Video Infilling with Rich Motion Prior

Xinyu Hou[1]
xinyu.hou@ntu.edu.sg
Liming Jiang[1]
liming002@ntu.edu.sg
Rui Shao[2]
shaorui@hit.edu.cn
Chen Change Loy[1]
ccloy@ntu.edu.sg

[1] S-Lab
Nanyang Technological University
Singapore

[2] School of Computer Science and Technology
Harbin Institute of Technology (Shenzhen)
Shenzhen, China

Abstract

Video infilling is a task of generating visually smooth and plausible intermediate frames in between given context frames. The infilling interval is usually large, and thus the intermediate contents to be filled experience significant and non-uniform changes in motion. To handle this challenging task, it is required for the model to learn robust motion dynamics to synthesize rich and plausible motion trajectories in between given contexts. In this work, we demonstrate the possibility of learning rich motion prior for video infilling via masked motion modeling. Our key insight is that the powerful ability of masked autoencoder to capture long-range dependencies could help us model and therefore generate rich and realistic in-between motions. Unlike previous multi-scale optical flow-based video interpolation methods, our framework is simple yet effective in longer-interval and larger-motion cases. In particular, we use the optical flow tokens learned by a pre-trained discrete tokenizer as the reconstruction target in masked motion modeling. With a random masking ratio over 0.5 during training, reasonable intermediate optical flows can be predicted by iterative decoding during inference. To demonstrate pixel-level infilling results, a dedicated bi-directional fusion of the warping results is applied. Through experiments conducted on the human action dataset, we demonstrate the effectiveness of our approach in predicting valid and diverse motions between given contexts. Quantitative results of pixel-level evaluation metrics show that our approach can outperform previous state-of-the-art methods even with the naïve fusion results.

1 Introduction

Video frame interpolation (VFI), given its broad array of applications in video enhancement, has garnered significant attention in computer vision. Most existing studies concentrate on interpolating uniform motion between consecutive frames. However, a more complex scenario, termed video infilling, remains somewhat uncharted. This issue presents a greater challenge than VFI as it necessitates interpolation across extended time intervals and wider motion gaps.

The primary distinction between video interpolation and video infilling stems from their respective objectives and input settings. Video interpolation algorithms are commonly employed to augment the video frame rate or produce slow-motion video playback. As a result, the inputs to such algorithms generally consist of adjacent frames of an existing video that exhibits high visual similarity and limited movements of objects. On the other hand, video infilling aims at filling in large temporal gaps between distant frames in a video, where significant scene changes can be captured.

Previous attempts have been made to tackle the issue of large-motion video interpolation [19][18], utilizing multi-scale optical flows in synthesizing valid intermediate frames. However, since the addressed motion range in these approaches still remains limited to extreme cases of motion between adjacent frames, the motion range that is feasible in these settings is not comparable to that required for video infilling tasks. More importantly, previous video interpolation approaches have generally relied on a uniform assumption regarding the intermediate motion between consecutive frames, where the objects move along a straight line at a constant speed. However, motions in real-world scenarios are diverse and can hardly meet the uniform assumption, especially in long temporal intervals. In Xu *et al*. [25], the issue of non-uniform motion was addressed for the first time in the video interpolation setting. This approach leveraged the acceleration information, enabling the prediction of curvilinear trajectories and variable velocities, which leads to more accurate interpolation results. Nevertheless, despite the attention given to the non-uniform motion, this method is unable to address the issue of large motion in the presence of distant frames, owing to the absence of a robust mechanism to capture and model complex motion prior.

To address the aforementioned challenges, we propose to learn rich motion prior from masked motion modeling to perform large and non-uniform motion video infilling. Inspired by Xu *et al*. [25], we exploit the acceleration information in context frames and model the long-range correlation of motion changes by a masked autoencoder (MAE) [8]. Our method follows a two-stage scheme: We first train a discrete vector quantizer in the optical flow space, which provides us with succinct representations to capture optical flow patterns. Then, an MAE is applied to reconstruct the masked discrete token indices extracted by the vector quantizer. We believe that the strong ability of MAE in modeling long-range dependencies can help us learn rich and plausible motion prior from real-world video sequences. Note that previous attempts have been made to address video completion via masked visual modeling [7]. However, since their modeling of videos remains in the pixel domain, the generated motions are rather uncanny. Instead, our approach of applying masked modeling on motion representation, optical flow specifically, allows the model to learn more smooth and realistic motions. The experimental results (Sec. 4) showcase the effectiveness of our proposed method in terms of both qualitative and quantitative performances.

Our contributions can be summarised as follows: 1) We use a pure reconstruction-based self-supervised learning framework to achieve longer-range and larger-motion video infilling; 2) We propose masked motion modeling to better model complex and non-uniform motion prior in the real world; 3) We introduce optical flows as explicit motion representations to capture rich and realistic motion, which serves as an alternative to using pixel reconstruction solely.

2 Related Work

Video Infilling. Video infilling has been addressed in several previous studies. To the best

of our knowledge, Xu *et al*. [24] is the first work that raises the problem. It formulates the infilling problem as a bi-directional constraint stochastic generation process and proposes an approach based on recurrent neural networks (RNN). Li *et al*. [13] argue that the problem can be effectively solved by fully convolutional neural networks (CNN). They learn a latent video representation by progressively up-sampling the context frame embeddings along the temporal dimension and directly decode the latent video representation into the output video. More recent works leverage the powerful tool of generative diffusion models [10, 23] to accomplish the task. Additionally, although Gupta *et al*. [7] primarily focus on the video prediction task, its ability to perform video infilling is also tested through real robot experiments. It is worth noticing that all previous video-infilling approaches model the representations of intermediate contents in the pixel domain.

Video Interpolation. Optical flows and related concepts have been utilized to synthesize new frames in video frame interpolation for a long time [11, 15, 16, 17]. Recent works in video interpolation have also tried to tackle more challenging scenarios like large motion and non-uniform motion. In particular, Sim *et al*. [19] and Reda *et al*. [18] exploit multi-scale structures to estimate intermediate optical flows in a coarse to fine manner. Xu *et al*. [25] and Liu *et al*. [14] employ higher-order acceleration information to estimate non-uniform motion patterns. However, these methods remain coping with a limited motion range and cannot learn meaningful motion trajectories between distant contexts.

Masked Modeling. Masked auto-encoders [8] have demonstrated their effectiveness as robust self-supervised representation learners, owing to the powerful global relation modeling ability of visual transformers. Through the use of various reconstruction targets, such as 3D video patches or motion trajectories [22] [21], masked auto-encoders are able to learn different representations from real-world distribution. The mask-and-reconstruct scheme has also been applied to generative tasks following a two-stage paradigm. First, a vector tokenizer and a discrete codebook are learned to effectively compress the data. Then, a valid composition of the quantized vectors is learned via masked modeling. Finally, realistic data following natural distribution can be generated by decompressing the vectors [8] [12] [7]. In this work, we make the first attempt to apply mask motion modeling on optical flows to generate plausible in-between motions.

3 Methodology

The video infilling task can be formulated as follows: given context frames $I_0, I_1, I_{T-2}, I_{T-1}$, we synthesize intermediate frames $\hat{I}_t$, $t \in [2,(T-3)]$ under the same frame rate. Our approach follows a conventional VAE-based two-stage synthesizing paradigm. During the tokenization stage, we train a VQ-GAN [4] on optical flows extracted from adjacent frames in both forward ($t \rightarrow t+1$) and backward ($t+1 \rightarrow t$) directions (Sec. 3.1). In the prior prediction stage, we use an MAE to perform masked motion modeling (Sec. 3.2). Finally, to better illustrate the effectiveness of our learned motion prior in the pixel domain, a non-learnable stage that naïvely fuses bi-directional warping results is deployed (Sec. 3.3). Note that since our work focuses on learning valid motion prior, we leave incorporating trainable fusion mechanisms to enhance the visual quality of the final intermediate frame predictions for future work. The entire process of our framework is illustrated in Fig. 1.

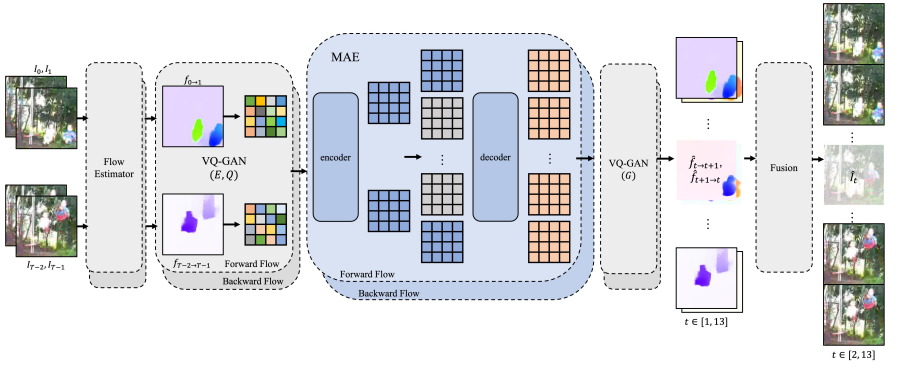

Figure 1: Framework of our proposed intermediate motion predictions via masked motion modeling. Inside the MAE block, the blue squares indicate context tokens, the gray squares are mask tokens, and the orange squares are decoded tokens. The gray background indicates frozen parts in the framework. Backward flows follow the same framework.

3.1 Optical Flow VQ-GAN

Vector Quantized Generative Adversarial Network (VQ-GAN) [2] has showcased its proficiency in synthesizing visually superior images. Vector quantization is an effective way of compressing data while preserving crucial details. To synthesize high-quality optical flows, we train a VQ-GAN in the optical flow domain. Given a single optical flow $f \in \mathbb{R}^{H\times W\times 2}$ of either forward or backward direction, an encoder E partitions and encodes it into non-overlapping patches $P = \{p_0, p_1, \cdots, p_N\} \in \mathbb{R}^{N\times D}$, where N denotes the number of patches and D is the embedding dimension. A quantizer Q then maps each patch to its nearest codebook entry. Meanwhile, the motion codebook $C = \{c_0, c_1, \cdots, c_K\} \in \mathbb{R}^{K\times D}$, with K being the codebook size, is updated accordingly throughout training. Finally, a decoder G is deployed to reconstruct realistic optical flows from the quantized vectors, denoting as $\hat{f} = G(c_q)$. Since the nearest-neighbor finding process is non-differentiable, we adopt the straight-through gradient estimator [1, 2] to directly copy the gradients across the quantization process for propagation. Similar to previous works [2, 4], the vector quantization loss is:

$$L_{VQ} = L_{rec} + \|\text{sg}[E(f)] - c_q\|_2^2 + \|\text{sg}[c_q] - E(f)\|_2^2, \tag{1}$$

where L_{rec} denotes L_1 loss for reconstruction, and sg$[\cdot]$ indicates stop-gradient operation. Adversarial loss is also applied for better visual quality of the reconstruction results:

$$L_G = \log(\text{Disc}(f)) + \log(1 - \text{Disc}(\hat{f})). \tag{2}$$

where Disc is a discriminator for adversarial training. The overall loss used for VQ-GAN training can be expressed as follows:

$$L_{total} = \min_{E,G,C} \max_{\text{Disc}} \mathbb{E}[L_{VQ} + \lambda L_G]. \tag{3}$$

To enhance the stability of the training procedure, the adversarial loss is only incorporated after a pre-defined number of steps. Besides, the weight parameter λ is learned adaptively during training.

3.2 Masked Motion Modeling with Optical Flow

Inspired by the previous success of masked visual modeling in image and video generations, we introduce masked motion modeling (MMM) to capture and model motion prior. Specifically, during training, we sample T consecutive frames from a video, denoted as $I = \{I_0, I_1, \cdots, I_{T-2}, I_{T-1}\} \in \mathbb{R}^{T\times H\times W\times 3}$. A pre-trained FlowFormer [9] is employed to extract optical flows in both directions from each pair of adjacent frames, yielding $(T-1)$ optical flows in each direction. Note that we will specifically describe the MMM framework for the forward flow direction in this section – the backward flow direction follows the same procedure. In the forward direction, the extracted optical flows can be denoted as $F = \{f_{0\to1}, f_{1\to2}, \cdots, f_{T-2\to T-1}\} \in \mathbb{R}^{(T-1)\times H\times W\times 2}$. Next, each optical flow is fed into the pre-trained VQ-GAN (described in Sec. 3.1) individually, and the nearest codebook entry indices of all tokens $Y = [y_i]_{i=1}^{(T-1)\times N}$ are obtained. To perform the video infilling task, we design an infilling mask $M \in \mathbb{R}^{(T-1)\times N\times 1}$ that always keeps the tokens from the first and last optical flows unmasked (blue squares in Fig. 1 MAE block), and masks the remaining tokens with a masking ratio $\in [0.5, 1)$ (gray squares in Fig. 1 MAE block). The training objective is to reconstruct the masked token conditioned on the unmasked ones. A cross-entropy loss is used to train the reconstruction task:

$$L_{MMM} = -\mathop{\mathbb{E}}_{Y\in trainset}\Big(\sum_{\forall i\in masked} \log p(y_i|Y_M)\Big), \tag{4}$$

where Y_M denotes the unmasked tokens from a video sequence. By training with varying masking ratios, iterative predictions are enabled in the inference stage. In particular, during inference, the infilling mask is initialized by masking all tokens except the context ones. A predetermined number of iterations τ is set. The number of tokens to be kept in each iteration χ_t is calculated by a cosine function. In each iteration, we sample a confidence score for every unmasked token from the predicted possibility distribution of it belonging to each pretrained codebook entry, but only the top χ_t tokens are kept and added as context tokens for predictions in future iterations. The mask is also updated accordingly at the end of every iteration. Finally, after all iterations, the entire set of intermediate tokens is filled and the reconstructed optical flows can be obtained by the VQ-GAN decoder.

3.3 Bi-directional Fusion

To illustrate the learned motion prior, we fuse the bi-directional frame-level warping results with a pre-determined non-learnable mechanism. With the predicted bi-directional optical flows $\hat{f}_{fwd}$ and $\hat{f}_{bwd} \in \mathbb{R}^{(T-1)\times H\times W\times 2}$, we assess the validity of bi-directional flows by a cycle consistency check proposed in Gao *et al.* [5]:

$$\varepsilon_{i\to j}(p) = \|\hat{f}_{i\to j}(p) + \hat{f}_{j\to i}(p + \hat{f}_{i\to j}(p))\|_2^2, \tag{5}$$

where i, j are frame indexes, and p is pixel position on the corresponding optical flow. During frame warping of either direction, only consistent regions with $\varepsilon(p) < \theta$ are warped, where θ is a pre-determined hyperparameter. The resultant consistent bi-directional warping results, denoted as $\hat{I}^c_{fwd}$ and $\hat{I}^c_{bwd} \in \mathbb{R}^{T\times H\times W\times 3}$, only contain consistent regions in every frame. To enhance the incomplete consistent warping results, we further obtain the current mask regions μ_{fwd} and $\mu_{bwd} \in \mathbb{R}^{T\times H\times W\times 1}$, and the raw warping results without consistency check $\hat{I}^r_{fwd}$ and $\hat{I}^r_{bwd} \in \mathbb{R}^{T\times H\times W\times 3}$. Subsequently, the final pixel values are approximated

by employing a carefully designed combination of consistent and raw warping results. The detailed equations of the fusion mechanism can be found in the supplementary material.

4 Experiments

4.1 Experimental Setups

Dataset. All experiments are conducted on the UCF101 human action dataset [20] since it contains real-world sequences with articulated non-uniform and complex motions. To ensure consistency, we strictly follow the train/test split provided on the official website of UCF101. The training set consists of 9537 videos and the test set consists of 3783 videos. Sequences from the training set are used for training both VQ-GAN and MMM models. Consecutive $T = 16$ frames are randomly sampled from each video sequence for both training and inference. The input size for both VQ-GAN and MMM is (224×224).

Network Architecture. For VQ-GAN encoder, we use latent shape of $(16 \times 16 \times 1)$, i.e., each (224×224) frame is encoded to $N = 14 \times 14$ patches. The codebook has a total size of $K = 1024$ with embedding dimension $D = 256$. The MAE encoder comprises 12 transformer blocks and the embedding dimension of the encoder is 768. The decoder consists of 4 blocks with 6 heads, and the embedding dimension is set to 384.

Implementation Details. The VQ-GAN is trained with a learning rate of 1.8×10^{-5} for 1.5×10^6 steps. The adversarial loss is added after 3×10^4 steps. As for MAE, the network is trained for 500 epochs with a base learning rate of 2.3×10^{-5}. The *AdamW* optimizer is employed with $betas = [0.9, 0.95]$. During inference, iterative decoding of 20 iterations is used for all experiments. The consistency check threshold θ is set to 5.

Metrics. For quantitative evaluation, Peak Signal-to-noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity(LPIPS) are adopted to evaluate the generated intermediate contents by frame following video interpolation algorithms. In addition, Frechét Video Distance (FVD) is used to evaluate the generated video as a whole for its quality in correlation with human perception.

4.2 Qualitative Results

The qualitative performance of our proposed method is shown in Fig. 2. As can be observed, our method can generate sharp and crisp intermediate optical flows that comply with given contexts. In addition, in the lower case of Fig. 2, though some ground truth optical flows fail to provide any motion information due to the limitation of the flow estimation technique, our model can still successfully model the in-between motions with valid contexts. Furthermore, since our masked modeling-based model does not require any domain-specific knowledge, it is able to generalize to data of any domain without being trained on it. Inference results of our UCF-101 trained model on the KITTI dataset [6] are illustrated in Fig. 3 to showcase such generalizability. More results are provided in the supplementary material.

4.3 Comparison with Recent Approaches

We compare our method with two previous approaches that specifically address the large-motion and non-linear motion challenges in the video interpolation tasks, respectively. 1)

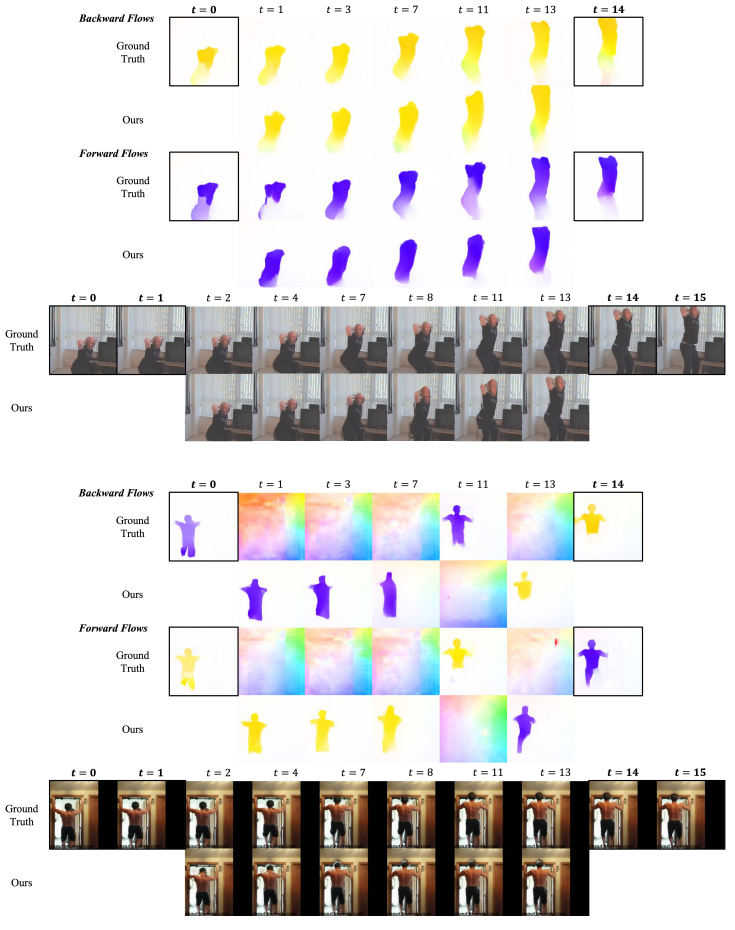

Figure 2: Qualitative evaluation of our method. Bi-direction optical flows are predicted by iterative decoding for 20 iterations. Pixel-level infilling results are obtained from non-learnable naïve fusion. Context flows and frames are marked with black outlines. At time t, forward flow denotes $f_{t\to t+1}$ and backward flow denotes $f_{t+1\to t}$.

FILM [18] uses a scale-agnostic feature pyramid to predict implicit flow residuals at multiple levels. The coarse-to-fine feature pyramid helps handle large motion gaps since large motion at a fine level is equivalent to small motion at a coarse level. FILM interpolates the center frame given two input frames as contexts, which is different from our 4-frame context settings. To ensure ground truth is available for the center interpolated frame while retaining a comparable motion range with ours, we adjust the inference setting of FILM to interpolate between frames I_1 and I_{13}, with target frame I_7. $\hat{I}_4$ and $\hat{I}_{10}$ can also be predicted with $(I_1, \hat{I}_7, I_{13})$; 2) QVI [25] takes in frame I_0, I_1, I_2, I_3 and calculates the accelerations and velocities at $t = 1, 2$ by substituting $(f_{1\to 0}, f_{1\to 2})$ and $(f_{2\to 1}, f_{2\to 3})$ into Equ. 6.

$$f_{0\to t} = \frac{1}{2}at^2 + v_0 t. \tag{6}$$

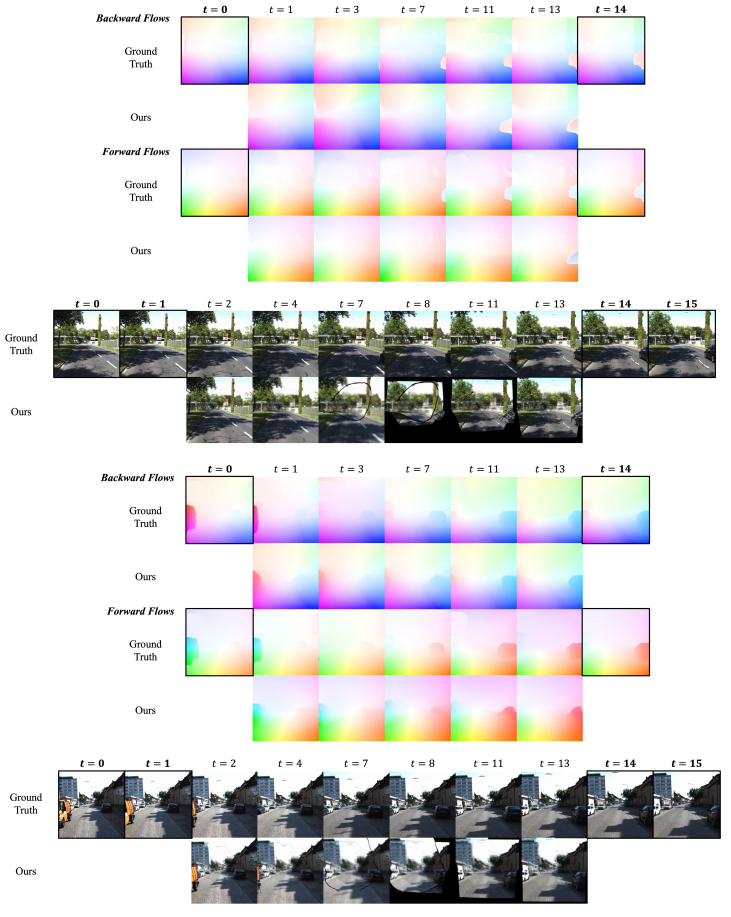

Figure 3: Cross-dataset qualitative evaluation of our method on the KITTI dataset [6]. Bi-direction optical flows are predicted by iterative decoding for 20 iterations. Pixel-level infilling results are obtained from non-learnable naïve fusion. Context flows and frames are marked with black outlines. At time t, forward flow denotes $f_{t\to t+1}$ and backward flow denotes $f_{t+1\to t}$.

The acceleration information is used to estimate the optical flows $f_{1\to t}, f_{2\to t}$ to the target intermediate frame, which are later used to predict the reversal optical flows $f_{t\to 1}, f_{t\to 2}$. To test the performance of QVI in long-range interpolation, we replace the adjacent 4 frame inputs by I_0, I_1, I_{14}, I_{15}. Likewise, we calculate the accelerations and velocities at $t = 1, 14$ following the same constant-acceleration assumption as [25]. The algorithm allows multiple intermediate frames to be generated with different t values. Therefore, we set t to twelve equally spaced time steps between $(1, 14)$, leading to the same configuration as our method. Since both previous works are trained on large-scale datasets and the pre-trained models are generalizable, we directly utilize the pre-trained weights of their methods in our inference setting.

The quantitative results are reported in Tab. 1. The reported numerical results are aver-

Table 1: Quantitative comparison with previous works on the UCF101[20] test set.

Method	By Frame			By Video
	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓
FILM [18]	0.7571	22.85	**0.1244**	-
QVI [25]	0.7254	22.13	0.1281	1206.89
Ours	**0.8051**	**22.97**	0.1310	**1104.69**

aged across five separate runs of inference. Note that FILM [18] is omitted in FVD comparison due to a different number of frames infilled per video, which leads to unfair per-video metric comparisons. As we can see, our method achieves poorer LPIPS scores than the two previous works. However, we think that they likely score better in LPIPS because the fading and re-appearing of objects in their generated video align well with perceptual evaluation when being assessed individually in each frame. On the contrary, our approach achieves a significantly lower FVD score, demonstrating its effectiveness in generating videos that align with human judgment of visual quality. To further demonstrate the superiority of our method with respect to two previous works, the qualitative comparison is illustrated in Fig. 4. Observably, in the upper case, both QVI and FILM fail to generate valid intermediate frames, resulting in the fading and re-appearing of the objects. In contrast, our method predicts a smooth motion trajectory throughout the video sequence. In addition, in the lower case of a man doing a push-up where the context frames are visually similar. Our method is the only one that manages to restore the upward-then-downward in-between motions. Such com-

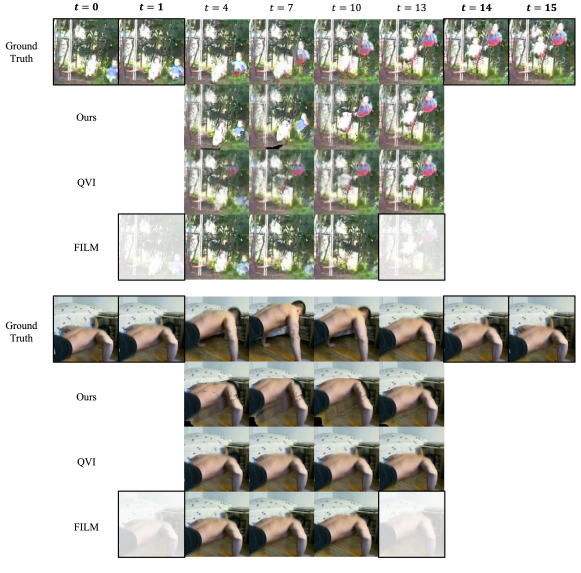

Figure 4: Qualitative comparison with previous works. As mentioned in Sec. 4.3, context frames at $t = 0, 1, 14, 15$ are used for both our method and QVI [25]. For FILM [18], frames I_1, I_{13} are context frames for prediction. All context frames are marked with black outlines.

pelling outcomes prove that masked motion modeling on optical flows can indeed help us learn rich motion prior and model large and non-uniform intermediate motion.

4.4 Ablation Study

To test the effectiveness of VQ-GAN in synthesizing high-quality optical flows, we conduct an ablation study on reconstructing the optical flow itself instead of VQ-GAN token indices. L_1 loss is used for the reconstruction task. Due to computational resource limitation, we conduct the ablation study on a small set of 4 classes out of 101 classes from the UCF101 [20] dataset. Our full model is trained on the same subset for a fair comparison. Predicted optical flows are illustrated in Fig. 5. As we can see, though the framework without VQ-GAN can learn valid motion trends (revealed by the color of the predicted optical flows), it fails to generate crisp results. We infer that the mosaic effect is an intrinsic characteristic of MAE reconstruction outputs since the original MAE is employed for representation learning.

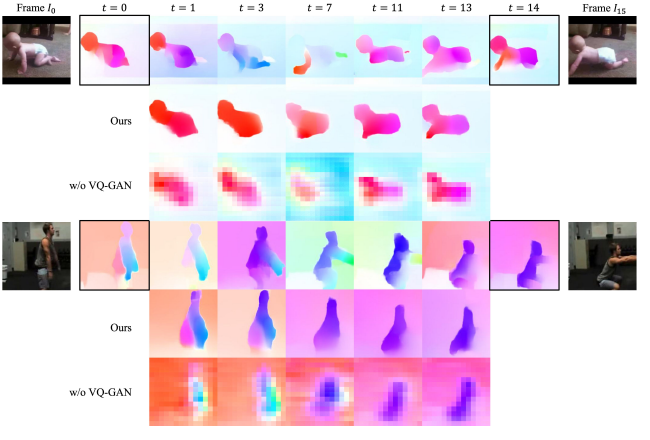

Figure 5: Ablation Study: Qualitative comparison with the framework without VQ-GAN. For each video sequence, the top row shows the ground truth optical flow. The two at $t = 0, 14$ with a black outline are the context optical flows used for prediction. Only backward flows are shown in this comparison.

5 Conclusion

In this work, we present an approach to learn motion correlations from natural videos using a mask-and-reconstruct scheme. To synthesize high-quality motion representation, we leverage a pre-trained VQ-GAN in the optical flow domain. In addition, mask motion modeling is used to capture rich motion prior embedded in reasonable compositions of the quantized vectors. Experimental results demonstrate that our method is able to learn rich and diverse motion prior of various real-world actions. Even with a simple non-learnable fusion of warping results, it achieves better numerical performance than previous methods. However, we believe that the performance can be further boosted with a learnable refinement network to generate visually better intermediate frames, which we leave as potential future work.

Acknowledgement

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[2] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *European Conference on Computer Vision (ECCV)*, 2022.

[3] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *European Conference on Computer Vision (ECCV)*, 2020.

[6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. In *International Journal of Robotics Research (IJRR)*, 2013.

[7] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. MaskViT: Masked visual pre-training for video prediction. In *International Conference on Learning Representations (ICLR)*, 2022.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[9] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A transformer architecture for optical flow. In *European Conference on Computer Vision (ECCV)*, 2022.

[10] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. In *Transactions on Machine Learning Research (TMLR)*, 2022.

[11] Huaizu Jiang, Deqing Sun, Varan Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. MAGE: Masked generative encoder to unify representation learning and image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] Yunpeng Li, Dominik Roblek, and Marco Tagliasacchi. From here to there: Video inbetweening using direct 3D convolutions. *arXiv preprint arXiv:1905.10240*, 2019.

[14] Yihao Liu, Liangbin Xie, Li Siyao, Wenxiu Sun, Yu Qiao, and Chao Dong. Enhanced quadratic video interpolation. In *European Conference on Computer Vision Workshop (ECCVW)*, 2020.

[15] Ziwei Liu, Raymond A. Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. FILM: Frame interpolation for large motion. In *European Conference on Computer Vision (ECCV)*, 2022.

[19] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. XVFI: extreme video frame interpolation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[20] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[21] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H. Li, Mingkui Tan, and Chuang Gan. Masked motion encoding for self-supervised video representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[22] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[23] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[24] Qiangeng Xu, Hanwang Zhang, Weiyue Wang, Peter N. Belhumeur, and Ulrich Neumann. Stochastic dynamics for video infilling. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[25] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.