MMVP: Motion-Matrix-based Video Prediction

USC Viterbi School of Engineering

Microsoft

Yiqi Zhong[1*†] Luming Liang[2*‡] Ilya Zharkov[2] Ulrich Neumann[1‡ 1]

[1]University of Southern California [2]Microsoft

[1]{yiqizhon, uneumann}@usc.edu [2]{lulian,zharkov}@microsoft.com

ICCV23

Introduction

- Video prediction has a multi-modal nature that needs to handle the low-level **appearance feature** and high-level **motion information** simultaneously;
- MMVP is a dual-stream video prediction pipeline. It decouples motion and appearance information by constructing **appearance-agnostic motion matrices**. The motion prediction module in MMVP is called matrix predictor, which takes motion matrices as the sole input. Later the predicted motion matrices and the appearance features will then be reunited through **matrix multiplication**.
- MMVP outperforms existing video prediction methods in both accuracy and efficiency, and largely reduces the model size.

Framework Overview

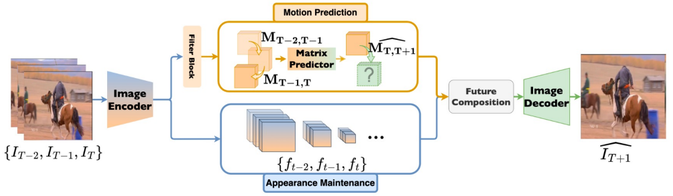
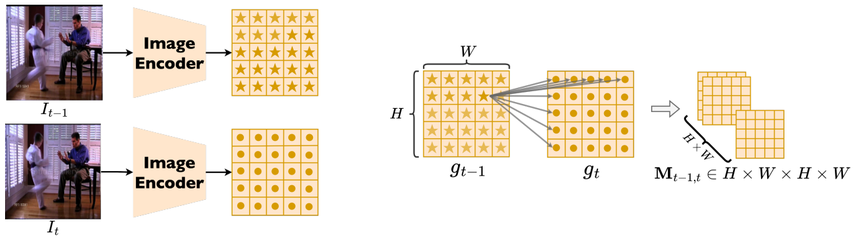


Figure 2: MMVP is a two-stream video prediction framework. It decouples motion prediction and appearance maintenance, and it reunites motion and appearance features through feature composition operation.

Core Components

Motion Matrix: Given two images I_p and I_q, an image encoder Θ encode the two consecutive images to a down-sample hidden space as image feature maps g_p and g_q. The temporal similarity matrix $\mathrm{M}_{p,q} \in \mathbb{R}^{H\times W\times H\times W}$ is defined as the pair-wise cosine similarity of the two feature maps, where H and W are the height and width of the feature map.



Multi-scale Future Composition: The future composition step generates information for the future frames using the observed information and the motion matrices. It is formulated as:

$$\widehat{\mathcal{X}_{T+j}} = \sum_{i=1}^{T} (\mathcal{X}_i \times \prod_{n=i}^{T-1} \mathbf{M}_{n,n+1} \times \widehat{\mathbf{M}_{T,T+j}}).$$

The χ in the equation represents the observed information of the past frames. The information can be the output features of the image encoder with different scales $f_i \in \mathbb{R}^{H_s\times W_s\times C}$, where C is the feature length; the information can also be the observed frames $I_i \in \mathbb{R}^{H\times W\times 3}$.

Evaluation Results

Table 6: Ablation study on sources for future composition and the comparison with other SOTA methods on UCF Sports.

Method	Composition source					Full set			Easy (SSIM ≥ 0.9)			Intermediate (0.6 ≤ SSIM < 0.9)			Hard (SSIM < 0.6)			Param#
	Img	1	1/2	1/8	1/16	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	
STIPHR [4]			-			0.8817	28.17	0.1626	0.9491	30.65	0.1066	0.8351	23.97	0.2271	0.4673	15.97	0.4450	18.05M
SimVP [12]			-			0.9189	29.97	0.1326	0.9664	32.87	0.0584	0.8845	25.79	0.1951	0.6267	18.99	0.5600	3.47M
MMVP	×	×	×	✓	✓	0.9000	28.31	0.1874	0.9375	30.43	0.1342	0.8759	25.36	0.2304	0.6593	19.90	0.4992	2.75M
	×	×	✓	✓	✓	0.9284	30.14	0.1115	0.9667	32.79	0.0603	0.8937	26.11	0.1693	0.7159	20.71	0.3570	2.79M
	×	✓	✓	✓	✓	0.9296	30.22	0.1064	0.9669	32.87	0.0576	0.8965	26.26	0.1571	0.7199	20.76	0.3555	2.80M
	✓	✓	✓	✓	✓	0.9296	30.29	**0.1051**	**0.9675**	32.99	**0.0567**	0.8958	26.22	**0.1554**	0.7175	20.76	0.3517	2.80M
	✓	×	✓	✓	✓	**0.9300**	**30.35**	0.1062	0.9674	**33.05**	0.0580	**0.8970**	**26.29**	0.1569	**0.7203**	**20.84**	**0.3510**	2.79M

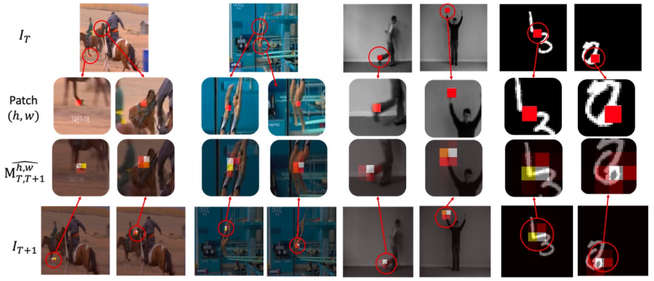


Figure 5: Predicted motion matrix visualization. We highlight the selected feature patch(es) at (h, w) in the last observed frame I_T in red and visualize their corresponding predicted motion matrices $\mathbf{M}^{h,w}_{T,T+1} \in \mathbb{R}^{H\times W}$ overlaying with the first future frame I_{T+1}. A brighter color indicates a higher predicted value. We select two samples for each dataset. From left to right, samples originate in the validation set of UCF Sports, KTH, and Moving MNIST.

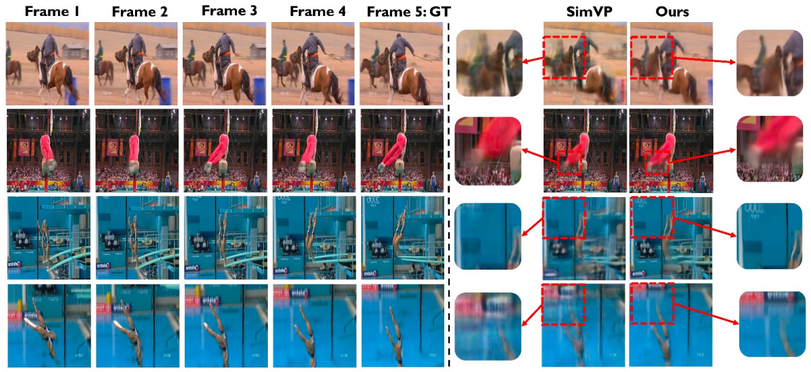


Figure 6: Qualitative results on our own splits of the UCF Sports dataset.