/TemplateVersion (2026.1)

面向轨迹的移位状态空间模型用于在线视频超分辨率

Qiang Zhu[1], Xiandong Meng[1], Yuxuan Jiang[2], Fan Zhang[2], David Bull[2], Shuyuan Zhu[3], Bing Zeng[3]

[1]Pengcheng Labartory, [2]University of Bristol, [3]University of Electronic Science and Technology of China
{ zhuqiang,mengxd } @pcl.ac.cn, { yuxuan.jiang,fan.zhang,dave.bull } @bristol.ac.uk, { eezsy,eezeng } @uestc.edu.cn

Abstract

在线视频超分辨率（VSR）是一项用于许多实际视频处理应用的重要技术，其目标是基于时间上之前的帧来恢复当前的高分辨率视频帧。现有的大多数在线 VSR 方法仅使用一个邻近的前一帧来实现时间对齐，这限制了视频的长距离时间建模。最近，状态空间模型（SSMs）被提出，具有线性计算复杂度和全局感受野，显著提高了计算效率和性能。在此背景下，本文提出了一种基于轨迹感知偏移 SSMs（TS-Mamba）的新型在线 VSR 方法，利用长期轨迹建模和低复杂度 Mamba 实现高效的时空信息聚合。具体来说，TS-Mamba 首先在视频中构建轨迹以从前一帧中选择最相似的标记。然后，使用由所提出的偏移 SSMs 块组成的轨迹感知偏移 Mamba 聚合（TSMA）模块来聚合所选择的标记。偏移 SSMs 块基于 Hilbert 扫描和相应的偏移操作设计，以补偿扫描损失并增强 Mamba 的空间连贯性。此外，我们提出了一种轨迹感知损失函数来监督轨迹生成，确保在训练我们的模型时标记选择的准确性。对三个广泛使用的 VSR 测试数据集进行的大量实验表明，与六个在线 VSR 基准模型相比，我们的 TS-Mamba 在大多数情况下实现了最先进的性能，并且在 MACs 中降低了超过 22.7% 的复杂度。TS-Mamba 的源代码将在 https://github.com 提供。

引言

在各种视频超分辨率（VSR）应用场景中，由于实时视频会议和直播应用的日益普及，在线 VSR 最近引起了广泛的关注（Fuoli et al. 2023; Xiao et al. 2023）。在在线 VSR 中，当前的高分辨率（HR）视频帧通常仅使用其低分辨率（LR）对应帧和之前的帧进行恢复。这受到这些在线实时应用固有的低延迟和低计算复杂度需求的限制。

在一个视频超分辨率（VSR）模型中，时间对齐或聚合是一个核心模块，用于在生成当前的高分辨率（HR）帧之前补偿来自邻近帧的信息。最近开发了先进的时间对齐或聚合模块，这些模块基于可变形卷积网络（DCN），基于光流引导的可变形对齐模型，非局部注意力机制或基于视觉 Transformer 的时空信息聚合。尽管它们提供了优越的 VSR 性能，但这些方法通常具有较高的复杂性，因此不适合在线 VSR。

为了缓解这些限制，近期的在线 VSR 方法采用了更高效的时间对齐模块，例如轻量级光流网络（Sajjadi, Vemulapalli, and Brown 2018; Xiao et al. 2023）、可变形注意力机制（Fuoli et al. 2023; Yang et al. 2023）和时间运动传播模块（Zhang et al. 2024b）。尽管它们具有高效性，这些方法主要基于卷积神经网络（CNN）使用短期时间信息——通常来自前一帧，这限制了它们进一步提高重建质量的能力。虽然引入长期时间对齐可以提升性能，但它通常会引入显著的计算开销，从而对实时或资源受限的应用构成挑战。

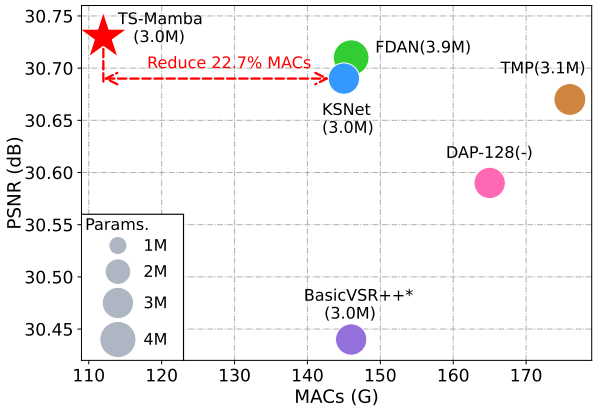


Figure 1: 在 REDS4 数据集上，我们的方法 TS-Mamba 与现有的在线 VSR 方法在 PSNR 和 MACs 方面进行了比较。我们的 TS-Mamba 优于这些 SOTA 方法，并在 MACs 方面显著降低了复杂性。

最近，低复杂度状态空间模型（SSMs）（Gu, Goel, and Ré 2021; Gu and Dao 2023）已被提出，具有线性计算复杂度和相对较大的感受野，这可能在有限复杂度下提高性能。受此启发，我们提出了一种用于在线 VSR 的轨迹感知移位 Mamba，记为 TS-Mamba，它利用长期轨迹建模和低复杂度 Mamba 来实现 token 级别的时空聚合。在 TS-Mamba 中，首先构建视频中的轨迹，以从之前的帧中选取最相似的 token。然后采用轨迹感知的移位 Mamba 聚合（TSMA）模块，该模块由移位 SSMs 块组成，用于聚合所选 token。这些移位 SSMs 块基于 Hilbert 扫描和四种不同的移位操作进行设计，以补偿扫描损失并增强 Mamba 的空间连续性。此外，我们提出了一种轨迹感知损失函数来监督轨迹生成，优化训练模型时 token 选择的准确性。所提出的 TS-Mamba 模型可以实现高效的长期视频建模，并显著减少计算复杂度。主要贡献总结如下：

- TS-Mamba 是第一个基于 SSMs 的在线 VSR 模型，它通过在 token 级别聚合来自之前帧的长期时空信息来恢复当前 HR 帧。这与现有的在线 VSR 方法不同，后者通常使用基于 CNN 的时间对齐来从单个之前帧中利用时间信息。
- 这也是首次将视频轨迹引入 Mamba，以选择前帧中最相似的令牌，并构建新的轨迹感知移位 Mamba 模型，以高效地聚合令牌级别的时空信息。
- 新颖的移位 SSM 块设计基于四种不同的移位操作和 Hilbert 扫描，以有效补偿 Hilbert 扫描的窗口内和窗口间损失，并增强 Mamba 的局部空间连续性。

所提出的方法在三个广泛使用的测试数据集上进行了基准测试，并展示了卓越的 VSR 性能，与五种最先进的在线 VSR 方法相比，在 MACs 方面的计算复杂度降低了超过 22.7 %（如图 1 所示）。

相关工作

视频超分辨率

视频超分辨率 (VSR) 是一项基础的低层次视觉任务，旨在从其低分辨率对应视频中复原高分辨率视频。现有的 VSR 方法通常是基于学习的，利用各种深度神经网络 (Teed and Deng 2020; Zhu et al. 2019; Arnab et al. 2021; Ho et al. 2022) 。例如，基于光流的方法 (Chan et al. 2021; Liu et al. 2022b) 探索帧间的时间运动以对齐它们；基于可变形卷积网络 (DCN) 的方法 (Tian et al. 2020; Wang et al. 2019) 学习帧间的运动偏移以进行特征对齐。此外，流引导的基于可变形卷积的方法 (Chan et al. 2022; Zhu et al. 2024b) 结合光流和 DCN 以实现更好的特征对齐。基于非局部注意力的方法 (Li et al. 2020; Yi et al. 2019) 汇聚全局信息进行特征聚合。基于视觉 Transformer 的方法 (Liu et al. 2022a; Tang et al. 2023) 在视频中聚合长期时空信息以恢复超分辨率帧。然而，这些方法往往与高复杂度相关，因此不太适合在线 VSR。

在线视频超分辨率

由于在线应用的特殊要求，在线视频超分辨率（VSR）方法需要轻量化和低延迟。因此，大多数现有的在线 VSR 方法已经提出了具有高效特征对齐模块的 (Fuoli et al. 2023; Sajjadi, Vemulapalli, and Brown 2018; Xiao et al. 2023) 。例如，DAP (Fuoli et al. 2023) 设计了一个可变形注意金字塔模块，以动态关注于帧间最显著的位置。FRVSR (Sajjadi, Vemulapalli, and Brown 2018) 和 CKBG (Xiao et al. 2023) 利用轻量化光流网络来估算帧间运动并进行运动补偿。KSNet (Jin et al. 2023) 提出了一种核分裂方式以在高值通道上对卷积核进行重新参数化，从而能够表征动态信息并降低通道维度上的复杂性。FDAN (Yang et al. 2023) 提出了一种流导向的可变形注意传播模块，以高效利用帧间的时间信息。TMP (Zhang et al. 2024b) 采用了一种高效的时间运动传播方法，利用运动场的连续性来实现快速特征对齐。然而需要注意的是，由于复杂性限制，这些在线 VSR 方法在特征对齐中仅基于一个前帧，这限制了 VSR 性能的进一步提高。

状态空间模型

状态空间模型 (Gu, Goel, and Ré 2021; Gu and Dao 2023)，例如 Mamba，由于其线性计算复杂性和建模全局依赖关系的能力，被广泛应用于视觉任务 (Liu et al. 2024; Zhu et al. 2024a) 。Mamba 通常通过扫描 (Qiao et al. 2024; Shi et al. 2025) 将二维图像转换为一维令牌，导致图像固有的空间连续性丧失。最近，出现了高级的扫描技术来解决该问题，如双向扫描 (Hu et al. 2024; Shi et al. 2025) 、交叉扫描 (Liu et al. 2024) 、连续二维扫描 (Yang et al. 2024) 和局部扫描 (Huang et al. 2024) 。据我们所知，Mamba 尚未被研究用于在线视频超分辨率任务。与现有基于 Mamba 的研究不同，我们引入了复杂的移位操作以增强 Hilbert 扫描，从而增强 Mamba 维持局部空间连续性的能力。

在在线视频超分辨率中，当重建低分辨率视频中的 t^{th} 帧时，我们将当前的低分辨率帧记为 I_{LR}^{t} ，并将时间上之前的低分辨率帧记为 $\{I_{LR}^{k}, k \in [t-T, t-1]\}$ 。图 2 中展示了我们提出的轨迹感知移位状态空间模型，TS-Mamba。在这里，所有这些低分辨率视频帧 $\{I_{LR}^{k}, k \in [t-T, t]\}$ 首先被输入到标记和轨迹生成模块 $\mathrm{G}(\cdot)$ 中，以提取当前帧标记 $\mathcal{Q}$ 和之前低分辨率帧的标记 $\mathcal{V}$ ：

$$\mathcal{Q} = \mathrm{G}\left(I_{LR}^{t}\right) = \left\{q_i^t\right\}, i \in [1, N], \tag{1}$$

$$\mathcal{V} = \mathrm{G}\left(\{I_{LR}^{k}\}\right) = \left\{v_i^k\right\}, i \in [1, N], k \in [t-T, t-1], \tag{2}$$

其中 $\mathrm{G}(\cdot)$ 由一个卷积层和 N_1 个残差块组成，用于从视频帧中生成特征和标记，N 是标记数，T 是时间窗口大小。

基于生成的令牌 $\{q_i^t\}$ ，I_{LR}^{t} 帧的轨迹 $\mathcal{T}^t$ 可以被表述为一组轨迹，

$$\mathcal{T}^t = \left\{\tau_i^k = \left(x_i^k, y_i^k\right)\right\}, i \in [1, N], k \in [t-T, t], \tag{3}$$

，其中 $x_i^k \in [1, H], y_i^k \in [1, W]$ ，H 和 W 分别表示特征的高度和宽度（针对 LR 帧）。每条轨迹 τ_i^k 包含一系列坐标 $\left\{\left(x_i^k, y_i^k\right), i \in [1, N]\right\}$ ，轨迹 τ_i^t 的终点与令牌 q_i^t 的坐标 (x_i^t, y_i^t) 相关联。

然后，我们沿着轨迹选择 s 个最相似的标记 $\mathcal{V}_s$ ，并将它们与标记 $\mathcal{Q}$ 一起输入到所提出的轨迹感知移动 Mamba 聚合（TSMA）模块中，以实现时空信息聚合：

$$F_{LR}^{t} = \mathrm{TSMA}(\mathcal{Q}, \mathcal{V}_s). \tag{4}$$

最后，聚合特征 F_{LR}^{t} 和当前的低分辨率帧 I_{LR}^{t} 分别被送入重建网络 $\mathrm{R}(\cdot)$ 和上采样网络 $\mathrm{U}(\cdot)$ ，以生成超分辨率帧 I_{SR}^{t} ：

$$I_{SR}^{t} = \mathrm{R}(F_{LR}^{t}) + \mathrm{U}(I_{LR}^{t}), \tag{5}$$

，其中 $\mathrm{R}(\cdot)$ 由两个卷积层、N_2 个残差块和一个像素混排层组成。这里的 $\mathrm{U}(\cdot)$ 表示双三次上采样操作。

为了在轨迹上选择最相似的符号，我们首先对与轨迹关联的符号 $\mathcal{Q}$, $\mathcal{V}$ 重新表述。基于方程 3 中轨迹的公式，符号 $\mathcal{Q}$ 和 $\mathcal{V}$ 可以被表述为：

$$\begin{aligned} &\mathcal{Q} = \{q_{\tau_i^t}\},\ i \in [1, N], \\ &\mathcal{V} = \{v_{\tau_i^k}\},\ i \in [1, N],\ k \in [t-T, t-1]. \end{aligned} \tag{6}$$

我们计算符号 $\mathcal{Q}$ 和符号 $\mathcal{V}$ 之间的余弦相似性，以选择轨迹上最相似的符号。所选符号的索引和所选符号可以被表述为：

$$\begin{aligned} &\{h_j\}_{j=1}^{s} = \underset{k}{\text{Top-k}} \langle \frac{q_{\tau_i^t}}{\| q_{\tau_i^t} \|_2^2}, \frac{v_{\tau_i^k}}{\| v_{\tau_i^k} \|_2^2} \rangle,\ h_j \in [1, T-1], \\ &\mathcal{V}_s = \{v_{\tau_i^{h_j}}\}_{j=1}^{s},\ i \in [1, N]. \end{aligned} \tag{7}$$

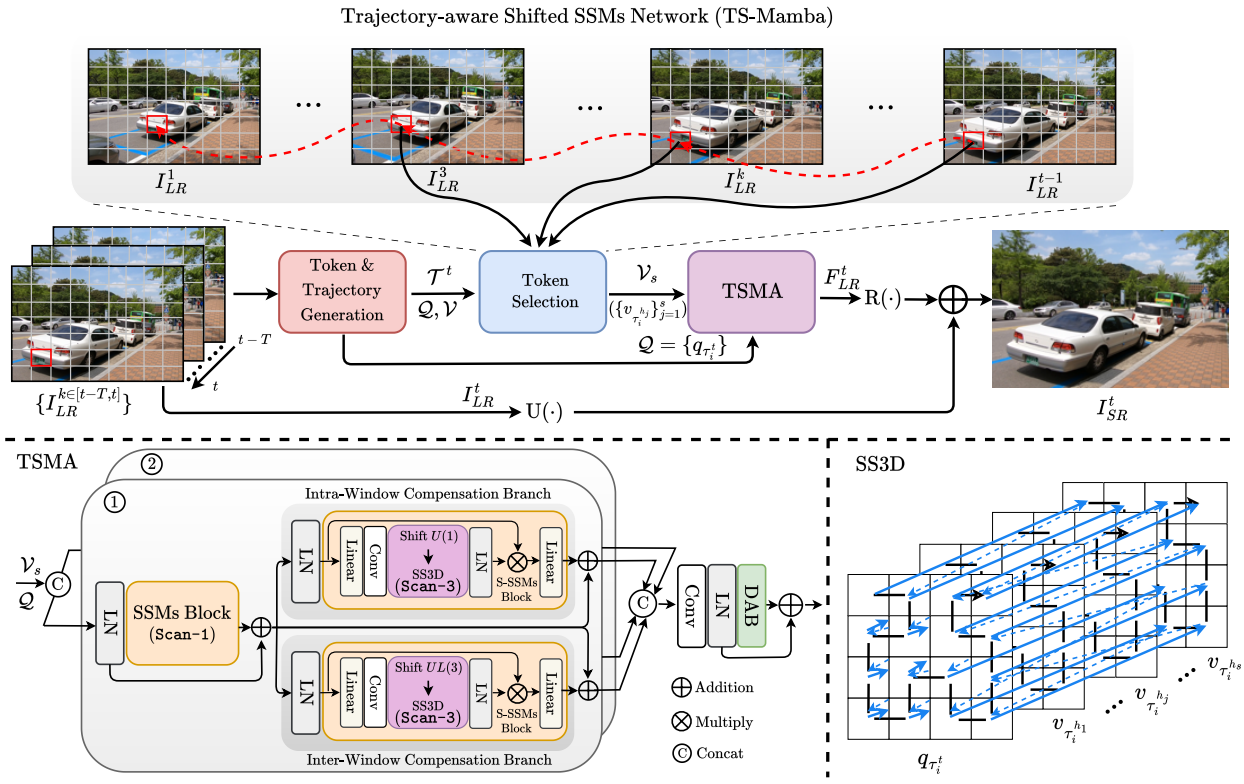
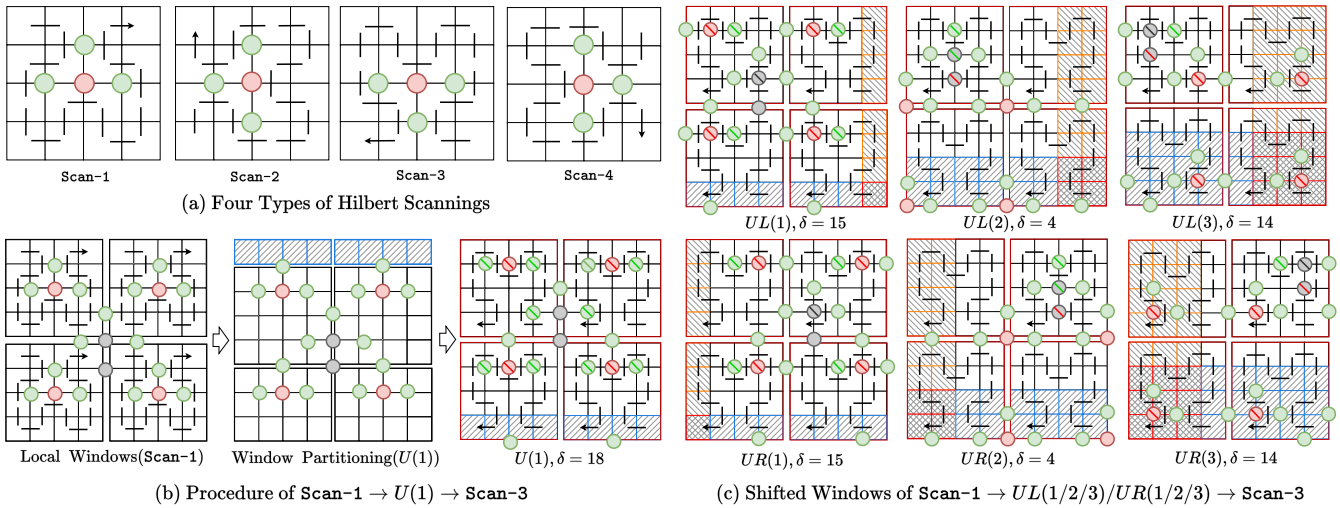


Figure 2: TS-Mamba 网络的架构。首先生成视频的轨迹，并沿轨迹选择来自前一帧的相似标记。然后，将选定的标记与当前帧标记一起输入轨迹感知移位 Mamba 聚合（TSMA）模块，以实现长期时空信息聚合。



(a) Four Types of Hilbert Scannings

(b) Procedure of Scan-1 → $U(1)$ → Scan-3

(c) Shifted Windows of Scan-1 → $UL(1/2/3)/UR(1/2/3)$ → Scan-3

Figure 3: 希尔伯特扫描和由七个过程生成的移位窗口示意图。(a) 四种类型的希尔伯特扫描。(b) Scan-1 → $U(1)$ → Scan-3 的过程和消除值 δ 。(c) 移位窗口和程序 Scan-1 → $UL(1/2/3)/UR(1/2/3)$ → Scan-3 对应的消除值 δ 。

因此，TS-Mamba 网络的过程描述如下：

$$I_{SR}^{t} = \text{TS-Mamba}(\mathcal{Q}, \mathcal{V}, \mathcal{T}^{t}) = \text{R}(\underset{\tau_i^k \in \mathcal{T}^t}{\text{TSMA}}(q_{\tau_i^t}, \{v_{\tau_i^{h_j}}\}_{j=1}^{s})) + \text{U}(I_{LR}^{t}). \quad (8)$$

考虑轨迹的移位曼巴聚合

曼巴网络通常用于通过扫描将二维图像转换为一维标记，从而导致图像固有的空间连续性损失。现有的研究 (Zhang et al. 2024a; Xiao and Wang 2025) 并未分析不连续区域的程度，而是反复使用多次扫描，使这些

方法难以维持图像的空间连续性，反而导致更大的复杂性。

为了解决这个问题，在这项工作中，我们首先分析了希尔伯特扫描中的空间不连续性，然后提出了一种轨迹感知的移动曼巴聚合（TSMA）模块，该模块以“扫描-移动-扫描”的方式结合了标准 SSMs 模块和提出的移动 SSMs（S-SSMs）模块，以补偿希尔伯特扫描的窗口内和窗口间损失。如图 2 所示，在 TSMA 模块中，令牌 $\mathcal{Q}$ 和选定的令牌 $\mathcal{V}_s$ 首先沿着通道维度连接，并以特定的“扫描-移动-扫描”方式输入两个路径，即 ① 或 ②，每个路径由一个标准 SSMs 模块和两个并行的 S-SSMs 模块组成，以根据标准 SSMs 模块的扫描补偿损失。每个路径的输出被连接后，经过卷积层和可变形注意块（DAB） (Xia et al. 2022) 进行聚合以获得输出特征。每个 SSMs/S-SSMs 模块和 DAB 之前都有层归一化（LN），之后跟随残差连接。在每个 SSMs/S-SSMs 模块中，轨迹感知令牌基于时间维度上的空间希尔伯特选择性扫描（SS3D）进行扫描，以捕获长期时空特征。

希尔伯特扫描的不连续性

为了评估希尔伯特扫描在局部窗口中的空间不连续性，我们定义不连续度 D_d 如下。如果四个相邻区域连续被扫描，则由这四个扫描区域构成的区域被认为是连续的（$D_d = 0$）；否则，不连续度 D_d 等于未连续扫描的区域数。对于由四个相邻区域组成的区域，不连续度的范围是 $D_d \in \{0, 1, 2, 3\}$ 。这在图 3 (a) 中有所说明，其中展示了四个典型的希尔伯特扫描，即 Scan-1 ，Scan-2 ，Scan-3 和 Scan-4 ，位于一个 4×4 的网格上。这里，用绿色圆圈标记的区域 $D_d = 1$，用红色圆圈标记的区域 $D_d = 2$。

此外，我们将一般情况扩展到基于 8×8 网格的情况，以进一步讨论不连续度。一个 8×8 区域被划分为四个 4×4 的局部区域，我们在图 3 (b) 中展示了在 Scan-1 下局部窗口内和窗口间的不连续度 D_d 。可以观察到，窗口内不连续性和窗口间不连续性均存在。特别地，由于 Hilbert 扫描的特性，窗口之间的中心区域间距较大（即窗口间不连续性），导致了级间间隙。在这里，不连续度 D_d 等于 3——我们在图 3 (b) 中用灰色圆圈标记了这个区域。

移位 SSMs 块

为消除 Hilbert 扫描的不连续性，我们提出了“扫描-移动-扫描”的方法，该方法结合窗口移动和特定的 Hilbert 扫描，以增强 SSM 的连续性。移动可以基于移动方向和移动位置定义，例如，上移 1 个位置（U (1)），左上移 1 个位置（UL (1)）和右下移 2 个位置（DR (2)）。我们的“扫描-移动-扫描”方法是基于四个 Hilbert 扫描（如图 3 (a) 所示）和这些窗口移动过程设计出来的。如图 3 (b) 所示，我们以 Scan-1 $\rightarrow U(1) \rightarrow$ Scan-3 为例说明其过程。局部窗口首先由 $U(1)$ 移动操作划分，然后作为移动后的窗口依次循环输入。可以推断出，在移动后的窗口上进行的第二次扫描（Scan-3 ）可以消除第一次扫描（Scan-1 ）的不连续性。

为了评估不连续消除，我们设置了三个符号并定义了一个消除值 δ 来标记和计算消除。具体来说，在图 3 (b)-(c) 中，我们在圆上使用绿色“ \ ”、红色“ \ ”和灰色“ \ ”来表示分别消除 1、2 和 3 个不连续度。消除值 δ 通过对由窗内不连续消除和窗间不连续消除组成的已消除不连续度求和来计算，即 $\delta = \delta_{\text{intra}} + \delta_{\text{inter}}$ 。我们研究了许多移位操作和扫描的组合，并展示了由六种移位操作生成的代表性移位窗口，即 UL (1)、UL (2)、UL (3) 和 UR (1)、UR (2)、UR (3)，在图 3 (c) 中的第一次扫描（Scan-1 ）和第二次扫描（Scan-3 ）下的结果。从图 3 (b)-(c) 可以推断出，在第一次扫描下，Scan-1 $\rightarrow U(1) \rightarrow$ Scan-3 的过程达到了最佳消除（δ =18），即为 Scan-1 。该过程实现了最佳的窗内不连续消除（δ_{intra} =18），但未消除窗间不连续（δ_{inter} =0）。我们可以推断出，其他三个过程也可以实现最佳消除：Scan-2 $\rightarrow L(1) \rightarrow$ Scan-4 ，Scan-3 $\rightarrow D(1) \rightarrow$Scan-1, Scan-4$\rightarrow R(1) \rightarrow$Scan-2. Moreover, the procedure of Scan-1$\rightarrow UL(3)/UR(3) \rightarrow$Scan-3 has the best inter-window discontinuity elimination (δ_{inter}=6) but worse than the procedure of Scan-1$\rightarrow U(1) \rightarrow$Scan-3 for intra-window discontinuity elimination (δ_{inter}=8). UL and UR shift operations exhibit symmetry under the first scanning Scan-1 and the second scanning Scan-3 when the same shift positions. Different combinations of shift operations and scannings can bring significant different elimination performance (more detail in our supplementary).

Based on these observations, we elaborately select shift operations and Hilbert scannings to construct two S-SSMs blocks in parallel branches, i.e., intra-window compensation branch (IntraWCB) and inter-window compensation branch (InterWCB), to eliminate corresponding discontinuities. As illustrated in Figure 2, we set two procedures for the parallel SSMs blocks to construct our TSMA module: ①: Scan-1$\rightarrow U(1)+UL(3) \rightarrow$Scan-3 ； ② ： Scan-2$\rightarrow L$ (1)+LU (3) $\rightarrow$ Scan-4 ，以达到充分消除不连续的效果。

为了实现时间上的令牌聚合，我们沿时间维度实施基于空间 Hilbert 的选择性扫描，即 SS3D。如图 2 所示，我们展示了使用 Scan-1 的 SS3D 处理。当前令牌 $\{q_{\tau_i^t}\}$ 和选择的令牌 $\{v_{\tau_i^{h_j}}\}_{j=1}^{s}$ 被扫描，以将时空相邻像素转换为一维令牌序列。每个令牌序列根据局部窗口进行选择性扫描。此过程将选择的令牌与当前令牌交织在一起，使信息能在空间和时间维度上交互，以捕捉长期的时空特征。通过扫描时空相邻的像素，SS3D 保留了局部空间信息并逐步捕捉全局时间模式。

损失函数

我们采用 Charbonnier loss (Lai et al. 2018) 作为空间损失函数来监督 SR 帧生成：

$$\mathcal{L}_{spa} = \sqrt{\|I_{HR}^t - I_{SR}^t\|^2 + \epsilon^2}, \tag{9}$$

，其中 I_{HR}^t 为 HR 帧，ϵ 设置为 1×10^{-4} 。为了监督轨迹生成以确保令牌选择的准确性，我们首先采用公式中的 LR 视频轨迹（方程 3 ）以生成 HR 视频的轨迹：

$$\mathcal{T}_{HR}^t = \left\{ \tau_{i(HR)}^k = \left(x_i^k, y_i^k\right) \right\}, i \in [1, M], k \in [t - T, t]. \tag{10}$$

在此基础上，我们提出了我们的轨迹感知损失函数：

$$\mathcal{L}_{trj} = \left\| \mathcal{T}^t - ((\mathcal{T}_{HR}^t) \downarrow_{\hat{s}}) / \hat{s} \right\|, \tag{11}$$

其中 $\downarrow_{\hat{s}}$ 为下采样操作，缩放因子为 $\hat{s}$ ，按 LR 尺度对每个 $\hat{s}$ 坐标进行子采样。

总的来说，总损失为：

$$\mathcal{L}_{total} = \mathcal{L}_{spa} + \lambda \mathcal{L}_{trj}, \tag{12}$$

其中超参数 λ 设置为 0.1。

实验

实验设置

继之前的在线视频超分辨率研究 (Jin et al. 2023; Zhang et al. 2024b)，我们使用 REDS (Nah et al. 2019) 和 Vimeo-90K (Xue et al. 2019) 作为训练数据集。REDS4 用于评价在 REDS 数据集上训练的模型，而 Vimeo-90K-T 和 Vid4 (Liu and Sun 2013) 用于基准测试在 Vimeo-90K 数据集上训练的模型。两种降质，BI（双三次插值）和模糊降质（BD），用于执行下采样，且下采样因子设定为 $\hat{s} = 4$。对于 BI 下采样，高分辨率帧通过双三次插值过滤器下采样。对于 BD 下采样，高分辨率帧首先通过标准差为 $\sigma = 1.6$ 的高斯过滤器进行模糊处理，然后每隔 $\hat{s}$ 个像素进行子采样以生成低分辨率帧。采用 PSNR 和 SSIM 作为性能评价指标。对尺寸为 180 × 320 的低分辨率帧计算运行时间 (Run.)、FPS（每秒帧数）、MACs 和参数 (Params.) 以评估模型的复杂性和速度。在实验中，残差块的数量 N_1 和 N_2 分别设置为 2 和 13。令牌大小为 4 × 4，窗口大小为 8 × 8。选定令牌数量 s 设置为 3。进行数据增强时进行随机翻转、旋转和时间反转操作。在网络训练期间使用 Adam 优化器 (Kingma 2014) 和余弦退火方案 (Loshchilov and Hutter 2016)。高分辨率补丁大小为 256 × 256，批量大小为 8。总迭代次数为 600K。该方法在 PyTorch 平台上用两块 NVIDIA GeForce RTX 3090 GPU 实现。根据 (Liu et al. 2022a)，采用轻量光流网络 (Kong, Shen, and Yang 2021) 更新轨迹。在 REDS (Nah et al. 2019) 数据集上训练时，暂时窗口大小 T 设定为 15。对于 Vimeo-90K (Xue et al. 2019) 数据集，原始序列进行时间翻转以获得一个 14 帧序列。

我们将我们的方法与五种 SOTA 在线 VSR 方法进行比较，包括 RRN (Isobe et al. 2020)、DAP-128 (Fuoli et al. 2023)、FDAN (Yang et al. 2023)、KSNet (Jin et al. 2023) 和 TMP (Zhang et al. 2024b)，以及四种双向传播 VSR 方法，BasicVSR (Chan et al. 2021)、IconVSR (Chan et al. 2021)、BasicVSR++ (Chan et al. 2022) 和 SSL (Xia et al. 2023)。此外，我们实施了另一种方法，即 “BasicVSR++*”，通过去除 BasicVSR++ 的后向传播分支并减少其模型尺寸以适应在线 VSR 应用。我们使用 “P”、“F” 和 “N” 来表示那些具有之前支持帧、未来支持帧和无支持帧的方法。

整体表现

如表 1 所示，定量结果表明，所提出的方法在 PSNR 和 SSIM 方面的性能优于其他在线 VSR 模型。我们还基于公开的预训练模型和源代码，补充了 FDAN 和 KSNet 模型在 Vid4 和 Vimeo-90K-T 数据集上的结果，以便进行全面比较。这些结果在表 1 中用 “† ” 进行了报告。图 ?? 展示了定性比较，从中我们可以观察到，我们的方法在视觉质量上比其他在线 VSR 方法在 BI 和 BD 退化方面表现得更好。

根据 (Fuoli et al. 2023; Zhang et al. 2024b)，能够处理 720p（1280 × 720）视频且帧率至少为 24 的 VSR 方法被视为实时（R-T.）方法 (Fuoli et al. 2023)，我们在表格 1 中根据运行时间标记了所有测试的方法。需要注意的是，我们的 TS-Mamba 模型在所有在线 VSR 方法中实现了第二快的推理速度。TMP 的运行时间最快，因为其使用了 CUDA 加速器（高 MACs 但运行时间低），而 TS-Mamba 未使用。此外，TS-Mamba 还在 MACs 方面显著减少（约 36.3 % ），而相较于 TMP，在参数数量上略有减少，如图 1 所示。

消融研究

为了进一步验证我们贡献的有效性，我们在 REDS4 数据集上进行了消融研究。

我们首先通过创建以下变体确认了两个轨迹感知设计的贡献，即轨迹生成和轨迹感知损失。(v1.1) 无轨迹 - 移除了 TS-Mamba 中的 G(·) 和令牌选择模块；(v1.2) 无 $\mathcal{L}_{trj}$ - 在训练 TS-Mamba 模型时移除了轨迹感知损失函数。我们进一步验证了我们提出的 TSMA 模块在补偿分支和移位操作方面的效果，通过获得 (v1.3) 无 IntraWCB 和 (v1.4) 无 InterWCB - 分别从 TSMA 模块中移除了 IntraWCB 和 InterWCB；(v1.5) 无 IntraWCB + InterWCB - 从 TSMA 模块中移除了 IntraWCB 和 InterWCB。我们还测试了补偿分支中采用的移位操作，并实现了 (v1.6) 无 $U(1)/L(1)$ - 在 IntraWCB 中移除了 $U(1)/L(1)$ 移位操作，(v1.7) 无 $UL(3)/LU(3)$ - 在 InterWCB 中移除了 $UL(3)/LU(3)$ 移位操作，以及 (v1.8) 无 (v1.6)+(v1.7) - 在 TSMA 模块中移除了所有移位操作。如表 2 所示，所有这些变体的性能显著低于完整的 TS-Mamba，这充分验证了我们设计中每个关键组件的有效性。

为了确认我们 TS-Mamba 中代币数量 s 的值，我们在 TS-Mamba 模型上测试了不同的 s 值，并在表格 3 中展示了结果。需要注意的是，随着 s 的增加，VSR 性能提高，但同时模型的复杂度也增加。当 $s = 4$ 时，很难明显提高 VSR 性能。为了在复杂度和性能之间取得平衡，我们在这项工作中设置 $s = 3$。

结论

在这篇论文中，我们提出了一种用于在线 VSR 的轨迹感知移位 SSMs（TS-Mamba），通过长期轨迹建模和低复杂度的 Mamba 来实现高效的时空信息聚合。在 TS-Mamba 中，首先构建视频中的轨迹以从先前帧中选择最相似的标记。接着使用轨迹感知移位 Mamba 聚合模块，该模块由移位 SSMs 模块组成，用于聚合选定的标记。移位 SSMs 模块的设计基于 Hilbert 扫描和移位操作，以补偿扫描损失并增强 Mamba 的空间连续性。此外，我们提出了一种轨迹感知的损失函数来监督轨迹生成，优化在训练模型时标记选择的准确性。在三个广泛使用的 VSR 基准上进行的大量实验表明了我们方法的有效性和效率。

Category	Methods	Support Frame	R-T.	Run. ↓ (ms)	FPS ↑ (1/s)	MACs ↓ (G)	Params. ↓ (M)	BI degradation		BD degradation	
								REDS4(RGB) ↑ (PSNR/SSIM)	Vid4(Y) ↑ (PSNR/SSIM)	Vimeo-90K-T(Y) ↑ (PSNR/SSIM)	Vid4((PSNR/
Bidirectional	BasicVSR (Chan et al. 2021)	P+F	✗	63	15.9	397	6.3	31.42/0.8909	27.24/0.8251	37.53/0.9498	27.96/
	IconVSR (Chan et al. 2021)	P+F	✗	70	14.3	452	8.7	31.67/0.8948	27.39/0.8279	37.84/0.9524	28.04/
	BasicVSR++ (Chan et al. 2022)	P+F	✗	77	13.0	418	7.3	32.39/0.9069	27.79/0.8400	38.21/0.9550	29.04/
	SSL-bi (Chan et al. 2021)	P+F	✗	24	41.7	92	1.0	31.06/0.8933	27.15/0.8208	37.06/0.9458	27.56/
Online VSR	Bicubic	N	✓	-	-	-	-	26.14/0.7292	23.78/0.6347	31.30/0.8687	21.80/
	RRN (Chan et al. 2021)	P	✓	34	29.4	193	3.4	28.82/0.8234	25.85/0.7660	36.69/0.9432	27.69/
	BasicVSR++*	P	✓	40	25.0	146	3.0	30.44/0.8686	27.06/0.8173	37.11/0.9464	27.49/
	DAP-128 (Fuoli et al. 2023)	P	✓	38	26.3	165	-	30.59/0.8703	-	37.29/0.9476	-
	FDAN (Yang et al. 2023)	P	✓	34	29.4	146	3.9	30.71/0.8723	27.14/0.8206 †	37.36 /0.9483 †	27.76 /
	KSNet (Jin et al. 2023)	P	✓	31	32.3	145	3.0	30.69/0.8724	27.14/0.8208	37.34/ 0.9490	27.63/
	TMP (Zhang et al. 2024b)	P	✓	25	40.1	176	3.1	30.67/0.8710	27.10/0.8167	37.33/0.9481	27.61/
	TS-Mamba (ours)	P	✓	29	33.5	112	3.0	30.73/0.8727	27.17/0.8209	37.36 /0.9482	27.70/

Table 1: 与最新的在线 VSR 方法的比较。在三个基准测试上，运行时间、帧率、参数以及 PSNR（dB）/SSIM 值在 BI 和 BD 退化条件下进行了报告。

Models	PSNR/SSIM	Params.	Run.	MACs
(v1.1) w/o Trajectory	30.45/0.8678	1.7	20	84
(v1.2) w/o $\mathcal{L}_{trj}$	30.70/0.8721	3.0	29	112
(v1.3) w/o IntraWCB	30.58/0.8702	2.8	25	97
(v1.4) w/o InterWCB	30.61/0.8706	2.8	25	97
(v1.5) w/o IntraWCB+InterWCB	30.52/0.8689	2.4	21	85
(v1.6) w/o $U(1)/D(1)$	30.65/0.8710	3.0	27	112
(v1.7) w/o $UL(3)/DL(3)$	30.67/0.8714	3.0	27	112
(v1.8) w/o (v1.6) + (v1.7)	30.61/0.8702	3.0	25	111
TS-Mamba (ours)	30.73/0.8727	3.0	29	112

Table 2: 消融研究的结果。

s	PSNR/SSIM	Params.	Run.	MACs
1	30.64/0.8712	2.8	25	96
2	30.68/0.8720	2.9	27	104
3	30.73/0.8727	3.0	29	112
4	30.74/0.8727	3.1	31	120

Table 3: 选定令牌数量的消融研究 s 。

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 6836–6846.

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. Basicvsr: The search for essential components in video super-resolution and beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4947–4956.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5972–5981.

Fuoli, D.; Danelljan, M.; Timofte, R.; and Van Gool, L. 2023. Fast online video super-resolution with deformable attention pyramid. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1735–1744.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. Advances in Neural Information Processing Systems, 35: 8633–8646.

Hu, V. T.; Baumann, S. A.; Gui, M.; Grebenkova, O.; Ma, P.; Fischer, J.; and Ommer, B. 2024. Zigma: A dit-style zigzag mamba diffusion model. In Proceedings of the European Conference on Computer Vision, 148–166. Springer.

Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024. Localmamba: Visual state space model with windowed selective scan. In Proceedings of the European Conference on Computer Vision, 12–22. Springer.

Isobe, T.; Zhu, F.; Jia, X.; and Wang, S. 2020. Revisiting temporal modeling for video super-resolution. arXiv preprint arXiv:2008.05765.

Jin, S.; Liu, M.; Yao, C.; Lin, C.; and Zhao, Y. 2023. Kernel dimension matters: To activate available kernels for real-time video super-resolution. In Proceedings of the 31st ACM International Conference on Multimedia, 8617–8625.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kong, L.; Shen, C.; and Yang, J. 2021. Fastflownet: A lightweight network for fast optical flow estimation. In 2021 IEEE International Conference on Robotics and Automation, 10310–10316. IEEE.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2018. Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(11): 2599–2613.

Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; and Jia, J. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In Proceedings of the European Conference on Computer Vision, 335–351. Springer.

Liu, C.; and Sun, D. 2013. On Bayesian adaptive video super resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(2): 346–360.

Liu, C.; Yang, H.; Fu, J.; and Qian, X. 2022a. Learning trajectory-aware transformer for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5687–5696.

Liu, M.; Jin, S.; Yao, C.; Lin, C.; and Zhao, Y. 2022b. Temporal consistency learning of inter-frames for video super-resolution. IEEE Transactions on Circuits and Systems for Video Technology, 33(4): 1507–1520.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. Advances in Neural Information Processing Systems, 37: 103031–103063.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; and Mu Lee, K. 2019. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops, 1996–2005.

Qiao, Y.; Yu, Z.; Guo, L.; Chen, S.; Zhao, Z.; Sun, M.; Wu, Q.; and Liu, J. 2024. Vl-mamba: Exploring state space models for multimodal learning. arXiv preprint arXiv:2403.13600.

Sajjadi, M. S.; Vemulapalli, R.; and Brown, M. 2018. Frame-recurrent video super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6626–6634.

Shi, Y.; Xia, B.; Jin, X.; Wang, X.; Zhao, T.; Xia, X.; Xiao, X.; and Yang, W. 2025. Vmambair: Visual state space model for image restoration. IEEE Transactions on Circuits and Systems for Video Technology.

Tang, J.; Lu, C.; Liu, Z.; Li, J.; Dai, H.; and Ding, Y. 2023. CTVSR: Collaborative spatial-temporal transformer for video super-resolution. IEEE Transactions on Circuits and Systems for Video Technology.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In Proceedings of the European Conference on Computer Vision, 402–419. Springer.

Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3360–3369.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops, 1954–1963.

Xia, B.; He, J.; Zhang, Y.; Wang, Y.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Structured sparsity learning for efficient video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22638–22647.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 4794–4803.

Xiao, J.; Jiang, X.; Zheng, N.; Yang, H.; Yang, Y.; Yang, Y.; Li, D.; and Lam, K.-M. 2023. Online video super-resolution with convolutional kernel bypass grafts. IEEE Transactions on Multimedia, 25: 8972–8987.

Xiao, Z.; and Wang, X. 2025. Event-based Video Super-Resolution via State Space Models. In Proceedings of the Computer Vision and Pattern Recognition Conference, 12564–12574.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. International Journal of Computer Vision, 127: 1106–1125.

Yang, C.; Chen, Z.; Espinosa, M.; Ericsson, L.; Wang, Z.; Liu, J.; and Crowley, E. J. 2024. Plainmamba: Improving non-hierarchical mamba in visual recognition. arXiv preprint arXiv:2403.17695.

Yang, X.; Zhang, X.; Zhang, L.; and Zhang, L. 2023. Flow-guided deformable attention network for fast online video super-resolution. In 2023 IEEE International Conference on Image Processing, 390–394. IEEE.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; and Ma, J. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 3106–3115.

Zhang, G.; Liu, C.; Cui, Y.; Zhao, X.; Ma, K.; and Wang, L. 2024a. Vfimamba: Video frame interpolation with state space models. Advances in Neural Information Processing Systems, 37: 107225–107248.

Zhang, Z.; Li, R.; Guo, S.; Cao, Y.; and Zhang, L. 2024b. Tmp: Temporal motion propagation for online video super-resolution. IEEE Transactions on Image Processing.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024a. Vision mamba: Efficient visual representation learning with bidirectional state space model. In International Conference on Machine Learning, 62429–62442. PMLR.

Zhu, Q.; Chen, F.; Zhu, S.; Liu, Y.; Zhou, X.; Xiong, R.; and Zeng, B. 2024b. DVSRNet: Deep video super-resolution based on progressive deformable alignment and temporal-sparse enhancement. IEEE Transactions on Neural Networks and Learning Systems.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9308–9316.