Mamba Architectures for Video Frame Interpolation

2025년도 하계 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
Haeuk Lee

Outline

- VFIMamba: Video Frame Interpolation with State Space Models[1)]
  - NeurIPS 2024
- LC-Mamba: Local and Continuous Mamba with Shifted Windows for Frame Interpolation[2)]
  - CVPR 2025

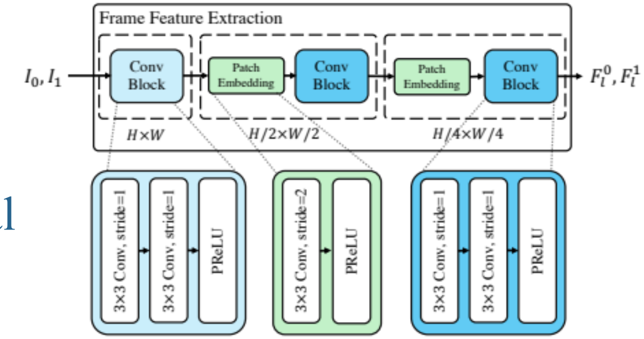
- VFIMamba: Video Frame Interpolation with State Space Models[1)]
    - NeurIPS 2024

1) ZHANG, Guozhen, et al. Vfimamba: Video frame interpolation with state space models. *Advances in Neural Information Processing Systems*, 2024, 37: 107225-107248.

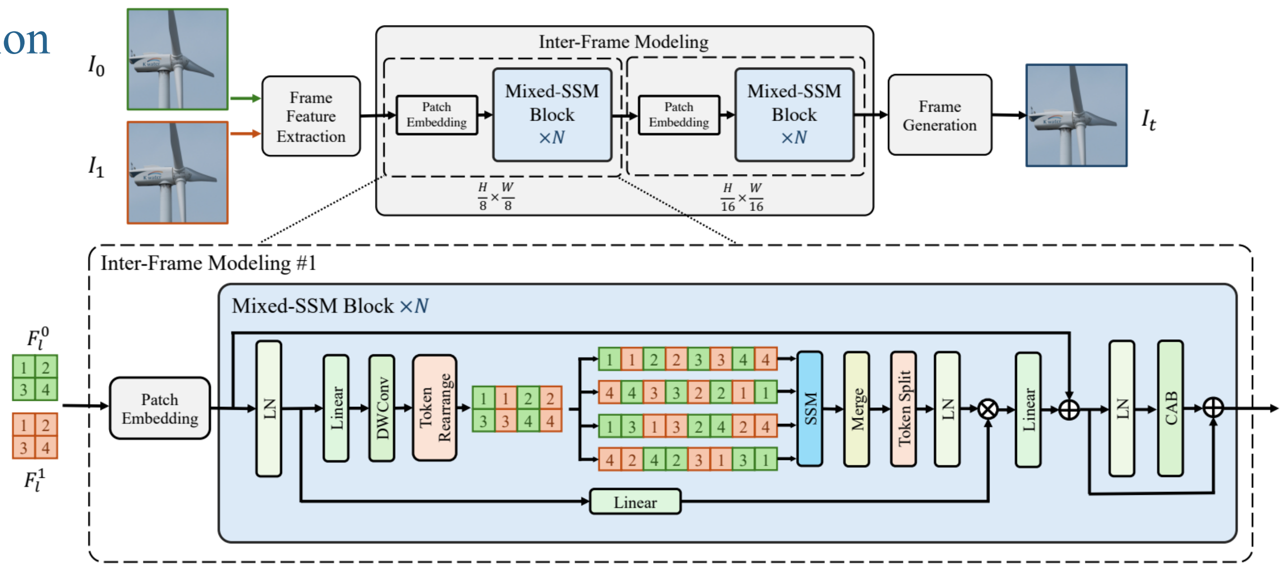
Introduction

- Video Frame Interpolation (VFI) generates intermediate frames for slow-motion and high-refresh-rate displays
- CNN-based methods lack global context
  - attention models have quadratic complexity
- S6 (Mamba) Structured State Space Models offer global receptive fields with linear complexity
- VFIMamba's Mixed-SSM Block applies multi-directional S6 scanning on interleaved frame tokens
- Curriculum learning strategy boosts performance

Overall Architecture



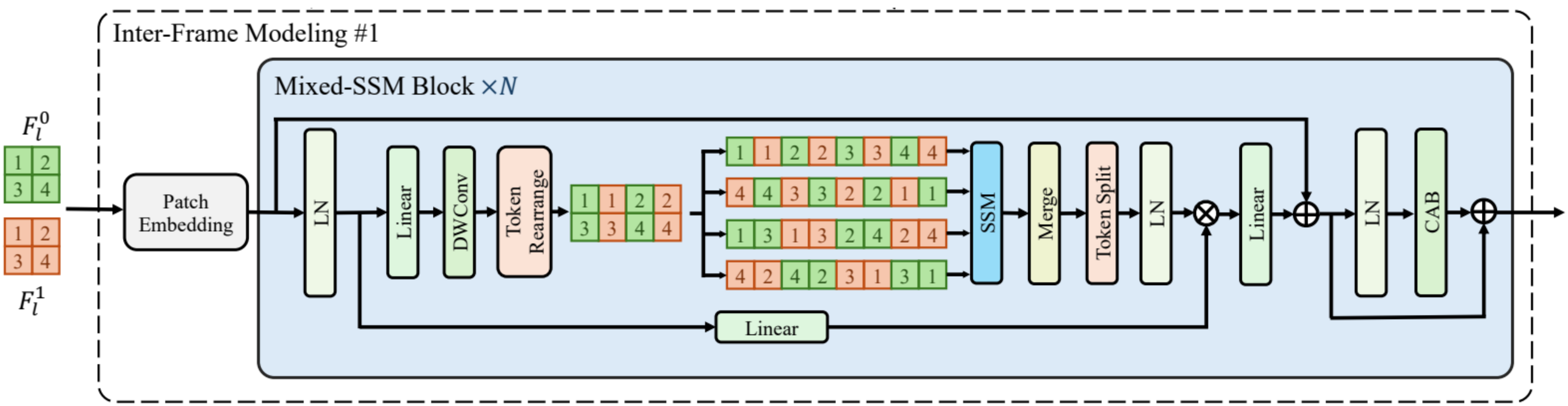
- Frame Feature Extraction
  - Input frames are processed with lightweight convolutional layers to extract shallow features at reduced resolutions
- Inter-Frame Modeling
  - Multi-scale inter-frame modeling is performed using the proposed Mixed-SSM Block (MSB) at each scale
- Frame Generation
  - Extracted inter-frame features are used for motion estimation and intermediate frame generation



1) ZHANG, Guozhen, et al. Vfimamba: Video frame interpolation with state space models. *Advances in Neural Information Processing Systems*, 2024, 37: 107225-107248.
2) GU, Albert; DAO, Tri. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

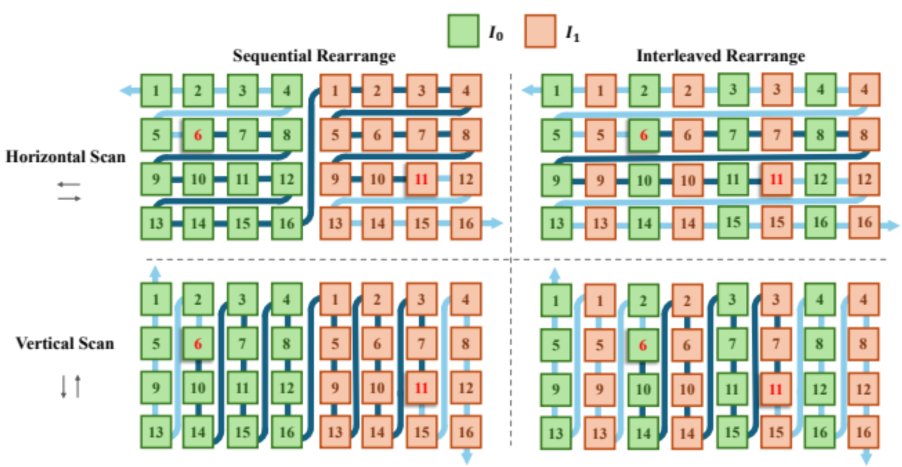
Mixed-SSM Block (MSB)

- Key module for inter-frame modeling based on the Selective State Space Model[2)] (S6, a.k.a. Mamba)
- Two main differences from Transformer blocks
  - Replaces Attention with enhanced S6 block for global, linear-complexity modeling
  - Channel-Attention Block (CAB) replaces MLP, promoting inter-channel interaction and local awareness
- The S6 block enables data-dependent, global information propagation with linear complexity



Frame Rearrangement & Scan Directions

- Interleaved Rearrangement
  - Tokens from two input frames are interleaved to form a "super image" for better spatiotemporal locality
  - This arrangement is better than simple concatenation for preserving neighborhood information
- Multi-directional Scanning
  - The super image is scanned in four directions (horizontal, vertical, and their reverses) using the S6 block
  - Each direction is processed independently, and results are merged back



Visualizations of different rearrangement methods and scan directions

1) ZHANG, Guozhen, et al. Vfimamba: Video frame interpolation with state space models. *Advances in Neural Information Processing Systems*, 2024, 37: 107225-107248.
2) BENGIO, Yoshua, et al. Curriculum learning. In: *Proceedings of the 26th annual international conference on machine learning*. 2009. p. 41-48.

Curriculum Learning[2)] Strategy

- Novel training scheme to fully unleash S6's potential for a wide range of motions
  - Start training on Vimeo90K (small motions), then gradually introduce large motion data from X-TRAIN
  - Start with Vimeo90K (small motions), then progressively introduce X-TRAIN samples
    - Resizing X-TRAIN frames (originally 512 × 512) to S × S and center-cropping to 256 × 256.
    - Every 50 epochs, increase S by 10% (starting from 256) and double the temporal interval between selected frames (starting from 1) to steadily ramp up motion magnitude
  - Effect
    - Maintains performance on small motion (low-res) while boosting high-motion (high-res) capabilities

Vimeo90K Dataset

X4K1000FPS dataset (X-TRAIN)

Experiminents

- Evaluation Datasets
  - Evaluated on diverse VFI benchmarks
    - Low-resolution: Vimeo-90K, UCF101, SNU-FILM (easy/medium/hard/extreme)
    - High-resolution: X-TEST, X-TEST-L, Xiph (all tested at 2K & 4K resolutions)

Vimeo90K

UCF101

SNU-FILM

X4K1000FPS (X-TRAIN, X-TEST-L)

Xiph

Experiminents

- Quantitative Results
  - State-of-the-art (SOTA) performance across most datasets and resolutions
  - Significant improvement in high-resolution and large-motion scenarios (e.g., X-TEST 2K/4K)
  - Comparable or better FLOPs and runtime versus existing efficient models

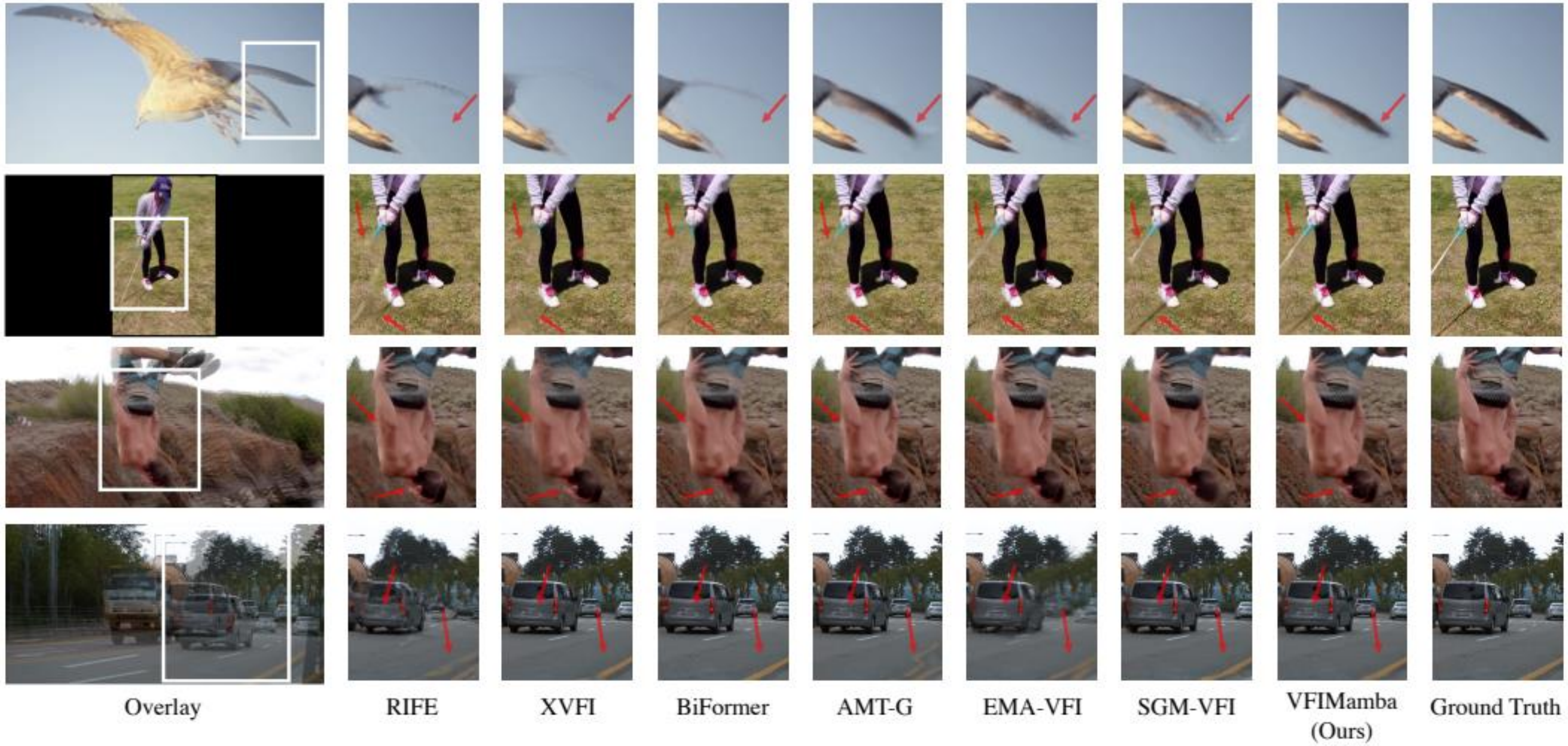
	Training Dataset	Vimeo-90K (Xue et al., 2019)	UCF101 (Soomro et al., 2012)	SNU-FILM (Choi et al., 2020) easy	medium	hard	extreme	Average	FLOPs (T)	Runtime (ms)
DAIN⋆ (Bao et al., 2019)	V	34.71/0.9756	34.99/0.9683	39.73/0.9902	35.46/0.9780	30.17/0.9335	25.09/0.8584	33.36/0.9507	5.51	897.8
AdaCof⋆ (Lee et al., 2020)	V	34.47/0.9730	34.90/0.9680	39.80/0.9900	35.05/0.9754	29.46/0.9244	24.31/0.8439	33.00/0.9458	0.36	85.1
CAIN⋆ (Choi et al., 2020)	V	34.65/0.9730	34.91/0.9690	39.89/0.9900	35.61/0.9776	29.90/0.9292	24.78/0.8507	33.29/0.9483	1.29	102.4
Softsplat (Niklaus & Liu, 2020)	V	36.13/0.9805	35.39/0.9697	40.26/0.9911	36.07/0.9798	30.53/0.9365	25.16/0.8604	33.92/0.9530	0.94	266.4
XVFI (Sim et al., 2021)	V	35.09/0.9759	35.17/0.9685	39.93/0.9907	35.37/0.9782	29.58/0.9276	24.17/0.8450	33.22/0.9477	0.37	165.2
M2M-VFI (Hu et al., 2022)	V	35.47/0.9778	35.28/0.9694	39.66/0.9904	35.74/0.9794	30.30/0.9360	25.08/0.8604	33.59/0.9522	0.26	60.9
RIFE (Huang et al., 2022)	V	35.61/0.9779	35.28/0.9690	39.80/0.9903	35.76/0.9787	30.36/0.9351	25.27/0.8601	33.68/0.9519	0.20	35.2
IFRNet-L (Kong et al., 2022)	V	36.20/0.9808	35.42/0.9698	40.10/0.9906	36.12/0.9797	30.63/0.9368	25.26/0.8609	33.96/0.9531	0.79	115.3
EMA-VFI-S (Zhang et al., 2023)	V	36.07/0.9797	35.34/0.9696	39.81/0.9906	35.88/0.9795	30.69/0.9375	25.47/0.8632	33.88/0.9534	0.20	76.4
EMA-VFI (Zhang et al., 2023)	V	**36.64/0.9819**	**35.48**/0.9701	39.98/0.9910	36.09/0.9801	30.94/0.9392	25.69/0.8661	34.14/0.9547	0.91	239.6
AMT-L (Li et al., 2023)	V	36.35/0.9815	35.39/0.9698	39.95/**0.9913**	36.09/**0.9805**	30.75/0.9384	25.41/0.8638	33.99/0.9542	0.58	183.42
AMT-G (Li et al., 2023)	V	36.53/0.9817	35.41/0.9699	39.88/**0.9913**	36.12/**0.9805**	30.78/0.9385	25.43/0.8644	34.03/0.9544	2.07	403.7
SGM-VFI (Liu et al., 2024a)	V+X	35.81/0.9793	35.34/0.9693	40.14/0.9907	36.06/0.9795	30.81/0.9375	25.59/0.8646	33.96/0.9535	1.78	942.9
VFIMamba-S	V+X	36.09/0.9800	35.36/0.9696	40.21/0.9909	36.17/0.9800	30.80/0.9381	25.59/0.8655	34.04/0.9540	0.24	128.0
VFIMamba	V+X	**36.64/0.9819**	35.45/**0.9702**	**40.51**/0.9912	**36.40/0.9805**	**30.99/0.9401**	**25.79/0.8682**	**34.30/0.9554**	0.94	310.9

	Training Dataset	X-TEST (Sim et al., 2021) 2K	4K	X-TEST-L (Liu et al., 2024a) 2K	4K	Xiph (Montgomery, 1994) 2K	4K	Average
XVFI (Sim et al., 2021)	X	31.15/0.9144	30.12/0.9045	29.82/0.8951	29.02/0.8866	34.76/0.9258	32.84/0.8810	31.29/0.9012
M2M-VFI (Hu et al., 2022)	V	32.13/0.9258	30.89/0.9138	30.90/0.9092	29.73/0.9001	36.44/0.9427	33.92/0.8992	32.34/0.9151
RIFE (Huang et al., 2022)	V	31.10/0.8972	30.13/0.8927	29.87/0.8805	28.98/0.8756	36.19/0.9380	33.76/0.8940	31.67/0.8963
FILM (Reda et al., 2022)	V	31.61/0.9174	OOM	30.18/0.8960	OOM	36.32/0.9343	33.27/0.8760	/
IFRNet-L (Kong et al., 2022)	V	31.78/0.9147	30.66/0.9050	30.76/0.8963	29.74/0.8884	36.21/0.9374	34.25/0.8946	32.23/0.9061
FLDR (Nottebaum et al., 2022)	X	31.12/0.9092	30.46/0.9041	29.90/0.8906	29.30/0.8879	34.80/0.9280	33.00/0.8862	31.43/0.9010
BiFormer (Park et al., 2023)	V+X	31.32/0.9200	31.32/0.9215	30.36/0.9068	30.14/0.9069	34.20/0.9246	33.49/0.8953	31.81/0.9125
EMA-VFI-S (Zhang et al., 2023)	V	30.91/0.9000	29.91/0.8951	29.51/0.8775	28.60/0.8733	36.55/0.9421	34.25/0.9020	31.62/0.8983
AMT-L (Li et al., 2023)	V	32.08/0.9277	30.96/0.9147	31.09/0.9103	30.12/0.9019	36.27/0.9402	34.49/0.9030	32.50/0.9163
AMT-G (Li et al., 2023)	V	32.35/0.9300	31.12/0.9157	31.35/0.9125	30.33/0.9036	36.38/0.9410	**34.63**/0.9039	32.69/0.9178
SGM-VFI (Liu et al., 2024a)	V+X	32.38/0.9272	31.35/0.9179	30.99/0.9072	29.91/0.8972	36.57/0.9424	34.23/0.9021	32.57/0.9157
VFIMamba-S	V+X	32.84/0.9328	31.73/0.9238	31.58/0.9169	30.50/0.9077	36.72/0.9428	34.32/0.9034	32.95/0.9212
VFIMamba	V+X	**33.34/0.9361**	**32.15/0.9246**	**32.22/0.9259**	**31.05/0.9159**	**37.13**/0.9451	34.62/**0.9059**	**33.42/0.9256**

Quantitative comparison with SOTA methods on the low(top)/high(bottom)-resolution datasets, in terms of PSNR/SSIM

Experiminents

- Qualitative Results
  - Sharper details and better motion estimation than previous SOTA methods, especially for large and complex motions
  - Clear preservation of object boundaries and textures



Visualizations from SNU-FILM and X-TEST

Experiminents

- Ablation Studies
  - S6 Block Effectiveness
    - Removing S6 or replacing with convolution/attention leads to noticeable performance drops
    - S6 achieves a strong balance of speed and accuracy
  - Frame Rearrangement
    - Interleaved rearrangement consistently outperforms sequential rearrangement for VFI

Model	Vimeo90K	X-TEST		SNU-FILM		Params (M)	720p Inference Time (ms)
		2K	4K	hard	extreme		
w/o S6	35.62/0.9771	28.94/0.8517	27.12/0.8436	30.41/0.9341	25.14/0.8567	16.1	**51**
Convolution	35.86/0.9790	31.58/0.9167	30.24/0.9044	30.61/0.9365	25.49/0.8631	23.4	55
Local Attention	35.92/0.9790	30.49/0.8917	30.00/0.8845	30.47/0.9338	25.46/0.8625	**15.6**	59
Full Attention	36.04/0.9798	OOM	OOM	30.55/0.9367	25.35/0.8602	**15.6**	336
S6	**36.12/0.9802**	**32.84/0.9328**	**31.73/0.9238**	**30.80/0.9381**	**25.59/0.8655**	16.8	77

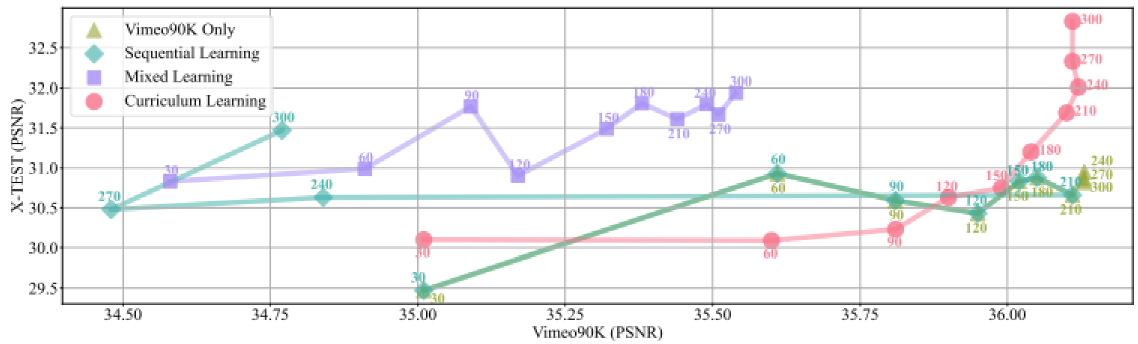
Horizontal Scan	Vertical Scan	Vimeo-90K	X-TEST		SNU-FILM	
			2K	4K	hard	extreme
Sequential	Sequential	35.55/0.9765	28.07/0.8327	26.75/0.8327	30.24/0.9319	25.03/0.8545
Sequential	Interleaved	35.76/0.9784	31.69/0.9226	30.45/0.9078	30.32/0.9342	25.21/0.8611
Interleaved	Sequential	35.79/0.9785	31.49/0.9221	30.35/0.9053	30.12/0.9331	25.11/0.8602
Interleaved	Interleaved	**36.12/0.9802**	**32.84/0.9328**	**31.73/0.9238**	**30.80/0.9381**	**25.59/0.8655**

Ablation on different models for inter-frame modeling(top) and different rearrangement methods(bottom)

1) ZHANG, Guozhen, et al. Vfimamba: Video frame interpolation with state space models. *Advances in Neural Information Processing Systems*, 2024, 37: 107225-107248.

Experiminents

- Ablation Studies
  - Learning Strategies
    - Curriculum learning delivers the best generalization on both small- and large-motion datasets
    - Mixed or sequential training is less effective



Performance of different learning methods, recorded every 30 epochs

	Curriculum Learning	Vimeo90K	X-TEST		SNU-FILM	
			2K	4K	hard	extreme
RIFE	✗	35.61/0.9797	31.10/0.8972	30.13/0.8927	30.36/0.9375	25.27/0.8601
	✓	35.60/0.9797	31.40/0.9142	30.23/0.9011	30.47/0.9376	25.38/0.8619
EMA-VFI-S	✗	36.07/0.9797	30.91/0.9000	29.91/0.8951	30.69/0.9375	25.47/0.8632
	✓	36.05/0.9797	31.15/0.9083	29.98/0.8988	30.73/0.9379	25.53/0.8652
VFIMamba-S	✗	**36.13/0.9802**	30.82/0.8997	29.87/0.8949	30.58/0.9378	25.30/0.8620
	✓	36.12/**0.9802**	**32.84/0.9328**	**31.73/0.9238**	**30.80/0.9381**	**25.59/0.8655**

Performance of different methods without or with curriculum learning

서강대학교
SOGANG UNIVERSITY

VDS LAB

- LC-Mamba: Local and Continuous Mamba with Shifted Windows for Frame Interpolation[1)]
    - CVPR 2025

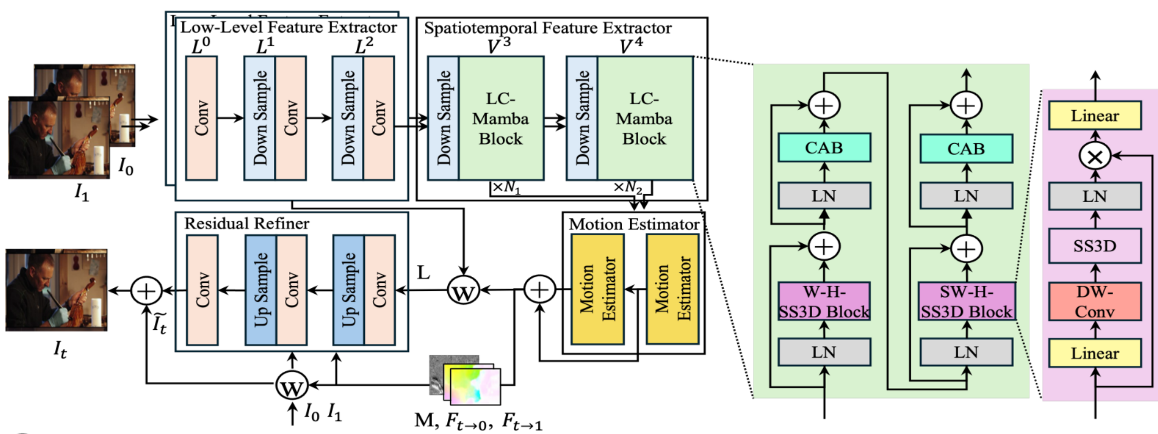
Introduction

- Key Ideas of LC-Mamba
  - Introduces Shifted-Window Hilbert-Scan (SW-H-SS2D) to maintain local continuity within each window
  - Proposes Interleaved 3D Scan (H-SS3D) for joint spatiotemporal feature fusion
- Core Contributions
  - Hybrid local-global feature extraction block combining SW-H-SS2D with lightweight attention
  - Superior quantitative gains (+0.03 dB PSNR on Vimeo-90K) and qualitative improvements over previous Mamba variants

1) JEONG, Min Wu; RHEE, Chae Eun. LC-Mamba: Local and Continuous Mamba with Shifted Windows for Frame Interpolation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025. p. 17671-17681.

Overall Architecture

- Low-Level Feature Extractor (LFE)
  - Pyramid of CNNs produces multi-scale features
- Spatiotemporal Feature Extractor (STFE)
  - Repeated LC-Mamba blocks
- Motion Estimator (ME)
  - Predicts dense flow and blending masks from fused features
- Reconstruction & Refinement (RR)
  - Warps input frames and blends using estimated masks
  - Final convolution for artifact removal



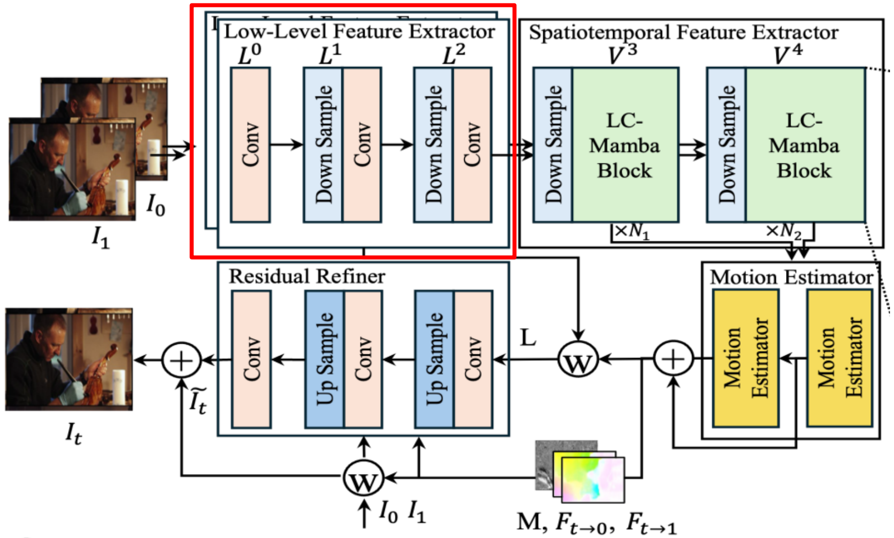
(a) Architecture (b) Two Successive LC-Mamba Blocks

서강대학교 SOGANG UNIVERSITY

VDS LAB

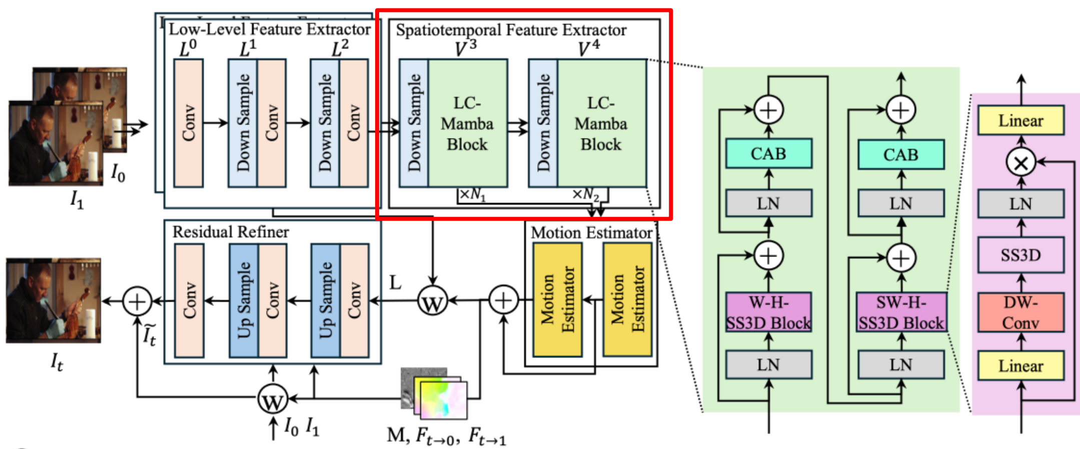
Low-Level Feature Extractor (LFE)

- LFE processes each input frame I_i to produce low-level feature maps $\{L_i^0, L_i^1, L_i^2\}$ at three pyramid levels
- At pyramid level l, feature L_i^l has spatial dimensions $\frac{H}{2^l} \times \frac{W}{2^l}$ and channel depth $2^l C$, halving resolution and doubling channels with each level
- Downstream Integration
  - The top-level features L_0^2 and L_1^2 from both frames are concatenated and passed to the Spatiotemporal Feature Extractor (STFE) for motion and context modeling



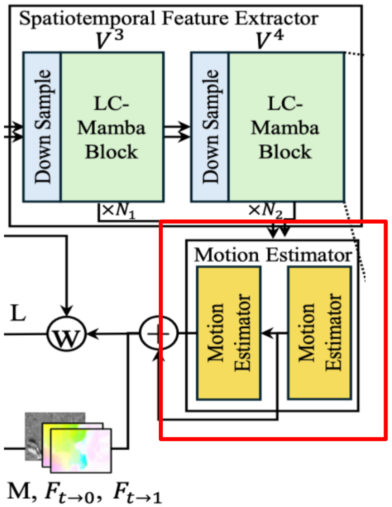
Spatiotemporal Feature Extractor (STFE)

- Input & Output
  - Takes the concatenated top-level LFE features $[L_0^2, L_1^2]$ and produces two levels of spatiotemporal motion features V^3 and V^4
- Module Composition
  - Built as a stack of hierarchical LC-Mamba blocks
- Downstream Role
  - Supplies V^3 and V^4 to the Motion Estimator (ME), which predicts bidirectional flows and blending masks for frame interpolation



(a) Architecture (b) Two Successive LC-Mamba Blocks

Motion Estimator (ME)



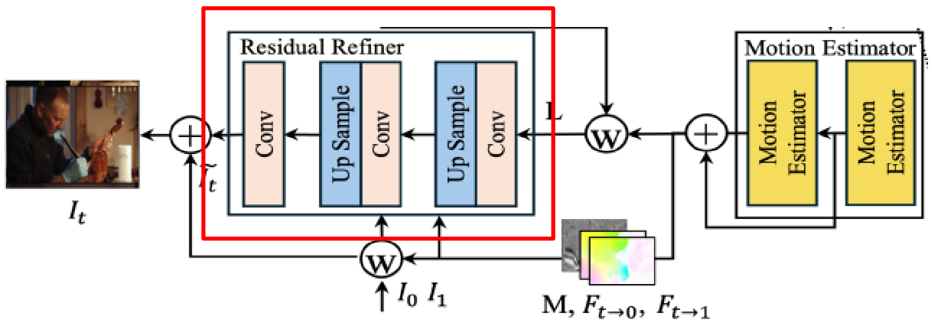
- Purpose & Inputs
  - Receives spatiotemporal features V^3 and V^4 from STFE
  - Aims to predict the coarse intermediate frame and the parameters for refinement
- Flow & Mask Estimation
  - Predicts bidirectional optical flows $F_{t\to0}$, $F_{t\to1}$ and a blending mask M
- Frame Synthesis
  - Applies backward warping and blends the two warped frames
- Output
  - Produces the preliminary interpolated frame $\widetilde{I}_t$, which is then passed to the Residual Refiner for final correction

-$\widetilde{I}_t = M \odot W_0 + (1 - M) \odot W_1$

1) JEONG, Min Wu; RHEE, Chae Eun. LC-Mamba: Local and Continuous Mamba with Shifted Windows for Frame Interpolation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025. p. 17671-17681.

Residual Refiner (RR)

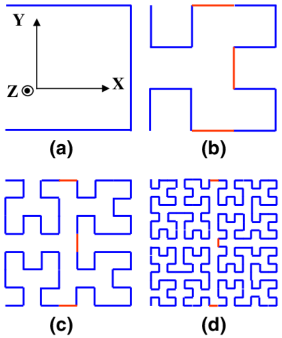
- Purpose
  - Refines the coarse interpolated frame by predicting and adding a residual image
- Inputs
  - Coarse frame $\widetilde{I}_t$ from the ME module
  - Original frames I_0, I_1
  - Low-level features L_i^l from LFE
  - Bidirectional flows $F_{t\to 0}$, $F_{t\to 1}$ and blending mask M



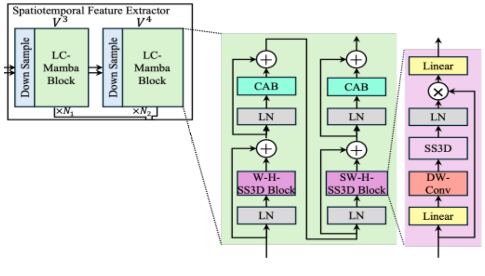
- Residual Prediction
  - A lightweight convolutional network processes $\{I, L, F, M\}$ to estimate a per-pixel residual R
- Final Output
  - Adds the predicted residual to the coarse frame

    - $I_t = \widetilde{I}_t + R$

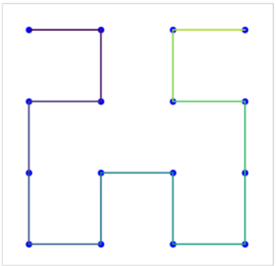
LC-Mamba Block (H-SS2D: Hilbert Curve)



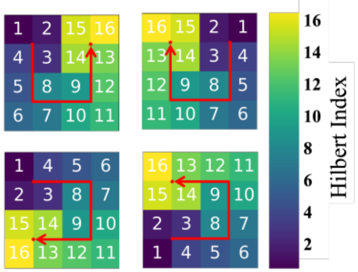
- Hilbert curve-based scanning
  - Converts 2D spatial data to 1D sequences while preserving local adjacency
- Four-path Hilbert structure
  - Uses multiple scanning directions to improve information propagation
- Window-based scanning
  - Limits the receptive field to local regions to reduce historical decay, especially effective for high-resolution input with dense motion.
- Shifted windows
  - Introduces cross-window interactions by alternating window partitions.



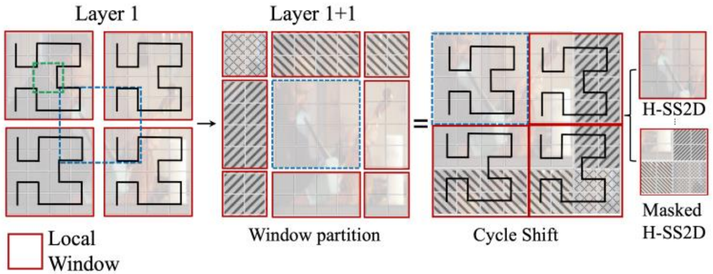
Two successive LC-Mamba block



Hilbert Curve based Scan



x4 scan paths

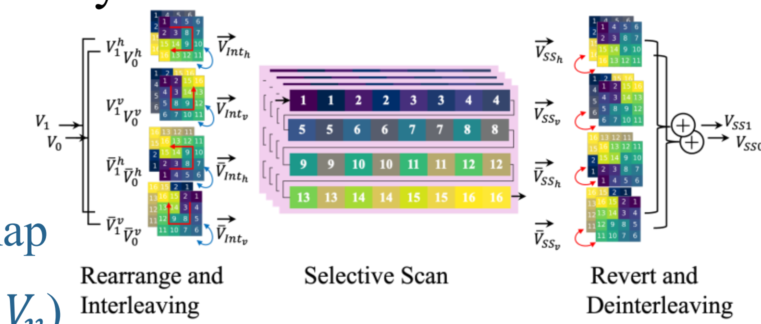


Shifted window based H-SS2D

1) JEONG, Min Wu; RHEE, Chae Eun. LC-Mamba: Local and Continuous Mamba with Shifted Windows for Frame Interpolation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025. p. 17671-17681.

LC-Mamba Block (H-SS3D)

- Interweaves tokens from two frames into a single 3D Hilbert curve–based sequence, preserving both spatial and temporal locality
- Workflow
  - Extract feature maps V_0, V_1 from frame 0 and 1
  - Apply H-SS2D in four Hilbert directions to each map
  - Interleave corresponding token sequences (e.g. V_h, V_v)
  - Perform windowed selective scans on interleaved sequences
  - Deinterleave and merge scanned outputs into spatiotemporal features



- Key Benefits
  - Captures fine-grained local details and long-range patterns across both space and time
  - Achieves linear complexity while modeling complex motions in high-resolution videos

Experimental Results

- Datasets
  - Vimeo90K: 3,782 frame triplets at 448 × 256 resolution
  - UCF101: 379 triplets at 256 × 256
  - Middlebury: OTHER set at ∼640 × 480
  - SNU-FILM: 1,240 triplets at 1280 × 720, split by motion difficulty
  - Xiph: tested on downsampled 2K and centrally-cropped 4K versions

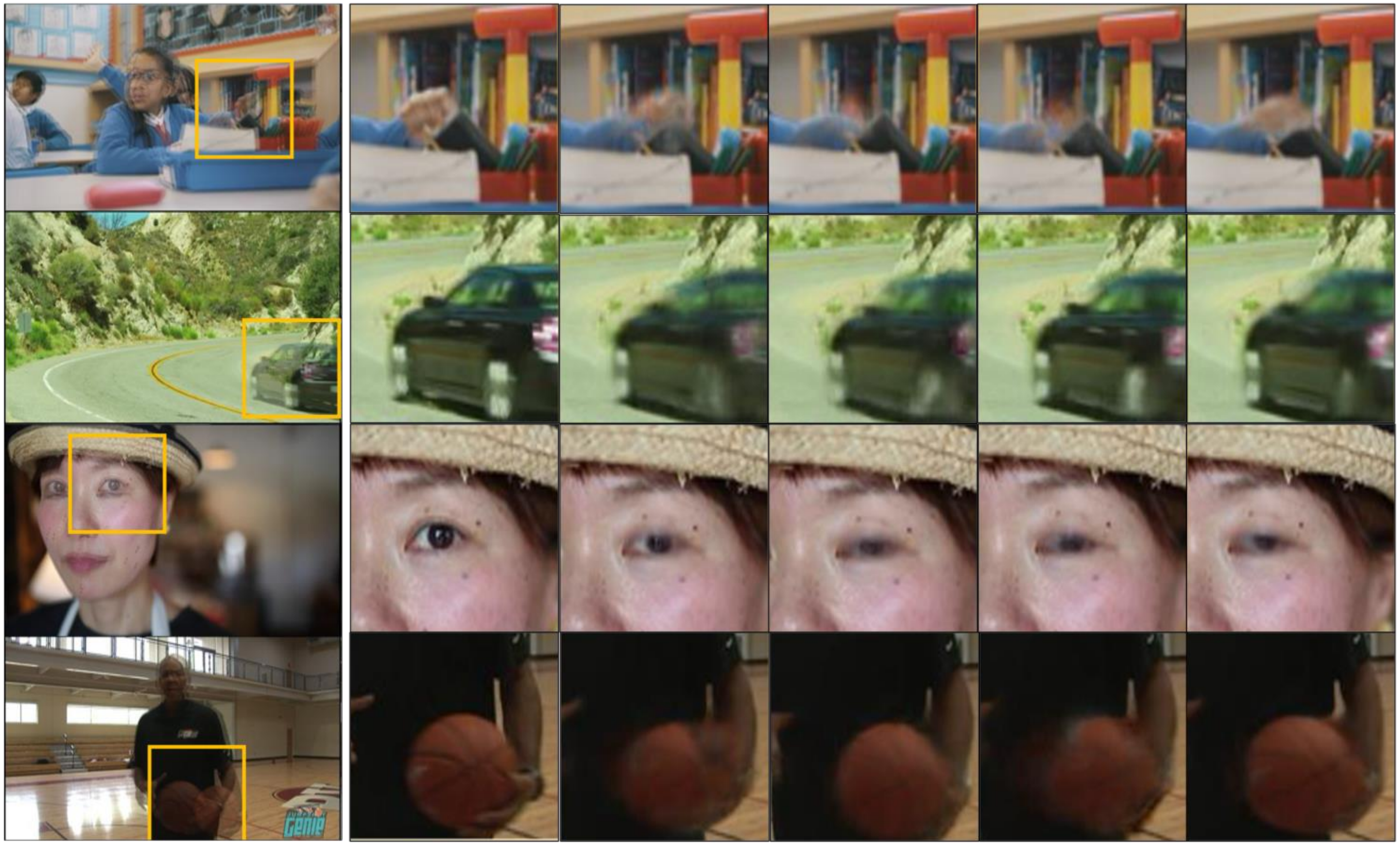
Experimental Results

- Quantitative Evaluation

Method	Vimeo90K	UCF101	Xiph		M.B.	SNU-FILM				Params (M)
			2K	4K		Easy	Medium	Hard	Extreme	
ToFlow [1]	33.73/0.9682	34.58/0.9667	33.93/0.922	30.74/0.856	2.15	39.08/0.9890	34.39/0.9740	28.44/0.9180	23.39/0.8310	1.4
IFRNet [15]	35.80/0.9794	35.29/0.9693	36.00/0.936	33.99/0.893	1.95	**40.03**/0.9905	**35.94**/0.9793	30.41/0.9358	25.05/0.8587	5
M2M [11]	35.47/0.9778	35.28/0.9694	36.44/0.943	33.92/0.899	2.09	39.66/0.9904	35.74/0.9794	30.30/0.9360	25.08/0.8604	7.6
SoftSplat [30]	36.10/0.9802	**35.39/0.9697**	36.62/**0.944**	33.60/0.901	**1.81**	39.88/0.9897	35.68/0.9772	30.19/0.9312	24.83/0.8500	7.7
RIFE [14]	35.61/0.9779	35.28/0.9690	36.19/0.938	33.76/0.894	1.96	39.80/0.9903	35.76/0.9787	30.36/0.9351	25.27/0.8601	9.8
BMBC [31]	35.01/0.9764	35.15/0.9689	32.82/0.928	31.19/0.880	2.04	39.90/0.9902	35.31/0.9774	29.33/0.9270	23.92/0.8432	11.1
Ours-E	**36.19/0.9803**	35.33/0.9695	**36.67**/0.943	**34.26/0.903**	1.98	39.82/**0.9907**	35.87/**0.9797**	**30.54/0.9373**	**25.33/0.8626**	6.7
EMA-S [46]	36.07/0.9794†	35.34/0.9696†	36.54/0.942†	34.24/0.902†	1.94†	39.81/0.9903†	35.88/0.9792†	30.68/0.9371†	25.47/0.8627†	14.5
VFIMamba-S [47]	36.09/0.9800†	35.35/0.9696†	36.71/0.942†	**34.26**/0.902†	1.97†	**40.21/0.9912**†	**36.17/0.9802**†	**30.80/0.9382**†	**25.59/0.8655**†	16.8
VFIFormer-S [26]	36.37/0.9810†	35.36/**0.9698**†	36.55/0.943†	33.37/0.899†	**1.89**†	40.02/0.9906†	35.91/0.9793†	30.22/0.9348†	24.80/0.8568†	17.1
ABME [32]	36.18/0.9805	35.38/**0.9698**	36.53/0.944	33.73/0.901	2.01	39.59/0.9901	35.77/0.9789	30.58/0.9364	25.42/0.8639	18.1
Ours-B	**36.43/0.9813**	**35.39/0.9698**	**36.90/0.945**	**34.26/0.904**	**1.89**	40.07/0.9909	36.08/0.9801	30.59/0.9375	25.35/0.8630	16.2
SGM-VFI-S-1/2 [19]	35.81/0.9785†	35.33/0.9692†	36.06/0.940†	33.26/0.897†	1.87†	**40.36**/0.9900†	36.12/0.9787†	30.62/0.9351†	25.38/0.8615†	20.8
SepConv [5]	33.79/0.9702	34.78/0.9669	34.77/0.929	32.06/0.880	2.27	39.41/0.9900	34.97/0.9762	29.36/0.9253	24.31/0.8448	21.7
AdaCoF [16]	34.47/0.9730	34.90/0.9680	34.86/0.928	31.68/0.870	2.24	39.80/0.9900	35.05/0.9754	29.46/0.9244	24.31/0.8439	21.8
DAIN [2]	34.71/0.9756	34.99/0.9683	35.95/0.940	33.49/0.895	2.04	39.73/0.9902	35.46/0.9780	30.17/0.9335	25.09/0.8584	24.0
VFIFormer [26]	36.50/0.9815†	**35.42/0.9699**†	OOM†	OOM†	**1.82**†	40.12/0.9907†	36.09/0.9798†	30.67/0.9378†	25.43/0.8643†	24.1
Ours-P	**36.53/0.9816**	**35.42/0.9699**	**36.99/0.946**	**34.49/0.906**	1.92	40.16/**0.9909**	**36.17/0.9802**	**30.72/0.9382**	**25.48/0.8645**	25.4
CAIN [6]	34.65/0.9730	34.91/0.9690	35.21/0.937	32.56/0.901	2.28	39.89/0.9900	35.61/0.9776	29.90/0.9292	24.78/0.8507	42.8
EMA [46]	36.50/0.9814†	35.38/0.9697†	36.74/0.944†	**34.54**/0.905†	**1.84**†	39.57/0.9905†	35.85/0.9797†	30.80/**0.9389**†	25.59/0.8650†	65.6
VFIMamba [47]	36.45/0.9807†	35.37/**0.9699**†	**37.02**/0.944†	34.39/0.904†	1.89†	**40.41**/0.9903†	**36.30**/0.9794†	**30.89**/0.9387†	**25.68/0.8661**†	66.1
Ours-P	**36.53/0.9816**	**35.42/0.9699**	36.99/**0.946**	34.49/**0.906**	1.92	40.16/**0.9909**	36.17/**0.9802**	30.72/0.9382	25.48/0.8645	25.4

Experimental Results

- Qualitative Evaluation



Overlay Ground-Truth EMA-S SGM-VFI-S-1/2 VFIMamba-S Ours-E

1) JEONG, Min Wu; RHEE, Chae Eun. LC-Mamba: Local and Continuous Mamba with Shifted Windows for Frame Interpolation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025. p. 17671-17681.

Experimental Results

- Ablation Study

Scanning	Vimeo90K	Xiph-2K	Xiph-4K	SNU-FILM(avg.)
Bidirection w/ ILV	35.41/0.9799	36.00/0.9381	33.13/0.8937	32.32/0.9405
Cross w/ ILV	36.07/0.9799	35.73/0.9362	33.80/0.8947	32.53/0.9413
Continuous w/ ILV	36.09/0.9800	36.57/0.9428	33.99/0.9010	24.59/0.8335
Local w/ ILV	36.11/0.9801	36.38/0.9415	34.01/0.9008	32.62/0.9411
Z-order w/ ILV	36.13/0.9800	35.91/0.9371	33.30/0.8932	32.36/0.9417
SW-H-SS3D	**36.19/0.9803**	**36.67/0.9437**	**34.26/0.9036**	**32.89/0.9426**

Performance comparison of different scanning methods

Settings	Vimeo90K	Xiph-2K	Xiph-4K	SNU-FILM(avg.)
8 w/ shift	36.43/**0.9813**	**36.90**/0.9452	**34.26**/0.9046	33.02/**0.9429**
8 w/o shift	36.45/**0.9813**	36.78/0.9448	34.15/0.9042	32.95/0.9428
16 w/ shift	36.44/**0.9813**	36.88/**0.9454**	34.15/**0.9047**	33.02/**0.9429**
16 w/o shift	**36.46/0.9813**	36.88/0.9449	34.23/0.9045	**33.05/0.9429**

Ablation studies for window settings

N_1 / N_2 / C	Vimeo90K	Xiph-2K	Xiph-4K	SNU-FILM(avg.)
4 / 4 / 16	36.19/0.9803	36.67/0.9437	34.26/0.9036	32.89/ 0.9426
2 / 2 / 32	36.43/0.9813	36.90/0.9452	34.26/0.9046	33.02/0.9429
4 / 4 / 32	**36.53/0.9816**	**36.99/0.9459**	**34.49/0.9061**	**33.13/0.9435**

Ablation study on the scalable capability of LC-Mamba blocks