Exploring Spatial-Temporal Multi-Frequency Analysis for High-Fidelity and Temporal-Consistency Video Prediction

Beibei Jin[1,2], Yu Hu *[1,2], Qiankun Tang[1,2], Jingyu Niu[1,2], Zhiping Shi[3], Yinhe Han[1,2], and Xiaowei Li[1,2]

[1]Research Center for Intelligent Computing Systems, State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{jinbeibei, huyu, tangqiankun, niujingyu17b, yinhes, lxw}@ict.ac.cn
[3]Capital Normal University
shizp@cnu.edu.cn

Abstract

Video prediction is a pixel-wise dense prediction task to infer future frames based on past frames. Missing appearance details and motion blur are still two major problems for current models, leading to image distortion and temporal inconsistency. We point out the necessity of exploring multi-frequency analysis to deal with the two problems. Inspired by the frequency band decomposition characteristic of Human Vision System (HVS), we propose a video prediction network based on multi-level wavelet analysis to uniformly deal with spatial and temporal information. Specifically, multi-level spatial discrete wavelet transform decomposes each video frame into anisotropic sub-bands with multiple frequencies, helping to enrich structural information and reserve fine details. On the other hand, multi-level temporal discrete wavelet transform which operates on time axis decomposes the frame sequence into sub-band groups of different frequencies to accurately capture multi-frequency motions under a fixed frame rate. Extensive experiments on diverse datasets demonstrate that our model shows significant improvements on fidelity and temporal consistency over the state-of-the-art works. Source code and videos are available at https://github.com/Bei-Jin/STMFANet.

*Corresponding author: Yu Hu, huyu@ict.ac.cn. This work is supported in part by the National Key RD Program of China under grant No. 2018AAA0102701, in part by the Science and Technology on Space Intelligent Control Laboratory under grant No. HTKJ2019KL502003, and in part by the Innovation Project of Institute of Computing Technology, Chinese Academy of Sciences under grant No. 20186090.

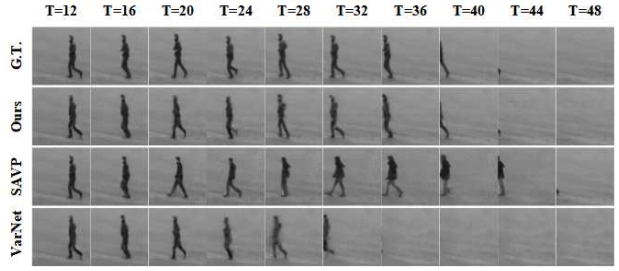

Figure 1. A comparison of long-term prediction on a KTH [34] motion sequence. Our model generates predictions with higher fidelity and temporal consistency than the state-of-the-art methods, SAVP [22] and VarNet [19]. In the other two methods' predictions, the person gradually blurs to distortion and runs out of the image too fast or too slowly, which is inconsistent to the ground truth.

1. Introduction

Unsupervised video prediction has attracted more and more attention in the research community and AI companies. It aims at predicting upcoming future frames based on the observation of previous frames. This looking-ahead ability has a broad application prospect on video surveillance [11], robotic systems [12] and autonomous vehicles [48]. However, building an accurate predictive model still remains challenging because it requires to master not only the visual abstraction model of different objects but also the evolution of various motions over time. Many recent deep learning methods [22, 47, 36, 3, 40, 39, 44, 21] have brought about great development on the video prediction task. However, there still exists a clear gap between their predictions and the ground-truth (GT), as shown in Figure 1. The predictions of the compared methods suffer from deficient retention of high-frequency details and insuf-

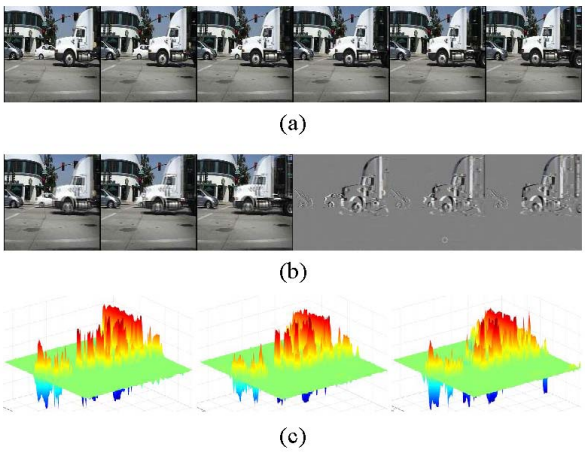

Figure 2. Discrete Wavelet Transform (DWT) on time axis can capture the different motion frequencies between the slower car and the faster truck. (a) is a video sequence with length six. DWT of (a) on time axis results in the sub-bands in (b). (c) is the heat maps of the right three sub-bands in (b), which can clearly show the difference between their movements.

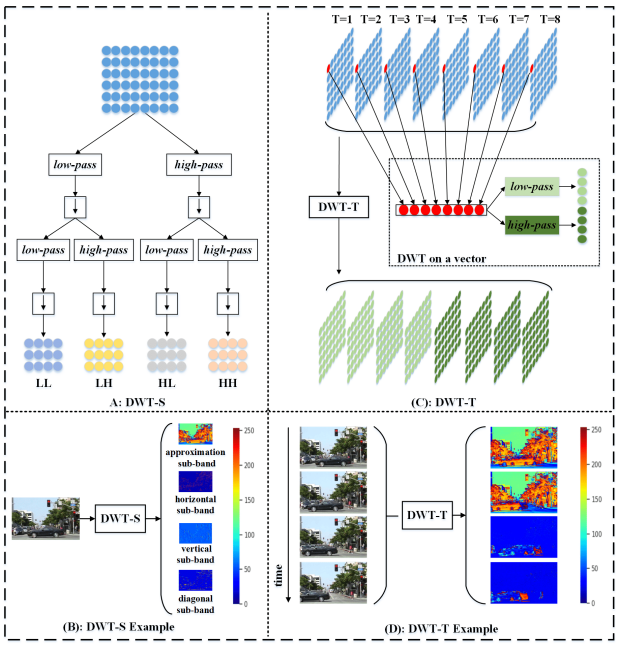

Figure 3. (A): Discrete Wavelet Transform in Spatial dimension (DWT-S) decomposes an image into one low frequency sub-band (LL) and three high frequency sub-bands of different directions (LH, HL, HH) which represent sub-bands of different directions (horizontal, vertical, diagonal). (B): An visualization example of (A). (C): Discrete Wavelet Transform in Temporal dimension (DWT-T) decomposes an image sequence into low frequency sub-bands and high frequency sub-bands on time axis. (D): An visualization example of (C). The sub-bands are visualized in heatmap style.

ficient use of motion information, resulting in distortion and temporal inconsistency:

Loss of details. Down-sampling is commonly adopted to enlarge the receptive field and extract global information, resulting in inevitable loss of high-frequency details. However, video prediction is a pixel-wise dense prediction problem. Sharp predictions would not be made without the assistance of fine details. Although dilated convolution can be employed to avoid using down-sampling, it has the problem of grid effect and is not friendly to small objects, which hinders the application to video prediction.

Insufficient exploitation of temporal motions. Dynamic scenes are composed of motions with more than one temporal frequency. In Figure 2, the lower temporal motion of the smaller car in the left and the faster temporal motion of the bigger truck in the right. They have different moving frequencies. However, previous methods usually process them one by one at a fixed frame rate. Although Recurrent Neural Networks (RNNs) are used to memorize dynamic dependencies, it has no ability to distinguish motions at different frequencies and cannot analyze time-frequency characteristics of temporal information.

Therefore, it is necessary to introduce multi-frequency analysis into video prediction task. Biological studies [16, 4] have shown that Human Visual System (HVS) exhibits multi-channel characteristics for spatial and temporal frequency information. The retinal images are decomposed into different frequency bands with approximately equal bandwidth on a logarithmic scale for processing [29], which includes a low frequency band and multiple high frequency bands. Besides spatial dimension, there also is a similar frequency band decomposition in temporal dimension. These characteristics enable the Human Visual System (HVS) to process visual content with better discrimination of detailed information and motion information. Wavelet analysis [6, 1] is a spatial-scale (temporal-frequency) analysis method, which has the characteristic of multi-resolution (frequency) analysis and can well represent the local characteristics of spatial (temporal) frequency signal, which is very similar to HVS.

Discrete Wavelet Transform (DWT) is a common wavelet analysis method for image processing. As shown in Figure 3(B), the Discrete Wavelet Transform in Spatial dimension (DWT-S) (Figure 3(A)) can decompose an image into one low frequency sub-band and three anisotropic high frequency sub-bands of different directions (horizontal, vertical, diagonal). Figure 3(D) shows the Discrete Wavelet Transform in Temporal dimension (DWT-T) (Figure 3(C)) decomposes a video sequence of length four into two high-frequency sub-bands and two low-frequency sub-bands on time axis. The frequency on time axis here can be viewed as how fast the pixels change with time, which is related to temporal motions. Inspired by the characteris-

tics of HVS and wavelet transform, we propose to explore the multi-frequency analysis for high-fidelity and temporal-consistency video prediction. The main contributions are summarized as follows:

1) To the best of our knowledge, we are the first to propose a video prediction framework based on multi-frequency analysis that is trainable in an end-to-end manner.

2) To strengthen the spatial details, we develop a multi-level Spatial Wavelet Analysis Module (S-WAM) to decompose each frame into one low-frequency approximation sub-band and three high-frequency anisotropic detail sub-bands. The high-frequency sub-bands represent the boundary details well and are in favor of sharpening the prediction details. Besides, multi-level decomposition forms a spatial frequency pyramid, helping to extract objects' features with multi scales.

3) To fully exploit the multi-frequency temporal motions of objects in dynamic scenes, we employ a multi-level Temporal Wavelet Analysis Module (T-WAM) to decompose buffered video sequence into sub-bands with different time frequencies, promoting the description of multi-frequency motions and helping to comprehensively capture dynamic representations.

4) Both quantitative and qualitative experiments on diverse datasets demonstrate a significant performance boost than the state-of-the-art. Ablation studies are made to show the generalization ability of our model and the evaluation of sub-modules.

2. Related Work

2.1. Video Generation and Video Prediction

Video generation is to synthesize photo-realistic image sequences without the need to guarantee the fidelity of the results. It focuses on modeling the uncertainty of the dynamic development of video to produce results that may be inconsistent with the ground truth but reasonable. Differently, Video prediction is to perform deterministic image generation. It needs not only to focus on the per-frame visual quality, but also to master the internal temporal features to determine the most reliable development trend that is closest to the ground truth.

Stochastic Video Generation. Stochastic Video Generation models focus on handling the inherent uncertainty in predicting the future. They seek to generate multiple possible futures by incorporating stochastic models. Probabilistic latent variable models such as Variational Auto-Encoders (VAEs) [20, 33] and Variational Recurrent Neural Networks (VRNNs) [7] are the most commonly used structures. [2] developed a stochastic variational video prediction (SV2P) method that predicted a different possible future for each sample of its latent variables, which was the first to provide effective stochastic multi-frame generation for real-world videos. SVG [8] proposed a generation model that combined deterministic prediction of the next frame with stochastic latent variables, introducing a per-step latent variables model(SVG-FP) and a variant with a learned prior (SVG-LP). SAVP [22] proposed a stochastic generation model combining VAEs and GANs. [5] extended the VRNN formulation by proposing a hierarchical variant that used multiple levels of latents per timestep.

High-fidelity Video Prediction. High-fidelity Video Prediction models aim to produce naturalistic image sequences as close to the ground truth as possible. The main consideration is to minimize the reconstruction error between the true future frame and the generated future frame. Such models can be classified as direct prediction models [35, 47, 44, 21, 3, 40, 30, 39, 18, 25] and transformation-based prediction models [50, 41, 38, 32]. Direct prediction models predict pixel values of future frames directly. They use a combination of forward neural network and recurrent neural network to encode spatial and temporal features, and then perform decoding to get the prediction with the corresponding decoding network. Generative adversarial networks (GANs) are often employed to make the predicted frames more realistic. Transformation-based prediction models aim at modeling the source of variability and operate in the space of transformations between frames. They focus on learning the transformation kernels between frames which are applied to the previous frames to synthesize the future frames indirectly.

Here, latent variables in stochastic video generation models is not considered in our model. Such models learn and sample from a space of possible futures to generate the subsequent frames. Although reasonable results can be generated by sampling different latent variables, there is no guarantee of consistency with the ground truth. Moreover, the quality of generation results vary from sample to sample, which is uncontrollable. This limits the application of such models in some practical tasks requiring a high degree of certainty, such as autonomous driving. We focus on high-fidelity video prediction, aiming to construct a prediction model to predict realistic future frame sequences as close to the ground truth as possible. To overcome the challenges of lack of details and motion blur, we propose to explore multi-frequency analysis based video prediction by incorporating wavelet transform with generative adversarial network.

2.2. Wavelet Transform

Wavelet Transform (WT) has been widely applied in image compression [6] and image reconstruction [17]. In image processing, Discrete Wavelet Transform (DWT) is often

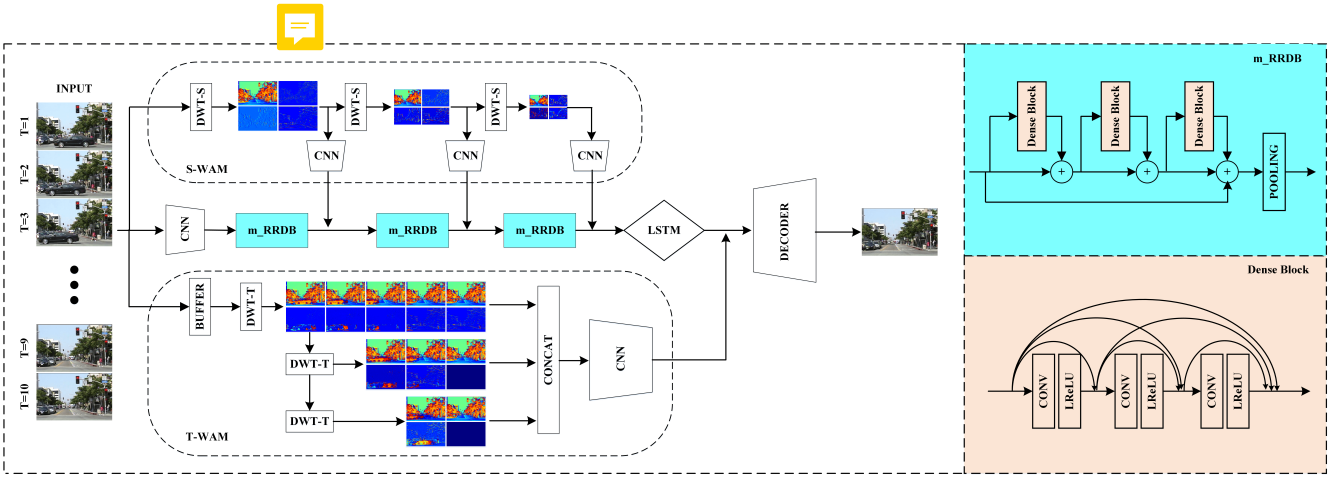

Figure 4. The pipeline architecture of our network. Note that the diagram takes the next frame prediction as an example. Multi-frame prediction can be done by feeding the predicted frame into the encoder network.

used. A fast implementation of it by using filter bank is proposed in [28]. The filter bank implementation of wavelets can be interpreted as computing the wavelet coefficients of a discrete set of child wavelets for a given mother wavelet. According to [28], we illustrate the process of DWT on space axes of an image and DWT on time axis of a video sequence in Figure 3. Multi-level DWT can be done by repeating a similar process on a sub-band images. The multi-resolution (frequency) analysis of DWT is consistent with Human Visual System (HVS), which provides a biological basis for our approach. We recommend to refer to [28] to learn more about Discrete Wavelet Transform (DWT).

3. Method

3.1. Problem Statement

We aim to synthesize future frames of high fidelity and temporal consistency by observing several beginning frames. Let $X = \{x_i\}, (1 \leq i \leq m)$ be the input of length m. $x_i \in \mathbb{R}^{H \times W \times C}$ represents the ith frame. H, W and C are the height, width and channel number. Let $Y = \{y_j\}, (1 \leq j \leq n)$ represents the ground truth of future frame sequence of length n and $\hat{Y} = \{\hat{y}_j\}, (1 \leq j \leq n)$ represents the prediction of Y. The goal is to minimize the reconstruction error between $\hat{Y}$ and Y. We will take the next frame prediction as an example.

3.2. Network Architecture

We adopt generative adversarial network as the model structure. The Generator G and discriminator D are trained with competing goals: G aims to predict frames that can fool D, while D aims to distinguish whether the input samples are real (from the training dataset) or fake (from G).

Figure 4 demonstrates the overall block diagram of the generator G to predict frame $t+1$ at time step t. It follows an encoder-decoder architecture. The encoder aims to transform the input sequence into a hidden feature tensor, while the decoder is in charge of decoding the feature tensor to generate the prediction of the next frame. The encoder consists of three part: stem CNN-LSTM, cascaded Spatial Wavelet Analysis Modules (S-WAMs) and Temporal Analysis Module (T-WAM). The decoder is composed of deconvolution and up-sampling layers.

The stem encoder is a *'CNN-LSTM'* structure. At each time step t $(t \geq 1)$, the frame x_t is passed through the stem network to extract multi-scale spatial information under different receptive fields. To pursue a better expression of appearance features, we refer to the Residual-in-Residual Dense Block (RRDB) proposed by [42] in the design of our stem structure. It is a combination of multi-level residual network and dense connections. We make a modification: adding a down-sampling layer in each RRDB unit to reduce the size of feature maps.

To reserve more high-frequency spatial details, considering multi-resolution analysis of wavelet transform, we propose a Spatial Wavelet Analysis Module (S-WAM) to enhance the representation of high-frequency information. As illustrated in Figure 4, S-WAM consists of two stages: Firstly, the input is decomposed into one low-frequency sub-band and three high-frequency detail sub-bands by DWT on Spatial dimension (DWT-S); Secondly, the sub-bands are fed into a shallow CNN to do further feature extraction and obtain consistent number of channels with the corresponding m_RRDB unit. We cascade three S-WAMs to do multi-level wavelet analysis. The output of each level of S-WAM is added with the corresponding feature tensors of the m_RRDB unit. The cascaded S-WAMs provide the compensation of details to the stem network under multiple frequencies, which promotes the prediction of fine details.

On the other side, to model the temporal multi-frequency motions in video sequences, we design a multi-level Temporal Wavelet Analysis Module (T-WAM) decomposing the

sequence into sub-bands under different frequencies on time axis. In our experiments, we conduct multi-level DWT on temporal dimension (DWT-T) on the input sequence until the number of low-frequency sub-bands or high-frequency sub-bands equals two. We take three DWT-T as an example in Figure 4. Then we concatenate those sub-bands as the input of a CNN to extract features and adjust the size of feature maps. The output is fused with the historical information from LSTM cell to strengthen the ability to distinguish multi-frequency motions for the model. The fused feature tensors from the encoder network are fed to the decoder network to generate the prediction of the next frame. We conduct a discriminator network as [30] and train the discriminator to classify the input $[X, \hat{Y}]$ into class 0 and the input $[X, Y]$ into class 1.

3.3. Loss Function

We adopt multi-module losses which consists of the image domain loss and the adversarial loss.

Image Domain Loss. We combine $\mathcal{L}_2$ loss with the Gradient Difference Loss (GDL) [30] as the image domain loss:

$$\mathcal{L}_{img}(Y, \hat{Y}) = \mathcal{L}_2(Y, \hat{Y}) + \mathcal{L}_{gdl}(Y, \hat{Y}). \quad (1)$$

$$\mathcal{L}_2(Y, \hat{Y}) = ||(Y - \hat{Y})||_2^2 = \sum_{i=1}^{n} \|(y_i - \hat{y}_i)\|_2^2. \quad (2)$$

$$\begin{aligned}\mathcal{L}_{gdl}(Y, \hat{Y}) = &\sum_{i=1}^{n} \sum_{i,j} \big| |y_{i,j} - y_{i-1,j}| - |\hat{y}_{i,j} - \hat{y}_{i-1,j}| \big|^{\alpha} \\ &+ \big| |y_{i,j-1} - y_{i,j}| - |\hat{y}_{i,j-1} - \hat{y}_{i,j}| \big|^{\alpha},\end{aligned} \quad (3)$$

where α is an integer greater or equal to 1, and $|.|$ is the operation of absolute value function.

Adversarial Loss. Adversarial training involves a generator G and a discriminator D, where D learns to distinguish whether the frame sequence is from the real dataset or produced by G. The two networks are trained alternately, thus improving until D can no longer discriminate the frame sequence generated by G. In our model, the prediction model is regarded as a generator. We formulate the adversarial loss on the discriminator D as:

$$\mathcal{L}_D^A = -logD([X, Y]) - log(1 - D(X, \hat{Y})), \quad (4)$$

and the adversarial loss for the generator G as:

$$\mathcal{L}_G^A = -logD([X, \hat{Y}]). \quad (5)$$

Hence, we combine the losses previously defined for our generator model with different weights:

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{img} + \lambda_2 \mathcal{L}_G^A, \quad (6)$$

where λ_1 and λ_2 are hyper-parameters to trade off between these distinct losses.

4. Experiments

4.1. Experiment Setup

Datasets. We perform experiments on diverse datasets widely used to evaluate video prediction models. The KTH dataset [34] contains 6 types of actions from 25 persons. We use person 1-16 for training and 17-25 for testing. Models are trained to predict next 10 frames based on the observation of previous 10 frames. The prediction range of testing is extended to 20 or 40 frames. The hyper parameters in the loss function on KTH dataset are: $\lambda_1 = 1$ and $\lambda_2 = 0.01$. The BAIR dataset [10] consists of a random moving robotic arm that pushes objects on a table. This dataset is particularly challenging due to the high stochasticity of the arm movements and the diversity of the background. We follow the setup in [22] and the hyper parameters in the loss function on the BAIR dataset are: $\lambda_1 = 1$ and $\lambda_2 = 0.001$. In addition, following the experiments settings in [24], we validate the generalization ability of our models on the car-mounted camera datasets (train: KITTI dataset [14], test:Caltech Pedestrian dataset [9]). The hyper parameters are: $\lambda_1 = 1$ and $\lambda_2 = 0.001$.

Metrics. Quantitative evaluation of the the accuracy is performed based on Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) metrics [46]. Higher values indicate better results. To measure the realism of predicted results, we employ the metric of Learned Perceptual Image Patch Similarity (LPIPS) [49]. Fréchet Video Distance (FVD) [37] is also adopted to evaluate the distribution over entire videos.

Table 1. The average comparison results over predicted 20 time steps ($10 \rightarrow 20$) and 40 time steps ($10 \rightarrow 40$) based on 10 time steps on the KTH dataset. The best results under each metric are marked in bold.

Method	KTH					
	$10 \rightarrow 20$			$10 \rightarrow 40$		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
MCNET [39]	25.95	0.804	-	23.89	0.73	-
fRNN [31]	26.12	0.771	-	23.77	0.678	-
PredRNN [45]	27.55	0.839	-	24.16	0.703	-
PredRNN++ [43]	28.47	0.865	-	25.21	0.741	-
VarNet [19]	28.48	0.843	-	25.37	0.739	-
E3D-LSTM [44]	29.31	0.879	-	27.24	0.810	-
MSNET [23]	27.08	0.876	-	-	-	-
SAVP [22]	25.38	0.746	9.37	23.97	0.701	13.26
SAVP-VAE [22]	27.77	0.852	**8.36**	26.18	0.811	**11.33**
SV2P time-invariant [2]	27.56	0.826	17.92	25.92	0.778	25.21
SV2P time-variant [2]	27.79	0.838	15.04	26.12	0.789	22.48
Ours	**29.85**	**0.893**	11.81	**27.56**	**0.851**	14.13
Ours (w/o S-WAM)	29.13	0.872	12.33	26.42	0.805	16.06
Ours (w/o T-WAM)	28.57	0.839	15.16	26.08	0.782	17.45
Ours (w/o WAM)	27.37	0.821	18.31	24.03	0.721	20.07

4.2. Quantitative Evaluation

The results of methods [39, 31, 45, 43, 19, 44, 23, 5] are reported in the reference papers [44, 19, 23, 5]. For the

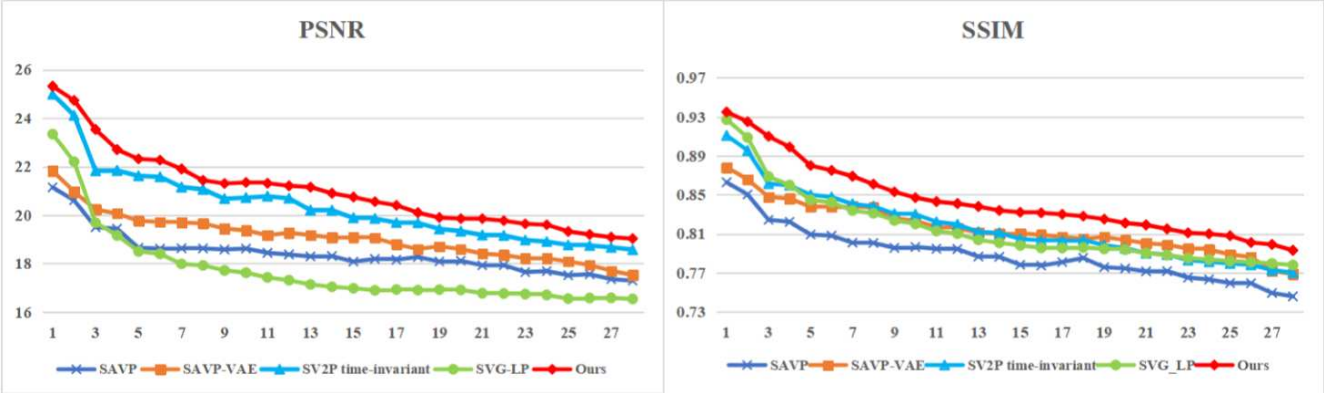

Figure 5. Quantitative comparison of different prediction models on BAIR datasets. Higher values for both PSNR and SSIM indicate better performance.

Table 2. Quantitative evaluation of different methods on the BAIR dataset. The metrics are averaged over the predicted frames. The best results under each metric are marked in bold.

Method	BAIR		
	PSNR	SSIM	LPIPS
SAVP [22]	18.42	0.789	6.34
SAVP-VAE [22]	19.09	0.815	6.22
SV2P time-invariant [2]	20.36	0.817	9.14
SVG-LP [8]	17.72	0.815	6.03
Improved VRNN [5]	-	0.822	**5.50**
Ours	**21.02**	**0.844**	9.36
Ours (w/o S-WAM)	20.22	0.825	11.23
Ours (w/o T-WAM)	19.87	0.819	11.72
Ours (w/o WAM)	18.15	0.784	13.13

models [22, 2, 8], we generate the results by running the pre-trained models the authors reported online. Table 1 reports quantitative comparison on the KTH dataset. We can see that our model achieves the best result on PSNR and SSIM in terms of prediction for both future 20 frames and 40 frames, which indicates that our results are more consistent with the ground truth. However, on LPIPS, SAVP and its variants SAVP-VAE perform better than us. We analyze that the introduction of latent variables in the stochastic generation methods focuses more on the visual quality of the generated results and less on the consistency with ground truth. Nevertheless, our model focuses more on fidelity and temporal consistency with the original sequences, which is in line with our original intention.

Figure 5 illustrates the per-frame quantitative comparison on the BAIR dataset. We also calculate the average results in Table 2. In consistent with the result on KTH dataset, we obtain the best PSNR and SSIM among the reported methods. While the Improved VRNN [5] achieves the highest on LPIPS. Because of the high stochasticity of the BAIR dataset, it is challenging to maintain fidelity and temporal consistency while making good visual effects. Besides frame-wise comparison, we adopt FVD (Fréchet video Distance) [37] to evaluate the distribution over entire sequences. As shown in Table 3, our FVD results are competitive to other methods on both datasets, which shows the consistency of the distribution of the predicted sequences.

Table 3. FVD (the smaller the better) evaluation on KTH and BAIR dataset. Baselines did not evaluate on KITTI and CalTech Pedestrian.

Dataset	SVG-FP	SV2P	SAVP	Ours
KTH	208.4 [37]	136.8 [37]	78.0 [37]	**72.3**
BAIR	315.5 [37]	262.5 [37]	**116.4 [37]**	159.6

4.3. Qualitative Evaluation

We report visualization examples on KTH dataset and BAIR datasets in Figure 6 and 7. The first row is the ground truth, where the initial frames represent the input frames. Our model makes more accurate predictions while maintaining more details of the arms in the handclapping example in first group of Figure 6. Meanwhile, we predict a walking sequence that is more consistent with the ground truth in the second group of Figure 6, while for other methods, the person in the image walks out of the scene too quickly (VarNet) or two slowly (SAVP and SV2P time-invariant). For the predictions on BAIR dataset, we are also the most consistent. Though the stochastic generation methods seem to generate more clear results, they are very different from the moving trajectories of the real sequence. This again confirms our belief that introducing more stochasticity in models will sacrifice fidelity. From the experiment results above, we can see that the multi-frequency analysis of discrete wavelet transform does help models to retain more detail information as well as temporal motion information.

4.4. Ablation Study

Evaluation of generalization ability. Consistent with the previous works to evaluate the generalization ability,

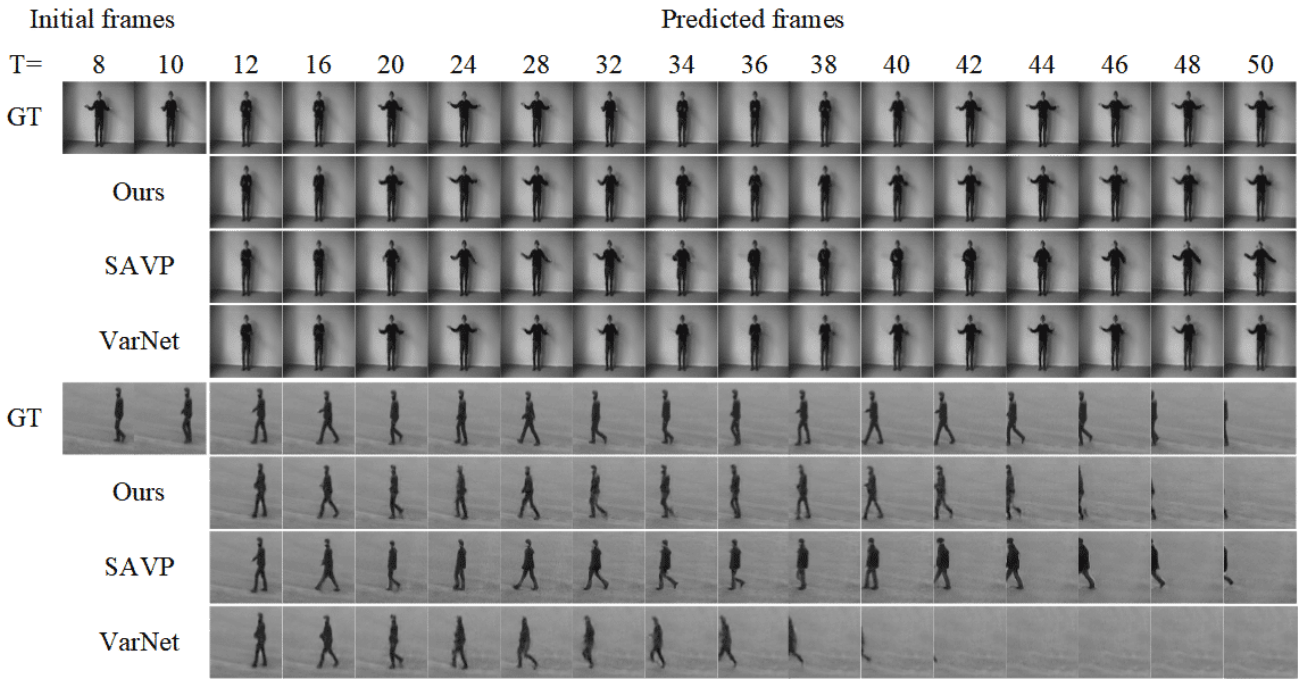

Figure 6. The prediction visualization of future 40 time steps based on the 10 frames on the KTH dataset.

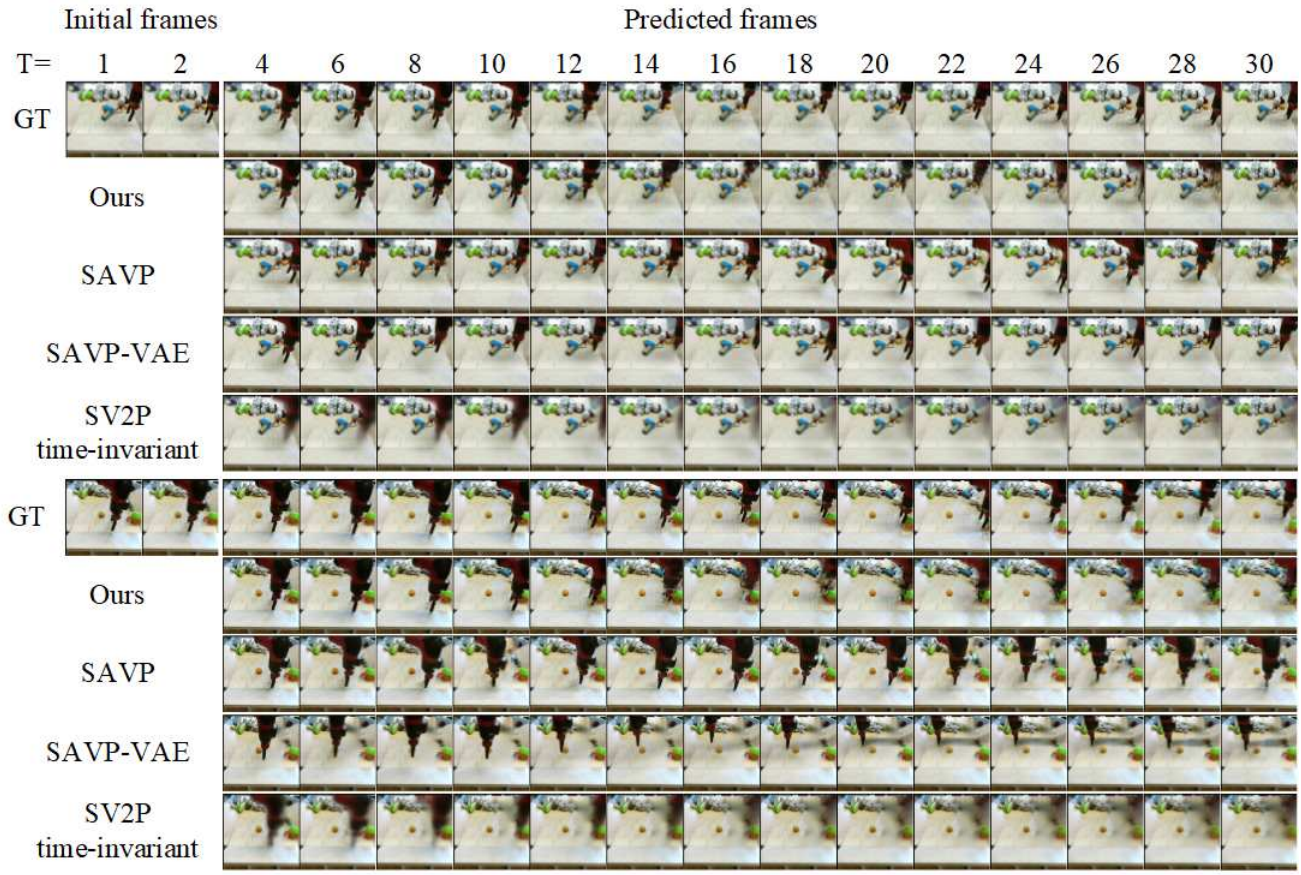

Figure 7. The prediction visualization comparison on the BAIR action free dataset. Our model predicts more consistent results to the ground truth.

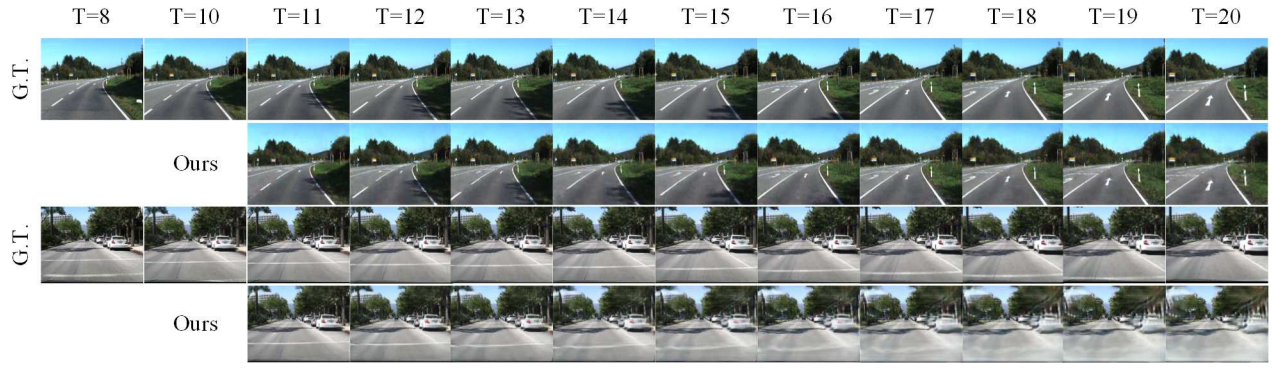

Figure 8. Visualization examples on KITTI dataset (the first group) and CalTech Pedestrian dataset (the second group).

Table 4. Evaluation of Next frame prediction on the CalTech Pedestrian dataset after trained on the KITTI dataset. All models are trained by observing 10 frames.

Method	PSNR	SSIM	LPIPS	#param
PredNet [27]	27.6	0.905	7.47	6.9M
ContextVP [3]	28.7	0.921	6.03	8.6M
DVF [26]	26.2	0.897	5.57	8.9M
Dual Motion GAN [24]	-	0.899	-	-
CtrlGen [15]	26.5	0.900	6.38	-
DPG [13]	28.2	0.923	**5.04**	-
Cycle GAN [21]	**29.2**	0.830	-	-
Ours	29.1	**0.927**	5.89	7.6M
Ours (w/o S-WAM)	28.6	0.919	6.90	7.2M
Ours (w/o T-WAM)	28.1	0.903	7.56	7.3M
Ours (w/o WAM)	26.8	0.897	7.89	6.9M

we test our model on the Caltech Pedestrian dataset after trained on KITTI dataset in Table 4. We achieve the state-of-the-art performance. Figure 8 shows the visualization examples on KITTI dataset (the first group) and Caltech Pedestrian dataset (the second group). We can see that our model predicts clearly the evolution of driving lines and the cars. The results remain consistent with the ground truth, which verifies the good generalization ability of the model. Besides, we report the number of model's parameters in Table 4. Compared to ContextVP [3] and DVF [26], our model achieves better results with fewer parameters.

Evaluation of sub-modules. To assess the impact of each sub-module, we do ablation studies in the absence of S-WAM and/or T-WAM. Results suggest that sub-modules, S-WAM and T-WAM, have both contributed to improving the prediction effect. Specifically the model without S-WAM gains more than the model without T-WAM. The visualization in Figure 9 is consistent. We analyze that the temporal motion information is of vital importance in the long-term prediction, especially for long-term prediction. Improving the expression of multi-frequency motion information in the model is the basis for making predictions with high-fidelity and temporal-consistency.

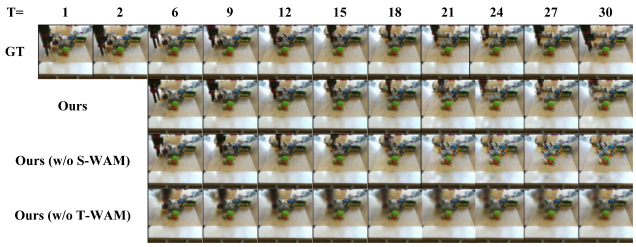

Figure 9. A BAIR Failure case. Best viewed by zooming.

Analysis of failure cases. As shown in Figure 9, for beginning motion under certain historical dependence, **Ours** model predicts accurately. Since an abrupt movement occurs (18th - 21th frame), predictions of robotic manipulator become incorrect. BAIR is indeed of high stochasticity due to the action variability. Our T-WAM module extracts the transient features of the sequence, in addition to decomposing the input into sub-band groups of different frequencies to accurately capture multi-frequency motions. However, maintaining high fidelity to accommodate abrupt motions is challenging, even for stochastic models, unless the corresponding action priors are added.

5. Conclusion

We discuss the issues of missing details and ignoring temporal multi-scale motions in current prediction models, which always lead to blurry results. Inspired by the mechanism in Human Visual System (HVS), we explore a video prediction network based on multi-frequency analysis, integrating spatial-temporal wavelet transform and generative adversarial network. The Spatial Wavelet Analysis Module (S-WAM) is proposed to reserve more details through multi-level decomposition of each frame. The Temporal Wavelet Analysis Module (T-WAM) is proposed to exploit the temporal motions through multi-level decomposition of video sequences on time axis. Extensive experiments demonstrate the superiority of our method over the latest methods.

References

[1] Milad Alemohammad, Jasper R Stroud, Bryan T Bosworth, and Mark A Foster. High-speed all-optical haar wavelet transform for real-time image compression. *Optics Express*, 2017. 2

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017. 3, 5, 6

[3] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. Contextvp: Fully context-aware video prediction. In *ECCV*, 2018. 1, 3, 8

[4] Fergus W Campbell and Janus J Kulikowski. Orientational selectivity of the human visual system. *The Journal of physiology*, 1966. 2

[5] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. *arXiv preprint arXiv:1904.12165*, 2019. 3, 5, 6

[6] Honggang Chen, Xiaohai He, Linbo Qing, Shuhua Xiong, and Truong Q Nguyen. Dpw-sdnet: Dual pixel-wavelet domain deep cnns for soft decoding of jpeg-compressed images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 2, 3

[7] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *NeurIPS*, 2015. 3

[8] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. *arXiv preprint arXiv:1802.07687*, 2018. 3, 6

[9] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 2012. 5

[10] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. *arXiv preprint arXiv:1710.05268*, 2017. 5

[11] Issam Elafi, Mohamed Jedra, and Noureddine Zahid. Unsupervised detection and tracking of moving objects for video surveillance applications. *Pattern Recognition Letters*, 2016. 1

[12] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, 2017. 1

[13] Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, and Trevor Darrell. Disentangling propagation and generation for video prediction. In *ICCV*, 2019. 8

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 5

[15] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *CVPR*, 2018. 8

[16] RF Hess and RJ Snowden. Temporal properties of human visual filters: Number, shapes and spatial covariation. *Vision research*, 1992. 2

[17] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, 2017. 3

[18] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. *CVPR*, 2018. 3

[19] Beibei Jin, Yu Hu, Yiming Zeng, Qiankun Tang, Shice Liu, and Jing Ye. Varnet: Exploring variations for unsupervised video prediction. *IROS*, 2018. 1, 5

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[21] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *CVPR*, 2019. 1, 3, 8

[22] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 1, 3, 5, 6

[23] Jungbeom Lee, Jangho Lee, Sungmin Lee, and Sungroh Yoon. Mutual suppression network for video prediction using disentangled features. In *BMVC*, 2019. 5

[24] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *ICCV*, 2017. 5, 8

[25] Wenqian Liu, Abhishek Sharma, Octavia Camps, and Mario Sznaier. Dyan: A dynamical atoms-based network for video prediction. In *ECCV*, 2018. 3

[26] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *ICCV*, 2017. 8

[27] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *ICLR*, 2017. 8

[28] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *TPAMI*, 1989. 4

[29] James Mannos and David Sakrison. The effects of a visual fidelity criterion on the encoding of images. *IEEE Transactions On Information Theory*, 2003. 2

[30] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 3, 5

[31] Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *ECCV*, 2018. 5

[32] Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *ECCV*, 2018. 3

[33] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 3

[34] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. *ICPR*, 2004. 1, 5

[35] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. *ICML*, 2015. 3

[36] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. *CVPR*, 2018. 1

[37] T. Unterthiner et al. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 5, 6

[38] Joost Van Amersfoort, Anitha Kannan, Marc'Aurelio Ranzato, Arthur Szlam, Du Tran, and Soumith Chintala. Transformation-based models of video sequences. *arXiv preprint arXiv:1701.08435*, 2017. 3

[39] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017. 1, 3, 5

[40] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. *arXiv preprint arXiv:1704.05831*, 2017. 1, 3

[41] C. Vondrick and A. Torralba. Generating the future with adversarial transformers. In *CVPR*, 2017. 3

[42] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV*, 2018. 4

[43] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. *PMLR*, 2018. 5

[44] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. In *ICLR*, 2019. 1, 3, 5

[45] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and S Yu Philip. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *NeurIPS*, 2017. 5

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5

[47] Henglai Wei, Xiaochuan Yin, and Penghong Lin. Novel video prediction for large-scale scene using optical flow. *arXiv preprint arXiv:1805.12243*, 2018. 1, 3

[48] Junqing Wei, John M Dolan, and Bakhtiar Litkouhi. A prediction- and cost function-based algorithm for robust autonomous freeway driving. *IV*, 2010. 1

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[50] Yipin Zhou and Tamara L Berg. Learning temporal transformations from time-lapse videos. In *ECCV*, 2016. 3