的一种新选择。

2 相关工作

本节回顾了三维重建的相关文献和现有方法，将它们分为传统方法和现代方法。多年来，诸如立体视觉、运动恢复结构（SfM）和多视图立体视觉（MVS）等传统方法被广泛用于从二维图像生成三维模型。这些技术通常依赖于在多个视图间匹配特征，并解决优化问题来生成三维点云或网格。然而，它们往往计算量大，并且为了保证精度需要大量的输入图像。

较新的方法利用深度学习和神经网络，例如神经辐射场（NeRF），从较少的输入图像生成高质量的三维重建。这些模型学习将三维场景表示为连续函数，这些函数可以被渲染成逼真的三维模型。尽管它们取得了成功，但计算成本和对大型训练数据集的需求仍然是显著的限制因素。

2.1 图像恢复

自从深度学习出现后，一些开创性的工作使图像复原取得了显著进展，例如用于图像超分辨率的 SRCNN [4]、用于图像去噪的 DnCNN [81]、用于减少 JPEG 压缩伪影的 ARCNN [3]

等。早期的尝试通常利用残差连接 [1]、密集连接 [16] 以及其他技术来精心设计卷积神经网络（CNN），以提高模型的表征能力。尽管取得了成功，但基于 CNN 的复原方法在有效建模全局依赖关系方面通常面临挑战。

由于 Transformer 已在多个任务中证明了其有效性，例如时间序列、三维云 [15] 和多模态，将 Transformer 用于图像复原似乎很有前景。尽管具有全局感受野，但 Transformer 仍然面临着自注意力机制带来的二次计算复杂度的特定挑战。为了解决这个问题，IPT [2] 将一幅图像分成若干小块，并使用自注意力机制独立处理每一块。SwinIR [41] 进一步引入了移位窗口注意力机制 [45] 来提升性能。此外，在为复原设计高效注意力机制方面不断取得进展 [17]。然而，高效注意力机制的设计通常是以牺牲全局感受野为代价的，高效计算和全局建模之间的权衡困境并未从根本上得到解决。

2.2 空间状态方程

源于经典控制理论的状态空间模型（SSMs）[7], [8], [14]，近期被引入到深度学习中，成为用于状态空间变换的一种颇具竞争力的网络架构。其在长距离依赖建模中能随序列长度线性缩放这一极具潜力的特性，引起了研究人员的极大兴趣。例如，结构化状态空间序列模型（S4）[7] 是深度状态空间模型在长距离依赖建模方面的开创性工作。随后，基于 S4 提出了 S5 层 [14]，该层引入了多输入多输出状态空间模型（MIMO SSM）以及高效的并行扫描机制。此外，H3 [4] 取得了令人瞩目的成果，几乎填补了状态空间模型与 Transformer 在自然语言处理方面的性能差距。[13] 通过引入门控单元进一步改进了 S4，得到了门控状态空间层，以此提升了模型的能力。最近，Mamba [5] 作为一种具备选择机制且有着高效硬件设计的依赖数据的状态空间模型，在自然语言处理方面的表现优于 Transformer，并且能随输入长度进行线性缩放。此外，也有一些开创性工作将 Mamba 应用于视觉任务，如图像分类 [12]]、视频理解 [10]、生物医学图像分割以及其他领域。在这项工作中，我们通过针对图像复原的特定设计来探索 Mamba 在图像复原方面的潜力，使其能够作为一个简单但有效的基准，为未来的相关工作提供参考。

3 本文方法

3.1 预备知识

结构化状态空间序列模型（S4）这一类模型近期取得的进展很大程度上受到连续线性时不变（LTI）系统的启发，该系统通过一个隐式潜在状态来映射一维函数或序列。形式上，该系统可表述为一个线性常微分方程（ODE）：

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) + Dx(t) \end{aligned} \tag{1}$$

其中，$h(t)$ 是状态变量，$x(t)$ 是输入变量，$y(t)$ 是输出变量，A、B、C、D 是相应的系数矩阵，t 表示时间。

之后，通常会采用离散化过程将公式（1）融入实际的深度学习算法中。具体来说，用 Δ 表示时间尺度参数，将连续参数 A、B 转换为离散参数 $\overline{A}$、$\overline{B}$。常用的离散化方法是零阶保持

（ZOH）规则，其定义如下：

$$\begin{aligned}\overline{A} &= \exp(\Delta A) \\ \overline{B} &= (\Delta A)^{-1}(\exp(A) - I) \cdot \Delta B\end{aligned} \tag{2}$$

经过离散化后，步长为 Δ 的公式（1）的离散化版本可以重写为如下的递归神经网络（RNN）形式：

$$\begin{aligned}h_k &= \overline{A}h_{k-1} + \overline{B}x_k \\ y_k &= Ch_k + Dx_k\end{aligned} \tag{3}$$

此外，公式（3）在数学上也可等价转换为如下的卷积神经网络（CNN）形式：

$$\begin{aligned}\overline{K} &\triangleq (C\overline{B}, C\overline{AB}, \cdots, C\overline{A}^{L-1}\overline{B}) \\ y &= x \circledast \overline{K}\end{aligned} \tag{4}$$

其中，L 是输入序列的长度，$\circledast$ 表示卷积运算，$\overline{K}$ 是一个结构化卷积核。

近期先进的状态空间模型——Mamba [5]，进一步将 $\overline{A}$、$\overline{B}$ 改进为依赖输入的形式，从而能够实现动态的特征表示。Mamba 用于图像复原的思路在于它对 S4 模型优势的拓展。具体而言，Mamba 与公式（3）有着相同的递归形式，这使得该模型能够记忆超长序列，以便激活更多像素来辅助复原。同时，并行扫描算法 [5] 使 Mamba 能够享有与公式（4）相同的并行处理优势，从而便于进行高效训练。

3.2 本文方法概述

如图 2 所示，我们的 MambaIR 由三个阶段构成：浅层特征提取、深层特征提取以及高质量重建。给定一个低质量（LQ）输入图像 $I_{LQ} \in \mathbb{R}^{H\times W\times 3}$，我们首先在浅层特征提取阶段使用一个 3×3 的卷积层来生成浅层特征 $F_S \in \mathbb{R}^{H\times W\times C}$，其中 H 和 W 分别代表输入图像的高度和宽度，C 是通道数量。

随后，浅层特征 F_S 进入深层特征提取阶段，以获取第 L 层（$l \in \{1, 2, \cdots, L\}$）的深层特征 $F_D^l \in \mathbb{R}^{H\times W\times C}$。该阶段由多个残差状态空间组（RSSGs）堆叠而成，每个残差状态空间组（RSSG）包含若干个残差状态空间块（RSSBs）。此外，在每个组的末尾引入了一个额外的卷积层，用于对从残差状态空间块（RSSB）中提取的特征进行细化。

最后，我们使用逐元素相加的方式来获取高质量重建阶段的输入 $F_R = F_D^L + F_S$，它被用于重建高质量（HQ）输出图像 I_{HQ}。

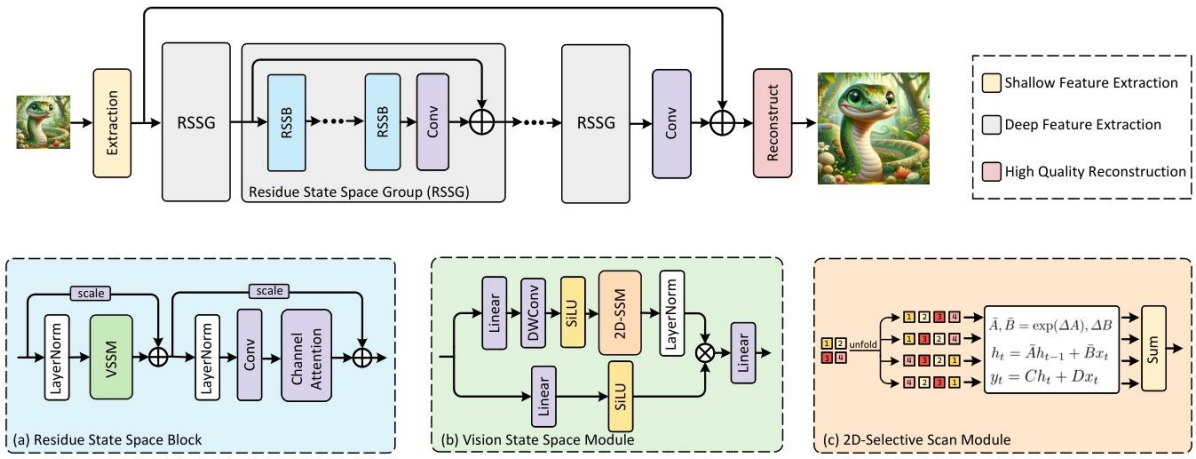

Figure 1. 我们的 MambaIR 的整体网络架构，以及（a）残差状态空间块（RSSB）、(b）视觉状态空间模块（VSSM）和（c）二维选择性扫描模块（2D - SSM）

3.3 残差状态空间块

先前基于 Transformer 的复原网络 [11], [16] 中的模块设计主要遵循“归一化（Norm）→ 注意力（Attention）→ 归一化（Norm）→ 多层感知机（MLP）”这一流程。尽管注意力机制和状态空间模型（SSM）都能够对全局依赖关系进行建模，然而，我们发现这两个模块的表现有所不同（更多细节请见补充材料），而且仅仅用状态空间模型替换注意力机制只能获得次优的结果。因此，为基于 Mamba 的复原网络定制一种全新的模块结构是很有前景的。

为此，我们提出了残差状态空间块（RSSB）。如图 **??** 所示，给定输入的深层特征 $F_D^l \in \mathbb{R}^{H\times W\times C}$，我们首先使用层归一化（LayerNorm，缩写为 LN），之后接入视觉状态空间模块（Vision State - Space Module，缩写为 VSSM [12] 来捕获空间上的长期依赖关系。此外，我们还使用可学习的缩放因子 $s \in \mathbb{R}^C$ 来控制来自跳跃连接的信息，其计算公式如下：

$$Z^l = VSSM\left(LN\left(F_D^l\right)\right) + s \cdot F_D^l$$

此外，由于状态空间模型（SSMs）将展平的特征图作为一维标记序列来处理，序列中相邻像素的数量会受到展平策略的极大影响。例如，当采用 [44] 中的四向展开策略时，对于关键像素而言只有四个最近邻像素可用（见图 2），也就是说，二维特征图中一些空间上邻近的像素在一维标记序列中彼此相距甚远，这种过远距离可能会导致局部像素遗忘问题。为此，我们在视觉状态空间模块（VSSM）之后引入了一个额外的局部卷积来帮助恢复邻域相似性。具体而言，我们先使用层归一化对 Z^l 进行归一化，然后使用卷积层来补偿局部特征。为了保持效率，卷积层采用瓶颈结构，即首先将通道数按因子 γ 进行压缩，以获得形状为 $\mathbb{R}^{H\times W\times \frac{C}{\gamma}}$ 的特征，然后进行通道扩展以恢复原始形状。

另外，状态空间模型（SSMs）通常会引入大量隐藏状态来记忆非常长距离的依赖关系，我们在图 2 中对不同通道的激活结果进行了可视化，发现存在明显的通道冗余。为了增强不同通道的表达能力，我们将通道注意力机制（Channel Attention，缩写为 CA）[9] 引入到残差状态空间块（RSSB）中。通过这种方式，状态空间模型（SSMs）能够专注于学习多样化的通道表示，之后关键通道会被后续的通道注意力机制选中，从而避免通道冗余。最后，在残差连接中使用另一个可调节的缩放因子 $s' \in \mathbb{R}^C$ 来获取残差状态空间块（RSSB）的最终输出

F_D^{l+1}。上述过程可以用以下公式表示：

$$F_D^{l+1} = CA\left(Conv\left(LN\left(Z^l\right)\right)\right) + s' \cdot Z^l$$

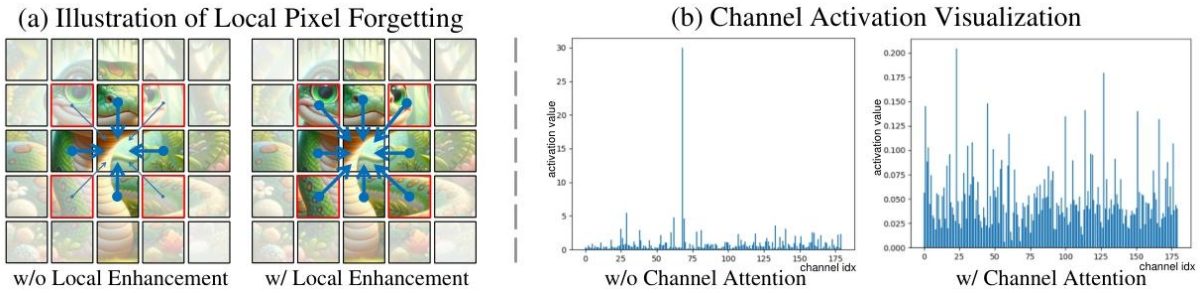

Figure 2. 我们的 MambaIR 的整体网络架构，以及（a）残差状态空间块（RSSB）、（b）视觉状态空间模块（VSSM）和（c）二维选择性扫描模块（2D - SSM）

3.4 视觉状态空间模块

为保持效率，基于 Transformer 的复原网络通常会将输入分割成小块 [8] 或者采用移位窗口注意力机制 [41]，这阻碍了全图像层面的交互。受 Mamba 在以线性复杂度进行长距离建模方面取得成功的启发，我们将视觉状态空间模块引入到图像复原中。

视觉状态空间模块（Vision State - Space Module，简称 VSSM）能够利用状态空间方程捕获长距离依赖关系，其架构如图 2(b) 所示。依照 [12] 中的做法，输入特征 X 将会经过两个并行分支。在第一个分支中，特征通道会通过一个线性层扩展到 $C \times r$（其中 r 是预先定义的通道扩展因子），随后依次经过深度可分离卷积（depth - wise convolution）、SiLU 激活函数、二维选择性扫描模块（2D - SSM layer）以及层归一化（LayerNorm）。在第二个分支中，特征通道同样会通过一个线性层扩展到 $C \times r$，然后经过 SiLU 激活函数。之后，来自这两个分支的特征通过哈达玛积（Hadamard product）进行聚合。最后，通道数量会被投影回 C，以生成与输入形状相同的输出 X_{out}，具体公式如下：

$$\begin{aligned} X_1 &= LN(2D - SSM(SiLU(DWConv(Linear(X))))) \\ X_2 &= SiLU(Linear(X)) \\ X_{out} &= Linear\left(X_1 \odot X_2\right) \end{aligned} \tag{5}$$

其中，DWConv 表示深度可分离卷积，$\odot$ 表示哈达玛积。

3.5 2D 选择扫描模块

标准的 Mamba [5] 按因果关系处理输入数据，因此只能捕获数据中已扫描部分内的信息。这一特性非常适用于具有顺序性质的自然语言处理（NLP）任务，但在迁移到诸如图像这类非因果数据时就会带来重大挑战。

为了更好地利用二维空间信息，我们参照 [12] 引入了二维选择性扫描模块（2D - Selective Scan Module，缩写为 2D - SSM）。如图 2(c) 所示，二维图像特征会沿着四个不同方向（从左上角到右下角、从右下角到左上角、从右上角到左下角以及从左下角到右上角）扫描并展平为一维序列。然后，依据离散状态空间方程来捕获每个序列的长距离依赖关系。最后，通过求和的方式合并所有序列，接着进行重塑操作以恢复二维结构。

3.6 损失函数定义

为了与先前的研究工作 [?], [17], [16] 进行公平比较，我们针对图像超分辨率（SR）任务使用 L_1 损失对我们的 MambaIR 进行优化，其计算公式如下：

$$\mathcal{L} = \|I_{HQ} - I_{LQ}\|_1$$

其中，$\|\cdot\|_1$ 表示 L_1 范数。

对于图像去噪任务，我们使用 $\epsilon = 10^{-3}$ 的 Charbonnier 损失，其计算公式如下：

$$\mathcal{L} = \sqrt{\|I_{HQ} - I_{LQ}\|^2 + \epsilon^2}$$

4 复现细节

4.1 与已有开源代码对比

在这项工作中，主要都是使用了 mambaIR 的代码进行的实验，但是在此之外还对于文本的数据进行了单独的实验。具体而言，在 basicSR 框架下设计了一个数据读取类，这个读取类是正对于 textzoom 的数据而设计的。对于 TextZOOM 来说，其高分辨率（HR）和低分辨率（LR）数据和 label 数据一同存储在 LMDB 文件中，这中存储方式和 lr 和 hr 分开单独存放是不一样的。为解决这一问题，使用了 textzoom 中对两名代表数据的处理方法，以便能准确地从这些 LMDB 文件中提取并预处理数据。这包括仔细解析文件结构，以及提取相关的高分辨率和低分辨率文本数据对，除了取出之外还要对于数据的长和宽进行处理，变为合适的长度和尺寸 128*32。

此外，我们还在 basicSR 框架内开发了一个文本识别率指标。我们利用三个预训练模型，即卷积循环神经网络（CRNN）、MROAN 以及自适应空间变换网络（ASTER）来进行文本识别。通过整合这些模型，我们能够更精准地计算文本识别率。这一指标为我们的方法在处理文本数据方面的性能提供了有价值的见解。通过这种方式，我们旨在增强文本识别在其他模型中的可复用性，并使其适用于 basicSR 框架内的每一个超分辨率（SR）模型。

4.2 实验环境搭建

主要的环境：Ubuntu 20.04，CUDA 11.7，Python 3.9，PyTorch 2.0.1 + cu118 具体的包在 requirements.txt 都有，其中一位要使用 mamba 需要 causal_conv1d,mamba_ssm 进行加速，需要单独下载适合环境的这两个包，对于图像超分这个工作来说，使用的是 DIV2K 和 Flickr2K 进行训练，使用 set5+set14+BSD100+Urban100+Mange100 实验是在 2*3090 的服务器上训练了 4 天，

4.3 创新点

对于 mamba 来说，他是个时间状态方程的基础上进行改进的结果，他在 RNN 上和 S4 上更进一步，对于 RNN 来说，他的推理很快但是他的训练因为他的线性结构导致梯度传递的缓慢，而 mamba 把所有的状态进行计算并且加入了选择性的门控机制，但是对于 mamba 来说它还未在非线性的图像领域有大的应用，mambaIR 这篇文章建立了一个 mamba 在图像恢

复领域的基础模型，使用和 VIT 相似的思路和针对 mamba 的通道冗余和线性扫描的问题进行设计，得到了一个效果很好的结果，我主要是在他的基础上使用文本图像的数据进行实验，希望可以在文本的超分辨率上使用 mamba 模型并希望取得效果，并这对这些 mamba 的实验进行探索，探究它在文本超分的特殊任务上的效果，对于文本超分来说，他不同于一般超分中主要难以解决的伪影问题，它的主要目的是把模糊的文字变得锐利清晰可辨认，方便下游任务的进行例如文本识别 OCR 等，所以来说在 SR 的侧重点上有一些不同。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，下图展示的是 mambaIR 和其他经典的图像恢复模型的对比结果。

表 1. Comparison of PSNR and SSIM values for different methods across datasets.

Method	**Scale**	**Set5**		**Set14**		**BSDS100**		**Urban100**		**Manga109**	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.30	0.9385	39.32	0.9792
HAN	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9386	39.33	0.9790
IGNN	×2	38.24	0.9613	34.07	0.9217	32.40	0.9025	33.23	0.9383	39.30	0.9788
CSNLN	×2	38.28	0.9616	34.05	0.9220	32.41	0.9027	33.34	0.9391	39.26	0.9793
NLSA	×2	38.34	0.9618	34.25	0.9230	32.44	0.9028	33.42	0.9394	39.59	0.9806
ELAN	×2	38.36	0.9620	34.30	0.9228	32.45	0.9030	33.44	0.9391	39.62	0.9793
IPT	×2	38.37	-	34.43	-	-	-	33.76	-	-	-
SwinIR	×2	38.35	0.9624	34.46	0.9250	32.53	0.9051	33.81	0.9427	39.92	0.9797
SRFormer	×2	38.51	0.9627	34.44	0.9251	32.57	0.9052	33.09	0.9449	40.07	0.9806
MambaIR	×2	**38.57**	**0.9627**	**34.67**	**0.9261**	**32.58**	**0.9048**	**34.15**	**0.9446**	**40.28**	**0.9806**
MambaIR+	×2	**38.60**	**0.9628**	**34.69**	**0.9260**	**32.60**	**0.9048**	**34.17**	**0.9443**	**40.33**	**0.9806**

表 2. Performance comparison of MambaIR variants.

Method	**Set5**		**Set14**		**Ban100**		**Urban100**		**Manga109**	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
lightSR×2	38.205	0.9611	34.015	0.921	32.333	0.902	32.966	0.935	39.277	0.978
SR×2	38.415	0.962	34.457	0.925	32.518	0.904	33.853	0.943	40.018	0.980

这是 mambaIR 的可视化结果，使用的是经典的 Set14 验证集

图 3. 可视化结果

这些是结果都表明 mambaIR 在超分上的效果是很不错的，作为一个基础模型来说，我们看表 1 可以知道，相比之前的 SRformer 之类的 transformer 模型来说，mambaIR 在 Set5 的 PSNR 上有 0.07db 的提升，而对于其他的测试集合提升也在 0.1 到 0.2 不等，，而对于轻量化的 Mamba 来说，效果也是很不错的。但是对于文本的任务来说，我跑出来的结果非常的不理想，在超分过后的图片的 PSNR 值在文本而南无中处于中间水平，而对于文本识别率来说结果并不理想，效果和比起 transformer 模型有差距，比起专门为文本是被而设计的模型效果更是有差距

6 总结与展望

实验结果可以看出来，对于传统的超分辨率领域来说，mamba 可以在 PSNR 和 SSIM 上去的不错的效果并且在参数量上比起 transformer 有优势，但是在文本超分的部分上看来，结果却并不是很好，这可能是因为对于文本超分这个任务来说，它主要的信息和要处理的部分集中在文本的部分和边缘之上，这和一般超分的平等的对待每张图片不同，而 mamba 它的线性扫描导致了一些原本 临近的像素在 mamba 中相隔非常远，虽然 mambaIR 使用了四向的扫描方式来减少这种线性扫描带来的感受野的盲区，但是对于文本任务来说还是有些不足，对于 mamba 在文本超分领域上的应用还有待探索，但是对于 mamba 来说在其他的领域还有很大的应用空间，可以进行探索，无论是在 LLM 上还是在 CV 领域中，mamba 可能还有着很多探索的领域，

参考文献

[1] Lukas Cavigelli, Pascal Hager, and Luca Benini. Cas-cnn: A deep convolutional neural network for image compression artifact suppression. In *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017.

[2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021.

[3] Chao Dong, Yubin Deng, ChenChange Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Apr 2015.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. *Learning a Deep Convolutional Network for Image Super-Resolution*, page 184 – 199. Jan 2014.

[5] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. Dec 2023.

[6] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Cornell University - arXiv,Cornell University - arXiv*, Aug 2020.

[7] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *Cornell University - arXiv,Cornell University - arXiv*, Oct 2021.

[8] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state-space layers. *Cornell University - arXiv,Cornell University - arXiv*, Oct 2021.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[10] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding, 2024.

[11] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021.

[12] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model.

[13] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. Jun 2022.

[14] JimmyT.H. Smith, Andrew Warrington, and ScottW. Linderman. Simplified state space layers for sequence modeling. Aug 2022.

[15] Yaohua Zha, Huizhen Ji, Jinmin Li, Rongsheng Li, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Towards compact 3d representations via point feature enhancement masked autoencoders.

[16] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. Oct 2022.

[17] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.