A Framework with Multi-Scale Hybrid Mamba Voxel Flow for Video Prediction

Muhao Xu, Baochen Fu, Dongyu Liu, Wenzhi Deng, Wei Yi, Yi Wan, Hua Wei and Weiye Song

***Abstract*—Video prediction is a critical task in video processing and generation, with far-reaching implications for various downstream applications. However, existing methods often produce blurred predicted frames and fail to maintain structural continuity in objects. To address these challenges, we propose a Multi-Scale Hybrid Mamba Voxel Flow framework that employs a progressive refinement strategy in combination with adaptive feature extraction modules. The framework begins by generating coarse optical flow estimates and predicted frames, which are progressively refined at lower resolutions to enhance detail and ensure temporal coherence. Specifically, Mamba Blocks are designed to capture complex global motion patterns, while Spatial Aggregation Blocks aggregate spatial context across different scales. Simam Modules further enhance feature representation by selectively focusing on significant spatial regions. Additionally, multi-level residual connections and depthwise channel separations are incorporated to reduce computational complexity. Experimental results show that the proposed method significantly improves the clarity and spatial consistency of predicted frames, outperforming state-of-the-art techniques.**

***Index Terms*—Video prediction, Multi-Scale Hybrid, Mamba, Voxel Flow.**

I. Introduction

Video prediction [1]–[3] aims to estimate future frames based on the current ones and holds substantial promise for improving representational learning [4] while supporting downstream tasks such as human motion forecasting [5], [6], autonomous driving [7]–[9], and anomaly detection [10]. This field has garnered growing attention in both academic and industrial communities [11]–[13].

Despite its significance, video prediction faces significant challenges posed by the diverse and complex motion patterns in real-world scenarios. Accurate motion estimation is essential for overcoming these challenges [14]–[17]. Early approaches primarily relied on recurrent neural networks (RNNs) [19] for modeling temporal motion patterns. To achieve more robust long-term predictions, some studies [18], [19] have incorporated semantic or instance segmentation maps to enable semantically consistent motion estimation in complex scenes. However, the practical availability of such segmentation maps is often limited, which restricts the real-world applicability of these approaches [15], [20], [21].

To overcome this constraint and minimize dependency on extra inputs, the OPT method [22] achieved significant results by leveraging only RGB images and estimating optical flow using an optimization-based approach. However, it is difficult to obtain pre-trained optical flow models and accurate optical flow, which affects its applicability. The optimization method based on VFI interpolation and extrapolation leads to huge computational overhead. DMVFN [23] introduced dynamic optical flow estimation to explicitly capture complex multi-scale motion patterns between adjacent frames. It employs a lightweight routing module that adaptively generates routing vectors based on the input frames, dynamically selecting sub-networks for efficient future frame prediction. However, because the CNNs in DMVFN are limited by their local receptive fields, their predictions for fast-moving objects often become ambiguous [24]. Real-world video prediction tasks often require the ability to manage substantial variations in spatial resolution. While methods such as DMVFN extract multi-scale motion features using varying receptive fields, they frequently struggle to preserve the structural continuity of objects in their predictions.

To address these challenges, we propose a Multi-Scale Hybrid Mamba Voxel Flow framework that aims to resolve blurred predictions and maintain structural consistency in objects. This framework models complex multi-scale motion between adjacent frames through a progressive refinement strategy, enabling efficient capture of broad displacements at lower resolutions and successive refinement of subtle local deformations at higher resolutions. It comprises a series of sequentially stacked multi-scale Hybrid Mamba Voxel Flow modules, each integrating a Mamba Block for global context modelling with a Spatial Aggregation Block for local detail enhancement. Mamba Block employs self-attention to capture long-range dependencies across the entire frame, enabling the model to relate distant regions and accurately infer broad motion patterns. Spatial Aggregation Block utilises a dual branch residual convolution design to emphasise texture fidelity and reinforce edge continuity, ensuring precise reconstruction of subtle structures and object boundaries. The SimAM module then adaptively fuses the global and local feature streams, weighting them according to spatial salience, to produce coherent and sharp frame predictions at all scales. Our experiments on three established public benchmarks - Cityscapes [25], KITTI [26], and UCF101 [27] - demonstrate that the proposed

This work was supported in part by the Youth Project of Natural Science Foundation of Shandong Province, China (ZR2023QC262); the National Natural Science Foundation of China (62205181); the Natural Science Foundation of Shandong Province (ZR2022QF017); and the Shandong Province Outstanding Youth Science Fund Project (Overseas) (2023HWYQ-023).

Muhao Xu and Baochen Fu contributed to the work equally and should be considered as co-first authors. Muhao Xu, Baochen Fu, Dongyu Liu, Wenzhi Deng, Wei Yi, Yi Wan, Hua Wei and Weiye Song are with the Department of Mechanical Engineering, Key Laboratory of High Efficiency and Clean Mechanical Manufacture of Ministry of Education, Shandong University, Jinan 250061, China. Baochen Fu is also with the School of Software, Shandong University, Shandong, Jinan 250061, China. Dongyu Liu is also with the Department of Bioengineering, The University of Texas at Dallas, Richardson, TX 75080, USA. (Corresponding authors: Weiye Song, e-mail:songweiye@sdu.edu.cn)

Manuscript received on February 15, 2025.

method achieves superior performance compared to state-of-the-art video prediction approaches. Extensive ablation studies further confirm the individual contribution of each framework component to the overall prediction accuracy.

Our contributions can be summarized as follows:

- We propose a multi-scale Hybrid Mamba Voxel Flow framework that employs a progressive refinement strategy to capture complex multi-scale motion. We introduce the Mamba Block, which leverages local self-attention to effectively extract global motion information.
- We introduce the Spatial Aggregation Block to improve local detail prediction via a dual-branch residual structure and the Simam Module to adaptively integrate features from different blocks, significantly enhancing the model's predictive capability.
- Our framework achieves state-of-the-art performance on benchmark datasets and is rigorously validated through extensive experiments and ablation studies.

II. Related Works

In this section, we provide an overview of existing research on short-term video prediction, spatiotemporal prediction, and optical flow estimation.

A. Short-term Prediction

Liu et al. [1] proposed the Deep Voxel Flow (DVF) network to address the problem of video frame synthesis. DVF is a fully convolutional encoder-decoder network that learns to synthesize video frames from existing ones using current pixel flows. DVF introduces a voxel flow layer, extending optical flow from two to three dimensions, enabling stable learning of motion for synthesizing or rendering objects, thus providing a novel approach to video prediction. Subsequently, to achieve better performance, video prediction methods began leveraging additional information. Wang et al. [28] proposed a video synthesis method, Vid2Vid, based on generative adversarial learning, which learns a mapping function from video input to output using a generative adversarial framework. By utilizing a sequence of semantic segmentation masks, Vid2Vid enables multimodal video synthesis. Pan et al. [29] proposed a video generation method based on a single semantic label map, dividing the task into two subtasks: generating an initial frame using an image generation model and then animating the frame using the predicted optical flow from a conditional variational autoencoder (cVAE). Extensive experiments show that semantic information significantly improves optical flow prediction and video frame generation. Wu et al. [18] proposed an object motion prediction-based method that separates dynamic objects from static backgrounds, predicting future motion paths, scaling, and shapes of dynamic objects to generate more accurate and realistic future videos.

Bei et al. [30] proposed a Semantic-Aware Dynamic Model (SADM) that decomposes scene layouts (semantic maps) and motions (optical flow) into layers, which are predicted and fused with their context to generate future layouts and motions. By detecting occlusion areas using predicted semantic maps and synthesizing these regions through content-aware inpainting, SADM generates more realistic and natural predicted videos.However, additional inputs like optical flow, semantic segmentation maps, and instance segmentation maps are often difficult to accurately obtain or estimate, typically requiring model training tailored to specific scenarios. When unrecognized objects appear in the scene, performance often degrades significantly. Hu et al. [23] proposed the Dynamic Multi-scale Voxel Flow Network (DMVFN), which achieves better prediction performance using only RGB images as input. The network comprises multiple Multi-scale Voxel Flow Blocks (MVFBs) constructed with convolutions and stacked sequentially. However, convolution-based methods rely on local receptive fields, limiting their ability to capture global spatial information and performing poorly on tasks requiring strong global dependencies.

B. Spatiotemporal Prediction

Shi et al. [31] proposed the Convolutional Long Short-Term Memory (ConvLSTM) network, which extracts spatial features from images through convolution operations while leveraging LSTM units to capture temporal dynamics, enabling joint modeling of spatial and temporal features and excelling in sequential data processing. Lotter et al. [32] developed the recursively structured PredNet, based on the predictive coding principles of neurobiology. This network generates predictions layer by layer, compares them with actual observations to produce error signals, and feeds the deviations back to subsequent layers to optimize predictive performance. Villegas et al. [33] designed a method to decouple motion and content, simplifying the video prediction task. They used two independent encoding paths to extract spatiotemporal dynamics and static layouts separately and combined motion features to transform content features into the next frame.

However, these methods are generally designed for long-sequence video prediction and are limited in modeling the details of high-resolution video frames. Liu et al. [34] proposed an encoder-decoder architecture based on dynamic atoms. This method uses sparse optimization to map consecutive frames into a dynamic feature space and employs a decoder to reconstruct the input data and predict future frames, effectively capturing dynamic characteristics of the data. Geng et al. [35] proposed a "correspondence loss" to reduce blurry results in video prediction. They aligned predicted images with ground truth using optical flow and calculated the loss for corresponding pixels, enhancing the model's focus on object positioning and significantly reducing prediction blur. Despite significant progress, these methods still face challenges in high-resolution video prediction, such as inaccurate frame prediction, reliance on long-sequence inputs, and high computational resource consumption. Voleti et al. [36] proposed the Masked Conditional Video Diffusion (MCVD) framework, which delivers an integrated model for video prediction, generation and interpolation by randomly masking past or future frames and employing a diffusion-based denoising process. Zhang et al. [37] introduced ExtDM, which uses a motion autoencoder and a distribution extrapolation module within a diffusion U-Net to explicitly predict future feature distributions. This enhances the accuracy and efficiency of long-range video prediction. Yuan et al.

[38] presented Spatio-Temporal Non-Autoregressive Model (STNAM), a model that enables the parallel generation of all future frames via spatio-temporal attention mechanisms. This significantly accelerates inference while mitigating error accumulation.

C. Optical Flow Estimation

Optical flow reflects the temporal motion changes of each pixel in an image. By estimating optical flow, it is possible to infer the motion of objects or the camera within a scene, thereby understanding the motion patterns in the image. Fischer et al. [39] proposed FlowNet, the first convolutional neural network-based framework for optical flow estimation. By introducing correlation layers into the network, it achieved cross-frame feature matching, significantly improving estimation accuracy. However, its computational efficiency is relatively low. To address this, Ranjan et al. [40] designed SPyNet, which processes optical flow estimation in a multi-scale manner based on a spatial pyramid structure, achieving both efficiency and accuracy, especially excelling in handling large motions. Sun et al. [41] further optimized the optical flow estimation process with PWC-Net. This method constructs learnable feature pyramids combined with warping operations, reducing computational redundancy while improving adaptability to large-motion scenarios. Teed et al. [42] proposed the RAFT framework, which achieves precise optical flow estimation by constructing multi-scale 4D correlation volumes and iterative optimization operations, demonstrating excellent performance on multiple benchmark datasets.

To tackle occlusion issues, Jiang et al. [43] introduced the Global Motion Aggregation (GMA) module, which effectively improves the quality of optical flow estimation in occluded regions by modeling long-range dependencies between pixels within a frame. Huang et al. [44] designed FlowFormer, which leverages a Transformer architecture combined with a cost volume encoder and cost memory decoder. This method exhibits strong global modeling capabilities and flexibility in optical flow estimation tasks.

III. Proposed Method

A. Introduction and Overall Pipeline

To enable video next-frame prediction, we propose a Multi Scale Hybrid Mamba Voxel Flow framework. As depicted in Fig. 1, the input frames $I_{t-1} \in \mathbb{R}^{h\times w\times 3}$ and $I_t \in \mathbb{R}^{h\times w\times 3}$ are initially processed at a coarse scale (Scale = 4) by the Hybrid Mamba Voxel Flow Modules, producing an initial flow estimation F^i as well as an intermediate predicted frame I^i_{t+1}. Subsequently, at finer scales (Scale = 2 and Scale = 1), the flow and intermediate results are iteratively refined, ultimately yielding the final high-resolution target frame I_{t+1}. This progressive approach enables the framework to effectively capture both coarse global motion and subtle local details, thereby improving prediction accuracy across multiple spatial scales. In Fig. 1, the term Planes denotes the number of convolutional channels within each module (i.e., the channel dimension C of feature maps. Planes controls the representational capacity of the convolutional layers at different stages of the network and thus determines the trade-off between modelling power and computational cost.

The Hybrid Mamba Voxel Flow Module is composed of several key components: the Mamba Block, the Spatial Aggregation Block (SA Block), the Simam Module, and feature downsampling and upsampling. At each spatial scale, the pipeline repeatedly applies an identical Mamba Block and the Spatial Aggregation Block to compute an initial flow estimate. Once the coarse flow at the lowest resolution (4× downsampled) has been predicted, it is upsampled (pixel-shuffled) to the next finer grid (2× downsampled). This corrected flow is then upsampled to full resolution and warped against the highest-resolution features. A final Mamba Block and the Spatial Aggregation Block pass then refines any remaining local misalignments. By progressively refining motion estimates from low to high resolutions, each module focuses on motions at its own scale—coarse modules capture large displacements efficiently, while finer modules recover detailed local movements.

In the coarse-scale stage, we first feed the frames I_{t-1} and I_t into a downsampling process to reduce their spatial resolution to one-quarter of the original. Specifically, we apply bilinear interpolation to obtain optical flow estimation. Meanwhile, the Hybrid Mamba Voxel Flow Modules produce an initial flow field F^i and an intermediate predicted frame I^i_{t+1}. Next, we concatenate these four features along the channel dimension to construct the input tensor X_{in}:

$$X_{in} = Concat\Big(Downsample(I_{t-1},\, I_t,\, I^i_{t+1},\, F^i)\Big), \quad (1)$$

where $Concat(\cdot)$ stacks I_{t-1}, I_t, I^i_{t+1}, and F^i along the channel dimension, $Downsample(\cdot)$ adopts bilinear interpolation to generate coarse feature representations. The resulting combined feature $X_{in} \in \mathbb{R}^{h\times w\times C'}$ simultaneously incorporates coarse-scale motion (i.e., optical flow) and predictive information (i.e., frames), enabling subsequent network layers to refine further both the flow and the predicted results under a broader receptive field and reduced computational load. Finally, we feed X_{in} into both the Mamba Block and the Spatial Aggregation Block for subsequent feature prediction. In each Multi-Scale Hybrid Mamba Voxel Flow module a residual connection is entered to improve the predictive power of the model.

B. Mamba Block

The Mamba Block integrates convolutional layers, a dual-branch feature extractor, PatchExpand, LayerNorm, local self-attention (SS2D), upsampling, and residual fusion into a unified processing pipeline. This architecture was chosen because it balances the fine-grained locality of convolution with the long-range context modeling of self-attention, while residual fusion ensure efficient resolution restoration and gradient flow—yielding high predictive accuracy at moderate computational cost.

First, the input feature X_{in} passes through two sequential convolution and activation layers (where GELU is denoted by

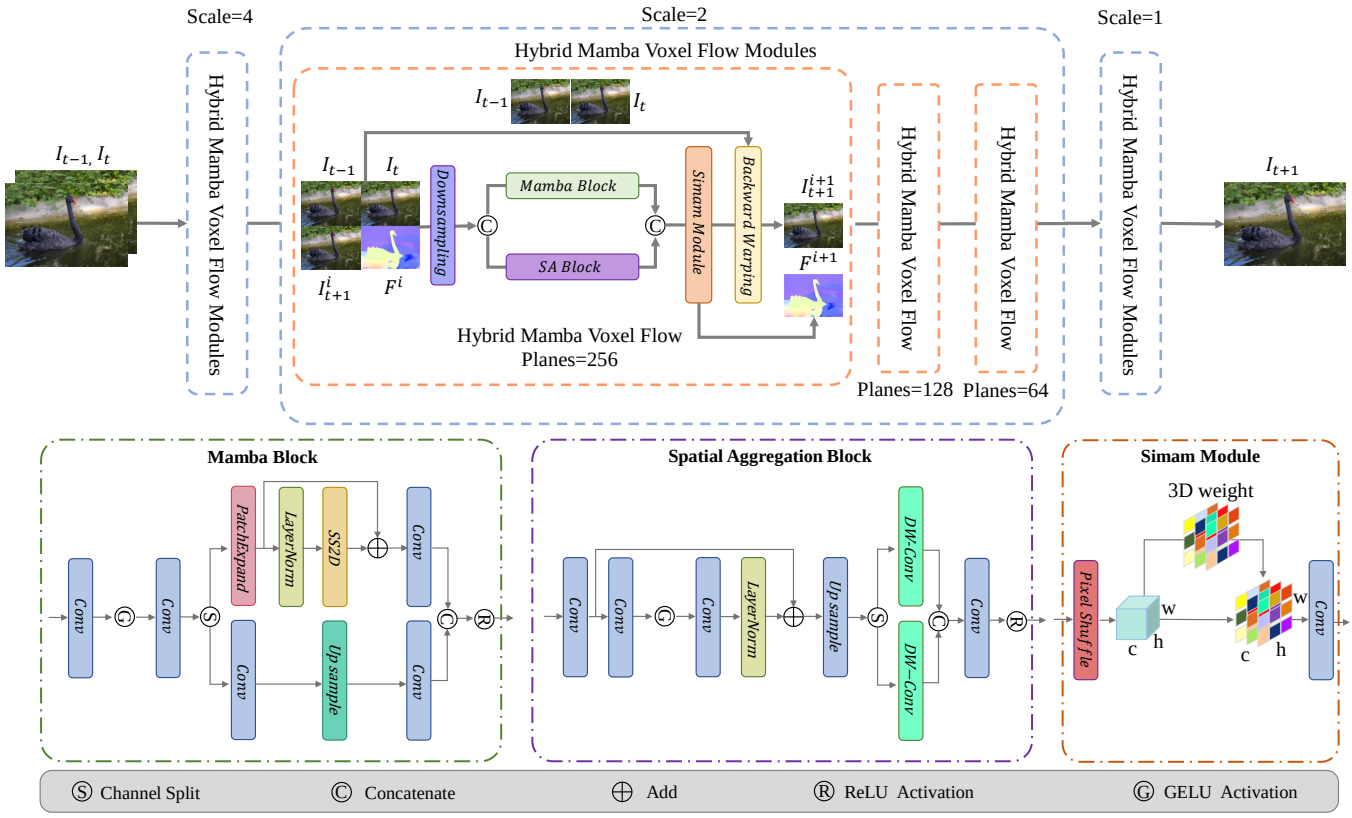

Fig. 1. Overview architecture of proposed Framework.

G), as well as Split (denoted by S) for channel splitting. It can be expressed by the formula:

$$\begin{aligned} X_1 &= GELU(Conv(X_{in})) \\ (X_{s1}, X_{s2}) &= Split(Conv(X_1)). \end{aligned} \tag{2}$$

where $Conv(\cdot)$ denotes a standard 3×3 convolution that extracts representation.

Next, PatchExpand linearly projects each token's features into a higher-dimensional channel space, doubling the channel capacity. These expanded channels are then rearranged into a finer spatial grid to restore resolution while preserving feature coherence. This is followed by a LayerNorm operation to normalise and balance the resulting feature distributions across all channels:

$$\begin{aligned} X_{patch} &= PatchExpand(X_{s1}), \\ X_{norm} &= LayerNorm(X_{patch}). \end{aligned} \tag{3}$$

Subsequently, the 2D-Selective-Scan module (SS2D) operates on X_{norm} to learn pixel-level dependencies, producing:

$$\begin{aligned} X_{attn} &= SS2D(X_{norm}), \\ X_{sa} &= Conv(X_{attn} \oplus X_{patch}), \end{aligned} \tag{4}$$

where $\oplus$ denotes element-wise addition. Meanwhile, the other branch convolution and upsamples X_{s2} to obtain the feature X_p and then concatenates it with X_{sa} along the channel dimension. This step is followed by a convolution and ReLU activation, resulting in:

$$X_{Mres} = ReLU\Big[Concat(Conv(X_{sa}),\, Conv(X_p))\Big]. \tag{5}$$

The final output X_{Mres} from Mamba Block provides high-quality feature representations for subsequent tasks such as optical flow estimation and frame prediction.

C. Spatial Aggregation Block

The proposed Spatial Aggregation Block effectively enriches spatial feature representations while controlling computational complexity. Unlike traditional structures, this module organically integrates multiple consecutive convolutions, upsampling, and parallel depthwise convolutions, complemented by both local residual connections. By adaptively balancing fine local cues with broader contextual information, this module significantly improves the representational power of the network. First, the input feature map is processed by a series of convolutions:

$$\begin{aligned} X_{1'} &= Conv(X_{in}), \\ X_{2'} &= GELU(Conv(X_{1'})), \\ X_{3'} &= Conv(X_{2'}) \end{aligned} \tag{6}$$

To stabilize the inter-channel distribution and expedite convergence, LayerNorm is applied. Additionally, a local residual connection is introduced to alleviate gradient vanishing and preserve early-stage information:

$$\begin{aligned} X_{ln} &= LayerNorm(X_{3'}), \\ X_{lr} &= X_{ln} \oplus X_{1'}, \end{aligned} \tag{7}$$

Next, X_{lr} is upsampled to increase its spatial resolution, and the resulting feature map is then split:

$$\begin{aligned} X_{up} &= Upsample(X_{lr}), \\ (X_{sp^1},\ X_{sp^2}) &= Split(X_{up}), \end{aligned} \tag{8}$$

Subsequently, the feature tensor is split into two logical branches along the channel dimension and executed sequentially in a single thread. Each branch then performs a depthwise convolution:

$$\begin{aligned} X_{sp^{1'}} &= DWConv(X_{sp^1}), \\ X_{sp^{2'}} &= DWConv(X_{sp^2}), \end{aligned} \tag{9}$$

where $DWConv(\cdot)$ reduces the computational cost by independently operating on each channel.

Finally, a convolutional layer and an activation function are applied to the fused features$X_{sp^{1'}}$ and $X_{sp^{2'}}$:

$$\begin{aligned} X_m &= Concat(X_{sp^{1'}},\ X_{sp^{2'}}), \\ X_{Sres} &= ReLU(Conv(X_m)). \end{aligned} \tag{10}$$

This structure efficiently integrates local details and global semantics via multi-level convolutions, upsampling, and parallel depthwise convolutions. The effect of local residual connections ensures effective gradient propagation and feature reuse, endowing the module with both high efficiency and powerful representational capability.

D. Simam Module

The Simam Module processes the concatenated outputs from the Mamba Block and the Spatial Aggregation (SA) Block. Specifically, the inputs X_{Mres} and X_{Sres} are first concatenated along the channel dimension to form the input tensor X_{Sin}:

$$X_S = Concat(X_{Mres}, X_{Sres}), \tag{11}$$

Next, the concatenated tensor X_{Sin} is fed into the Pixel Shuffle layer, which rearranges the feature map by spatially redistributing pixel information:

$$X_{shuffled} = PixelShuffle(X_S), \tag{12}$$

The pixel-shuffled features are first rearranged into a higher-resolution 3D grid and then passed through a learnable 3D-weight block, which applies element-wise weights across both spatial and channel dimensions to model their interdependencies. Specifically, each channel is mean-centered and its squared deviations are normalized by the spatial variance, then passed through a sigmoid to produce an attention map, which is applied element-wise to the original features to modulate each location according to its learned importance. The resulting weighted volume is finally refined by a standard convolutional layer:

$$\begin{aligned} X_{3D} &= 3DWeight(X_{shuffled}), \\ X_{out} &= Conv(X_{3D}), \end{aligned} \tag{13}$$

where the 3D weight block adjusts the spatial and channel dimensions based on the learned weight parameters.

The resulting output X_{out} contains the predicted optical flow field F^{i+1}, which facilitates more accurate frame alignment by correcting potential occlusions or large displacements, while also incorporating information from the next frame. Finally, X_{out} is passed through the backward warping block, where spatial transformations are applied to refine the alignment and synthesize the final output. The backward warping yields I_{t+1}^{i+1}, the predicted frame at time $t+1$.

E. Loss Function

The loss function consists of the Laplacian Loss (L_{lap}) and Perceptual Loss (L_{perc}), with the total loss being the weighted sum of these two terms. The loss is defined as:

$$L_{total} = \alpha L_{lap} + \beta L_{perc}, \tag{14}$$

$$L_{lap} = \frac{1}{N}\sum_{i=1}^{n} \gamma^{n-i} d(\tilde{I}_{t+1}^{i}, I_{t+1}), \tag{15}$$

$$L_{perc} = \sum_{i=1}^{k} \left\| F_i(\tilde{I}_{t+1}) - F_i(I_{t+1}) \right\|_2^2. \tag{16}$$

Where α and β represent the weights of the losses, we set them to 1 and 0.5, respectively, to balance structural similarity and visual quality. d denotes the ℓ_1 loss applied to the Laplacian pyramid representations derived from each pair of images. $\tilde{I}_{t+1}^{i}$ refers to the predicted frame output by the i-th optical flow block, while I_{t+1} represents the ground truth frame at time $t+1$. The perceptual loss is computed between the predicted frame $\tilde{I}_{t+1}$ from the output of the final block and the ground truth. K represents the number of selected convolutional layers, set to 5, and F_i denotes the feature representation of the i-th selected layer.

IV. Experiments

A. Dataset

Cityscapes [25] is a comprehensive dataset containing images of urban street scenes collected from 50 cities, designed to simulate diverse scenarios encountered in real-world autonomous driving tasks. The dataset captures a wide range of variations in weather conditions, time of day, seasons, and urban environments. These attributes make it a valuable resource for developing and evaluating algorithms intended for challenging autonomous driving scenarios. Cityscapes consists of 3,475 high-definition videos, each with a resolution of 2048×1024. Among these, 2,975 videos are allocated for training, while the remaining 500 are reserved for testing.

KITTI [26] serves as one of the most widely recognized benchmarks for assessing computer vision algorithms in autonomous driving contexts. This dataset was captured using high-resolution color and grayscale cameras mounted on vehicles, covering a diverse set of environments such as urban areas, rural roads, and highways. Each image in the dataset contains up to 15 vehicles and 30 pedestrians, offering a rich environment for motion prediction algorithms. The dataset comprises 28 driving videos, each with a resolution of 1242×375. Of these, 24 videos are designated for training,

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE CITYSCAPES, AND KITTI DATASETS. "RGB", "F", "S" AND "I" DENOTE THE VIDEO FRAMES, OPTICAL FLOW, SEMANTIC MAP, AND INSTANCE MAP, RESPECTIVELY. "N/A" MEANS NOT AVAILABLE.

Method	Input	Cityscapes						KITTI					
		MS-SSIM($\times 10^{-2}$)↑			LPIPS($\times 10^{-2}$)↓			MS-SSIM($\times 10^{-2}$)↑			LPIPS($\times 10^{-2}$)↓		
		t+1	t+3	t+5	t+1	t+3	t+5	t+1	t+3	t+5	t+1	t+3	t+5
Vid2vid [28]	RGB+S	88.16	80.55	75.13	10.58	15.92	20.14	N/A	N/A	N/A	N/A	N/A	N/A
Seg2vid [4]	RGB+S	88.32	N/A	61.63	9.69	N/A	25.99	N/A	N/A	N/A	N/A	N/A	N/A
FVS [18]	RGB+S+I	89.10	81.13	75.68	8.50	**12.98**	16.50	79.28	67.65	60.77	18.48	24.61	**30.49**
SADM [14]	RGB+S+F	**95.99**	N/A	**83.51**	**7.67**	N/A	**14.93**	**83.06**	**72.44**	**64.72**	**14.41**	**24.58**	31.16
PreNet [16]	RGB	84.03	79.25	75.21	25.99	29.99	36.03	56.26	51.47	47.56	55.35	58.66	62.95
MCNET [45]	RGB	89.69	78.07	70.58	18.88	31.34	37.34	75.35	63.52	55.48	24.05	31.71	37.39
DVF [1]	RGB	83.85	76.23	71.11	17.37	24.05	28.79	53.93	46.99	42.62	32.47	37.43	41.59
CorrWise [35]	RGB	92.80	N/A	83.90	8.50	N/A	15.00	82.00	N/A	66.70	17.20	N/A	25.90
OPT [22]	RGB	94.54	86.89	80.40	6.46	12.50	17.83	82.71	69.50	61.09	12.34	20.29	26.35
DMVFN [46]	RGB	95.73	89.24	83.45	5.58	10.47	14.82	88.53	78.01	70.52	10.74	19.27	26.05
MHMVF	RGB	**96.43**	**90.10**	**84.14**	**4.68**	**8.95**	**12.98**	**89.51**	**79.12**	**71.71**	**9.46**	**16.67**	**22.57**

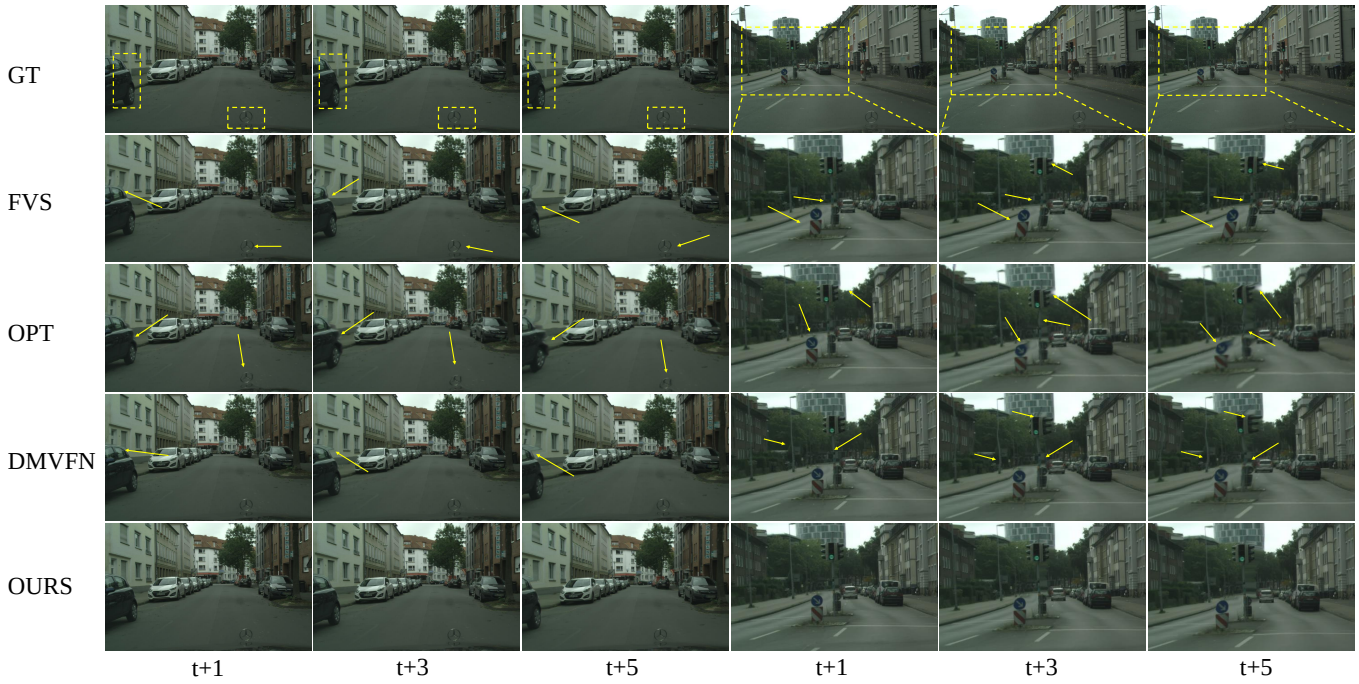

Fig. 2. Prediction comparison on the Cityscapes dataset.

and 4 are reserved for testing, ensuring a clear separation for evaluation purposes.

UCF101 [27] is a widely-used action recognition dataset comprising realistic action videos collected from YouTube, spanning 101 distinct action categories. The dataset consists of 13,320 videos across 101 action categories, offering the largest diversity in terms of actions and presenting significant challenges for action prediction tasks. These challenges arise from large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered backgrounds, and illumination conditions. As a result, UCF101 remains one of the most challenging datasets for action prediction research, providing an invaluable resource for advancing the field. In this work, we used the first subset for training and evaluation.

To ensure compatibility with prior studies and maintain consistency in experimental settings, we follow the resizing strategy adopted in FVS [18]. Video frames were sampled at 25 FPS following standard practice. Specifically, the images in the Cityscapes dataset are resized to 1024×512, while those in the KITTI dataset are resized to 832×256 and UCF101 dataset are resized to 256×256. This resizing facilitates efficient computation and ensures fair comparisons with existing methods, especially when evaluating performance across various benchmarks.

B. Implementation

The AdamW optimizer [47] was employed with a weight decay parameter of 10^{-3}. To regulate the learning rate, a

Fig. 3. Prediction comparison on the KITTI dataset.

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE UCF101 DATASETS.

Method	UCF101					
	MS-SSIM($\times 10^{-2}$)↑			LPIPS($\times 10^{-2}$)↓		
	t+1	t+3	t+5	t+1	t+3	t+5
DYAN	92.37	87.24	83.53	6.25	8.84	**10.83**
OPT	94.70	88.08	84.54	5.62	10.03	12.64
DMVFN	94.45	90.41	86.98	7.21	11.16	14.16
MHMVF	**94.91**	**90.84**	**87.30**	**5.39**	**8.70**	11.35

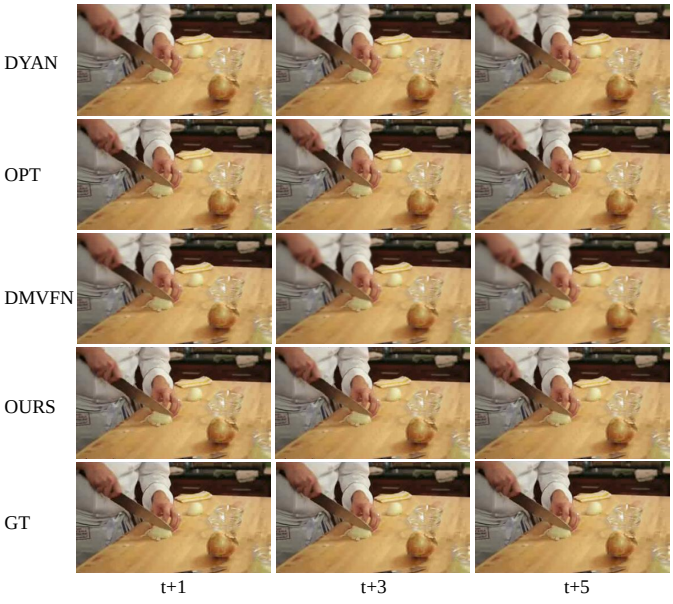

Fig. 4. Prediction comparison on the UCF101 dataset.

combination of linear warm-up and cosine annealing schedules was adopted, gradually adjusting it from 10^{-4} down to 10^{-5}. A batch size of 32 was used throughout the experiments. To increase data variability, input images were first randomly cropped to dimensions of 224×224 and then augmented using random rotations, alongside horizontal and vertical flipping techniques. The proposed method was trained for 300 epochs on all datasets. The implementation was carried out using the PyTorch framework, running on an Ubuntu 20.04 system, with both model training and evaluation executed on four NVIDIA A100 GPUs.

V. EXPERIMENTAL ANALYSIS

A. Results on the Cityscapes and KITTI datasets

The quantitative performance of our proposed Multi-Scale Hybrid Mamba Voxel Flow (MHMVF) framework, compared with various state-of-the-art methods, is presented in Table I on the Cityscapes and KITTI datasets. The compared methods are divided into two categories: those that rely solely on RGB images as input and those that utilize supplementary information such as semantic or instance maps. To evaluate model performance, we use multi-scale structural similarity (MS-SSIM) [48] and learned perceptual image patch similarity (LPIPS) [49]. Higher MS-SSIM scores and lower LPIPS values indicate better visual quality and perceptual consistency.

On the **Cityscapes** dataset, MHMVF achieves the best MS-SSIM scores across all evaluated time steps, recording 96.43 at $t+1$, 90.10 at $t+3$, and 84.14 at $t+5$. These results surpass those of DMVFN and semantic-guided SADM. In terms of LPIPS, MHMVF demonstrates its superiority with scores of 4.68, 8.95, and 12.98 at $t+1$, $t+3$, and $t+5$, respectively, significantly outperforming DMVFN. These results demonstrate that the multi-scale hybrid architecture of MHMVF, anchored by the global self-attention of the Mamba block and the

spatially adaptive local refinement of the Spatial Aggregation block, consistently preserves structure and enhances perceptual fidelity over long horizons. Specifically, at the coarsest scale, the Mamba block aggregates motion cues across the entire frame, capturing large displacements and global context with minimal overhead. At an intermediate scale, it blends these global flows with finer details. At the finest scale, the Spatial Aggregation block selectively emphasises salient edges and textures while suppressing artefacts. This progressive refinement strategy ensures that MS-SSIM remains high and LPIPS remains low, even at t+5, enabling the framework to robustly handle complex motion patterns, from rapid camera pans to subtle object deformations, in diverse urban scenes. On the **KITTI** dataset, MHMVF similarly outperforms all baseline methods. It achieves MS-SSIM scores of 89.51 at $t+1$, 79.12 at $t+3$, and 71.71 at $t+5$, surpassing DMVFN and OPT. MHMVF also achieves the lowest LPIPS values of 9.46, 16.67, and 22.57 at $t+1$, $t+3$, and $t+5$, demonstrating a significant improvement over DMVFN. In the KITTI scenes, which are characterised by rapid vehicle motion and varied lighting, MHMVF's coarse Mamba block first captures broad ego-motion and object displacements. Then, the fine-scale Spatial Aggregation block sharpens lane markings, car edges and pedestrian contours. This hierarchical refinement maintains high structural fidelity and suppresses blur at all time horizons, demonstrating robust temporal coherence under challenging real-world dynamics.

The comparative visualization results between our method and existing state-of-the-art approaches are presented in Fig. 2 and Fig. 3. For the **Cityscapes** dataset (Fig. 2), we selected two representative scenes containing both near-field and far-field variations. Yellow bounding boxes highlight regions prone to motion blur and distortion due to camera movement, with arrows pinpointing actual artifacts. As shown in Fig. 2, previous state-of-the-art methods are prone to deformation when predicting the rear side of the car on the left and the Mercedes-Benz logo. Our proposed Mamba Block can effectively solve the problem of difficult prediction of fast moving objects through global sensing field and significantly improve the accuracy of prediction.

The right side of Fig. 2 reveals magnified views of distant traffic signs and poles, where comparative methods (FVS, OPT, DMVFN) manifest varying degrees of warping and motion blur artifacts. In contrast, our model achieves superior shape preservation for distant objects with enhanced geometric accuracy. In the evaluation of the **KITTI** dataset (Fig. 3), we focused on dynamic scenes featuring high-speed vehicles and low-speed pedestrians to assess motion modeling capabilities. To facilitate detailed comparison, key areas are visualized using yellow bounding boxes and local magnification. Specifically, previous advanced methods exhibit noticeable motion artifacts in the vehicle contours and pose distortion in the rider's limb region. By modeling global information, our approach reliably maintains dynamic consistency for targets exhibiting diverse motion velocities.

In conclusion, the experiments confirm that MHMVF outperforms existing methods consistently across both the Cityscapes and KITTI benchmarks. It achieves an optimal balance between spatial fidelity and temporal consistency. MHMVF delivers superior frame quality and robust long-term prediction accuracy by combining coarse-to-fine progressive refinement with the complementary strengths of Mamba Blocks for global motion modelling, Spatial Aggregation Blocks for local detail preservation and SimAM Modules for adaptive feature weighting.

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE CITYSCAPES DATASETS.

METHOD	MS-SSIM($\times 10^{-2}$)↑			LPIPS($\times 10^{-2}$)↓		
	t+1	t+3	t+5	t+1	t+3	t+5
w/o SM	96.13	89.65	83.71	5.14	9.69	13.92
w/o MB	95.96	89.29	83.21	5.03	9.49	13.63
w/o SAB	90.21	81.99	75.64	8.87	15.59	21.00
MHMVF	**96.43**	**90.10**	**84.14**	**4.68**	**8.95**	**12.98**

B. Results on the UCF101 Dataset

Table II summarizes the performance of the proposed MHMVF method in comparison to three state-of-the-art approaches (DYAN [12], OPT [22], and DMVFN [46]) on the UCF101 dataset. MHMVF achieves the highest MS-SSIM scores across all prediction horizons ($t+1$, $t+3$, and $t+5$), with values of 94.91, 90.84, and 87.30, respectively. In terms of perceptual quality, MHMVF also demonstrates superior LPIPS performance. At $t+1$ and $t+3$, MHMVF achieves the lowest scores (5.39 and 8.70), outperforming all competing methods. For $t+5$, MHMVF obtains a score of 11.35, which is lower than OPT (12.64) and DMVFN (14.16) and comparable to DYAN's 10.83. These outcomes highlight how our progressive refinement framework — stacking multi-scale Hybrid Mamba Voxel Flow modules that combine global self-attention in the Mamba Block with local convolutional enhancement in the Spatial Aggregation Block and adaptive fusion by the SimAM module — robustly preserves both broad motion patterns and fine structural details, even over long-term prediction horizons. Fig. 4 presents a qualitative comparison on the UCF101 dataset, focusing specifically on hand joint movements. Our framework accurately reconstructs intricate hand motions, delivering sharper, more detailed predictions of finger articulations than DMVFN. Notably, MHMVF markedly reduces blurring and exhibits no visible distortions, demonstrating its superior ability to capture fine-grained dynamics.

Overall, the results on the UCF101 dataset demonstrate that MHMVF effectively addresses both structural and perceptual challenges in video prediction, achieving state-of-the-art performance across multiple time steps. The balanced improvements in MS-SSIM and LPIPS scores validate the effectiveness of our progressive multi-scale flow estimation and refinement strategy.

C. Ablation analysis and discussion

Table III provides quantitative results for the ablation study on the Cityscapes dataset, analyzing the contributions of

the Simam Module (SM), Mamba Block (MB), and Spatial Aggregation Block (SAB) within the MHMVF framework. Removing the Simam Module (w/o SM) results in a slight decrease in MS-SSIM, dropping from 96.43 to 96.13 at $t+1$, LPIPS scores increase from 4.68 to 5.14 at $t+1$. Excluding the Mamba Block (w/o MB) similarly impacts performance, reducing MS-SSIM to 95.96 at $t+1$ and 89.29 at $t+3$, while LPIPS values rise slightly, highlighting the importance of MB in capturing global motion through local self-attention. The absence of the Spatial Aggregation Block (w/o SAB) leads to the most significant performance drop, particularly at $t+5$, where MS-SSIM falls sharply to 75.64, and LPIPS increases dramatically to 21.0. These results demonstrate SAB's critical role in adaptively integrating features to maintain structural fidelity and perceptual quality

In contrast, the complete MHMVF framework achieves the best overall performance, with MS-SSIM scores of 96.43, 90.10, and 84.14 at $t+1$, $t+3$, and $t+5$, respectively, and LPIPS values of 4.68, 8.95, and 12.98. By synergistically leveraging the SimAM Module's adaptive feature weighting, the Mamba Block's global self-attention, and the Spatial Aggregation Block's localized fusion within our coarse-to-fine progressive pipeline, MHMVF consistently delivers superior MS-SSIM and LPIPS improvements across both short- and long-term horizons, demonstrating state-of-the-art predictive accuracy and perceptual fidelity.

As shown in Table IV, we conduct an ablation study on the Cityscapes dataset to investigate the role of the hyperparameters α and β in balancing short-term and long-term prediction performance. When $\alpha = 0$, the results clearly deteriorate for both MS-SSIM and LPIPS, demonstrating that relying solely on the branch governed by β is insufficient for capturing multi-scale motion information. Likewise, setting $\beta = 0$ provides only a small gain in MS-SSIM at $t+1$ but suffers from higher LPIPS values at $t+3$ and $t+5$, indicating that the absence of the β-controlled branch compromises perceptual quality. By contrast, both $\alpha = 1, \beta = 0.5$ and $\alpha = 1, \beta = 1$ yield notably better performance across all metrics; specifically, while $\alpha = 1, \beta = 0.5$ excels with a lower LPIPS at $t+1$, $\alpha = 1, \beta = 1$ shows a marginal edge at $t+5$. These findings suggest that an appropriate combination of α and β effectively captures global motion while also leveraging spatial detail compensation, thereby producing clearer and more structurally coherent predictions. Overall, we recommend $\alpha = 1, \beta = 0.5$ for a good trade-off between accuracy and perceptual quality.

Table V reports the latency and predictive accuracy of different video prediction methods evaluated on the Cityscapes dataset. Specifically, we compare the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) at future frames t + 1, t + 3 and t + 5, as well as the inference latency per frame in milliseconds (ms). As shown in the table V, our proposed method achieves the best performance in terms of both PSNR and SSIM across all prediction horizons. At frame t+1, our method achieves a PSNR of 30.52 and an SSIM of 90.93, outperforming both the DMVFN method and the MCVD method. This advantage persists at longer horizons, where it maintains relatively stable performance, demonstrating superior temporal consistency and robustness.

TABLE IV
RESULTS FOR DIFFERENT PROPORTIONS OF THE HYPERPARAMETERS α , β ON THE CITYSCAPES DATASET.

α	β	MS-SSIM($\times 10^{-2}$)↑			LPIPS($\times 10^{-2}$)↓		
		t+1	t+3	t+5	t+1	t+3	t+5
1	0	96.51	90.65	85.21	6.51	12.35	17.35
1	**0.5**	**96.43**	**90.10**	**84.14**	**4.68**	**8.95**	12.98
1	1	96.24	89.98	84.12	4.76	8.96	**12.83**
0.5	1	96.15	89.73	83.89	4.79	9.05	12.98
0	1	94.66	87.16	80.78	5.59	10.51	15.07

In addition to accuracy, our method is highly efficient, achieving a latency of only 60 ms per frame. This is markedly lower than the 327 ms latency incurred by MCVD and significantly lower than the 181 ms latency incurred by SimVP. Although DMVFN achieves the lowest latency (54 ms), our method offers a better balance between accuracy and speed, with substantial improvements in both PSNR and SSIM. These results clearly demonstrate the effectiveness of our architectural design in achieving high-quality predictions with low computational overhead, making our method a practical choice for real-time video prediction tasks.

TABLE V
LATENCY ANALYSIS OF DIFFERENT METHODS ON THE CITYSCAPES DATASET.

METHOD	PSNR↑			SSIM($\times 10^{-2}$)↑			Latency[ms]
	t+1	t+3	t+5	t+1	t+3	t+5	
SimVP	22.24	19.32	18.1	81.3	73.83	71.31	181
STNAM	22.38	18.28	16.37	82.75	72.85	68.94	201
MCVD	25.85	24.13	22.44	86.66	75.14	67.93	327
DMVFN	29.57	26.22	24.34	89.55	81.41	76.54	54
MHMVF	**30.52**	**26.97**	**24.95**	**90.93**	**82.69**	**77.47**	60

VI. CONCLUSION

In this paper, we introduced a novel multi-scale Hybrid Mamba Voxel Flow framework designed to address the challenges of blurry predictions and structural inconsistencies in video prediction tasks. By employing a progressive refinement strategy, our framework effectively models complex multi-scale motion between adjacent frames. The proposed Mamba Block leverages local self-attention to capture global motion information, while the Spatial Aggregation Block improves local detail prediction using a dual-branch residual structure. Additionally, the Simam Module adaptively integrates features from these blocks, significantly enhancing the overall predictive capability of the model. Experiments on Cityscapes, KITTI, and UCF101 datasets demonstrate that our method outperforms state-of-the-art techniques. Ablation studies further validate the effectiveness of each component in handling real-world complexities.By addressing challenges such as motion estimation and structural continuity, our framework contributes to the advancement of video prediction and its broader practical applications.

REFERENCES

[1] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4463–4471.

[2] X. Lin, Q. Zou, X. Xu, Y. Huang, and Y. Tian, "Motion-aware feature enhancement network for video prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 688–700, 2020.

[3] M. Li, L. Chen, J. Lu, J. Feng, and J. Zhou, "Order-constrained representation learning for instructional video prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5438–5452, 2022.

[4] J. Pan, C. Wang, X. Jia, J. Shao, L. Sheng, J. Yan, and X. Wang, "Video generation from single semantic label map," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, 2019, pp. 3728–3737.

[5] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, and H. Liu, "Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2133–2146, 2020.

[6] S. Yi, H. Li, and X. Wang, "Pedestrian behavior understanding and prediction with deep neural networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 263–279.

[7] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[8] X. Hu, W. Yin, M. Jia, J. Deng, X. Guo, Q. Zhang, X. Long, and P. Tan, "Drivingworld: Constructing world model for autonomous driving via video gpt," *arXiv preprint arXiv:2412.19505*, 2024.

[9] K. Zhang, Z. Tang, X. Hu, X. Pan, X. Guo, Y. Liu, J. Huang, L. Yuan, Q. Zhang, X.-X. Long *et al.*, "Epona: Autoregressive diffusion world model for autonomous driving," *arXiv preprint arXiv:2506.24113*, 2025.

[10] P. Wu, J. Liu, X. He, Y. Peng, P. Wang, and Y. Zhang, "Toward video anomaly retrieval from video anomaly detection: New benchmarks and model," *IEEE Transactions on Image Processing*, vol. 33, pp. 2213–2225, 2024.

[11] X. Sheng, L. Li, D. Liu, and H. Li, "Spatial decomposition and temporal fusion based inter prediction for learned video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6460–6473, 2024.

[12] W. Liu, A. Sharma, O. Camps, and M. Sznaier, "DYAN: A Dynamical Atoms-Based Network for Video Prediction," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, vol. 11216, pp. 175–191.

[13] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," International Conference on Learning Representations, ICLR, 2017.

[14] X. Bei, Y. Yang, and S. Soatto, "Learning semantic-aware dynamics for video prediction," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, 2021, pp. 902–912.

[15] W. Zhao, J. Chen, Z. Meng, D. Mao, R. Song, and W. Zhang, "Vlmpc: Vision-language model predictive control for robotic manipulation," *arXiv preprint arXiv:2407.09829*, 2024.

[16] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," *arXiv preprint arXiv:1605.08104*, 2016.

[17] J. Wang, W. Wang, and W. Gao, "Predicting diverse future frames with local transformation-guided masking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3531–3543, 2018.

[18] Y. Wu, R. Gao, J. Park, and Q. Chen, "Future video synthesis with object motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5539–5548.

[19] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[20] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep Learning-based Vehicle Behaviour Prediction For Autonomous Driving Applications: A Review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 33–47, 2022.

[21] Y. Yue, H. Qi, Y. Deng, J. Li, H. Liang, and J. Miao, "Infrastructure-side point cloud object detection via multi-frame aggregation and multi-scale fusion," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[22] Y. Wu, Q. Wen, and Q. Chen, "Optimizing video prediction via video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 814–17 823.

[23] X. Hu, Z. Huang, A. Huang, J. Xu, and S. Zhou, "A dynamic multi-scale voxel flow network for video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6121–6131.

[24] Y. Tang, P. Dong, Z. Tang, X. Chu, and J. Liang, "Vmrnn: Integrating vision mamba and lstm for efficient and accurate spatiotemporal forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5663–5673.

[25] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset," in *CVPR Workshop on the Future of Datasets in Vision*, vol. 2, 2015, p. 1.

[26] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[27] K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human action classes from videos in the wild," *Center for Research in Computer Vision*, vol. 2, no. 11, pp. 1–7, 2012.

[28] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," *arXiv preprint arXiv:1808.06601*, 2018.

[29] J. Pan, C. Wang, X. Jia, J. Shao, L. Sheng, J. Yan, and X. Wang, "Video generation from single semantic label map," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3733–3742.

[30] X. Bei, Y. Yang, and S. Soatto, "Learning semantic-aware dynamics for video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 902–912.

[31] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[32] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," *arXiv preprint arXiv:1605.08104*, 2016.

[33] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *International Conference on Learning Representations*, 2022.

[34] W. Liu, A. Sharma, O. Camps, and M. Sznaier, "Dyan: A dynamical atoms-based network for video prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 170–185.

[35] D. Geng, M. Hamilton, and A. Owens, "Comparing correspondences: Video prediction with correspondence-wise losses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3365–3376.

[36] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "Mcvd-masked conditional video diffusion for prediction, generation, and interpolation," *Advances in neural information processing systems*, vol. 35, pp. 23 371–23 385, 2022.

[37] Z. Zhang, J. Hu, W. Cheng, D. Paudel, and J. Yang, "Extdm: Distribution extrapolation diffusion model for video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 310–19 320.

[38] Y. Yuan and Z. Meng, "Stnam: A spatio-temporal non-autoregressive model for video prediction," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–7.

[39] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[40] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[41] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[42] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[43] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation," in *Proceedings*

of the IEEE/CVF international conference on computer vision, 2021, pp. 9772–9781.
[44] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li, "Flowformer: A transformer architecture for optical flow," in *European conference on computer vision*. Springer, 2022, pp. 668–685.
[45] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," *ICLR*, 2017.
[46] X. Hu, Z. Huang, A. Huang, J. Xu, and S. Zhou, "A dynamic multi-scale voxel flow network for video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6121–6131.
[47] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
[49] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

Muhao Xu received the M.S. degree in School of Information Science and Engineering, University of Jinan, Shandong, China in 2024. He is currently a Ph.D. degree candidate at the School of Mechanical Engineering, Shandong University, Shandong, China. His research interests include medical image processing, anomaly detection and multimodal information mining.

Baochen Fu received the M.S. degree from the School of Control Science and Engineering, Shandong University, Shandong, China, in 2024. He is currently a Ph.D.candidateattheSchoolofSoftware, Shandong University, Shandong, China. His research interests include computer vision, multimodal large models, and embodied AI.

Dongyu Liu received the B.S. degree in Control Science and Engineering from Shandong University, Shandong, China, in 2025. She is currently a Ph.D. student at the University of Texas at Dallas, Texas, USA. Her research interests include video prediction, deep learning, and optical imaging.

Wenzhi Deng is currently pursuing his B.S. degree in the School of Mechanical Engineering at Shandong University, with an expected graduation date of 2027. His research interests mainly focus on Mechatronic Engineering, the intersection of medicine and engineering, computer vision, etc.

Wei Yi received the Ph.D. degree in Mechanical Engineering and Automation from Shandong University, Jinan, China. He completed his postdoctoral research at Boston University, USA, where he served as a visiting scholar from 2018 to 2020. His work focused on the development and application of optomechanical systems for optical coherence tomography. He is currently Senior Experimentalist at Shandong University, with research interests in optical imaging and interdisciplinary applications of mechanical engineering.

Yi Wan received the Ph.D. degree from Shandong University, Jinan, China, in 2006. Since 2015, He is the Dean and Doctoral Supervisor of the School of Mechanical Engineering at Shandong University. His main research interests include Deep Learning, 3D printing, and control of robot. He is a project letter evaluation expert from the National Natural Science Foundation of China, as well as a peer reviewer for domestic and foreign journals such as International Journal of Advanced Manufacturing Technology, Materials Science Engineering C, Mechanical Systems and Signal Processing, Journal of South China University of Technology Natural Edition, and Journal of Southern Airlines.

Hua Wei received the B.Sc. degrees from Shandong University in 2016, and the Ph.D. degree from the University of Chinese Academy of Sciences in 2022, and engaged in post doctoral research in Institute of Software Chinese Academy of Sciences from 2022 to 2024. She is currently a laboratory technician at the School of Mechanical Engineering, Shandong University. Her research interests include computational optical imaging and image quality enhancement techniques, development of optical coherence tomography equipment, and intelligent analysis of image data.

Weiye Song received the Ph.D. degree from JiLin University, Changchun, China, in 2010. He is also a postdoctoral fellow at Boston University/Boston Medical Center and Harvard University/Massachusetts General Hospital. He is currently a professor at the School of Mechanical Engineering, Shandong University. His main research areas are opto electromechanical integrated medical equipment based on weak coherent optical imaging, spectroscopic analysis and other technologies, endoscopes, early medical diagnosis, intraoperative navigation equipment, etc.