Journal of Engineering and Sustainable Development
Vol. 29, No. 06, November 2025
https://jeasd.uomustansiriyah.edu.iq/index.php/jeasd
ISSN 2520-0917
Review Article
https://doi.org/10.31272/jeasd.2262

Deep Learning Video Prediction Theories and Their Architecture: A Review

Zahraa Talal. Al Mokhtar*, Shefa A. Dawwd

Computer Engineering Department, College of Engineering, University of Mosul, Mosul, Iraq

*Email: zahraatalal84@gmail.com

Article Info

Received 09/10/2023
Revised 06/04/2025
Accepted 26/04/2025

Abstract

The most important factor in making a suitable decision in video-based AI systems is the ability to forecast future outputs from computer vision data. Deep-learning (DL) architectures are considered a promising new direction in many fields of computer vision. The researchers have recently introduced several novel video prediction (VP) models that achieve high performance. However, before building any prediction model, the basic principles of VP architectures and theories must be understood to determine the appropriate datasets and evaluation metrics. This study reviews 51 peer-reviewed papers published that cover the major VP architectures, including CNN, RNN, Autoencoder, VAE, and GAN models. The comparative analysis shows that 3D-CNN and GAN-based architectures achieve superior performance, with SSIM = 0.97 and PSNR = 40.2 dB across standard datasets such as UCF101 and KITTI. The novelty of this work lies in providing a comprehensive quantitative comparison of architectures, metrics, and datasets, and in proposing a unified taxonomy that integrates spatial-temporal deep learning models, their evolution from 2D to 3D, and probabilistic approaches. The paper's main contribution is offering a structured classification of VP architectures and datasets, serving as a reference framework for researchers to evaluate and design novel video prediction systems.

Keywords: Convolutional neural network; Generative adversarial network; Recurrent Model; Three-dimensional CNN layers; Video prediction

1. Introduction

The Video Prediction (VP) theories rely on deep learning (DL) architectures to forecast changes between consecutive frames [1]. Many VP applications rely on the visual appearance of videos, objects, and scenes to anticipate the future, such as autonomous driving [2], surveillance [3], and saliency prediction [4]. All prediction models operate under the fundamental theories of vision prediction [5].

However, the relationship between vision and video prediction must be understood. Vision prediction is arranged into six groups, namely VP or Frame Prediction (FP) [6], Action Prediction (AP) [7], Trajectory Prediction (TP) [8], Pose Prediction (PP) [9], Motion Prediction [10], and other applications that involve various fields like map flow [11], visual weather prediction [12] and segmentation prediction [13]. VP is considered the most generic branch of vision applications. Other branches depend on predicting activities, poses, specific actions, and trajectories in the video. These applications do not apply prediction theories to the entire frame, as in the VP algorithms. Most of these applications combine two forms of prediction, such as combining pose and action prediction [14] or motion and trajectory prediction [15] within a single model.

VP models face challenges due to the processing required to compute each pixel across the current and past frames to produce the next frame(s). Moreover, a model must extract the relationship between the spatial and temporal domains in each pixel location. This relationship makes any VP model more challenging to implement than static image models. Many researchers have suggested processing spatial and temporal features separately [16], but these models must maintain consistency between the two results. Another researcher suggested designing a VP that simultaneously extracts spatial and temporal features to maintain consistency between features and reduce the number of parameters [17].

Before building any VP model, it is important to understand the concepts related to the type of DL-VP model, the dataset used, and an appropriate metric function. This review study provides an open horizon for understanding the classical models used to design VP models, the standard VP dataset, and the appropriate

qualitative analysis. The dataset's type plays an important role in determining the number of layers, the form of the VP, and the metric used to measure model performance.

Previous review studies on video prediction have often concentrated on a single class of deep learning architectures—for instance, CNN-based or RNN-based frameworks—without providing a unified understanding of how different model families address spatial and temporal dependencies. In addition, earlier works lacked a quantitative comparison of datasets, metrics, and performance results across various architectures. Many reviews also excluded recent probabilistic and transformer-based approaches that have emerged after 2020, which are critical for improving prediction stability and real-time applicability.

This paper overcomes these limitations by integrating and comparing the major deep learning families—CNN, RNN, Autoencoder (AE), Variational Autoencoder (VAE), and Generative Adversarial Network (GAN)—within a single analytical framework. It further synthesizes quantitative evaluation results and provides an updated review of 51 studies published between 2014 and 2023, establishing a unified taxonomy of architectures, datasets, and performance metrics.

The main objectives of this study are: to present a comprehensive review of existing deep learning architectures for video prediction, to classify and analyze VP models according to their structural type (CNN, RNN, AE, VAE, GAN, and hybrid), to compare benchmark datasets and evaluation metrics (SSIM, PSNR, MSE, IOU, etc.) across different model families, and o identify performance trends, challenges, and future research directions for developing more efficient and stable VP frameworks.

The remainder of this paper is organized as follows: Section 2 explains the spatial and temporal properties of videos; Section 3 defines the theoretical basis of video prediction; Section 4 presents common VP architectures; Section 5 and Section 6 summarize datasets and evaluation metrics, respectively; Section 7 and Section 8 analyze reviewed models and comparative results; Section 9 discusses key challenges and trade-offs; and finally, Section 10 concludes the study with insights and future recommendations.

2. Spatial and Time Dimension of Videos

There are many differences between image and video processing. Video processing requires complex transformations with motion or flow patterns in the time domain. If we study a small part of spatial location across subsequent time steps, as described in Fig. 1, a wide range of similarities in the local visual are observed. However, when viewed as a whole, the sequential frames would appear visually different. As a result, the VP produces spatial-temporal (ST) interconnections to describe the dynamic relationship in consecutive frames [1].

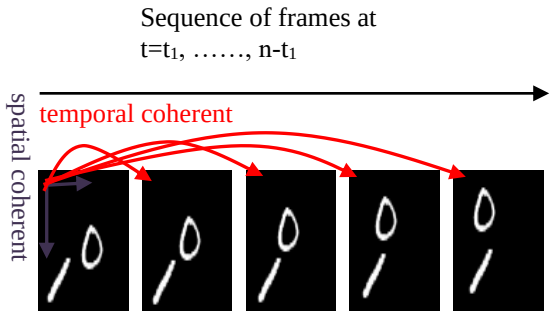

Figure 1. Spatial-temporal coherence in the sequence of frames.

3. Research Methodology

This review study used a structured, systematic approach to identify, evaluate, and synthesize the most relevant literature on deep learning-based video prediction (VP) architectures.

To ensure comprehensive coverage, the literature was collected from multiple major scientific databases, including IEEE Xplore Digital Library, Elsevier ScienceDirect, SpringerLink, and other open-access journals.

The search queries combined multiple keywords related to video prediction and deep learning, such as: “Video prediction,” “future frame forecasting,” “deep learning,” “convolutional neural network (CNN),” “recurrent neural network (RNN),” “autoencoder (AE),” “variational autoencoder (VAE),” “generative adversarial network (GAN),” and “transformer models.” Boolean operators (AND/OR) were used to combine these keywords according to the database's syntax. Each selected paper was analyzed based on standardized parameters to enable comparison across models:

- Model Type: CNN, RNN, AE, VAE, GAN, or hybrid.
- Dataset Used: KTH, UCF101, KITTI, Cityscapes, Caltech, etc.
- Performance Metrics: Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Intersection over Union (IOU), Frechet Video Distance (FVD), and others.
- Year of Publication and Application Domain: To identify temporal evolution and real-world applicability.

This methodology ensured a quantitative and reproducible review process covering 51 significant VP studies across various dataset categories and model architectures.

4. Video Prediction Definition

To define the VP theories, suppose Ft € $R^{C * W * H,}$ which represents the t^{th}-frame of the m-sequence of frames (F), where F = (Ft-m, ……., Ft-1, Ft). C, W, and H denote the number of channels and the width and height of frames, respectively. The output predicts the subsequent frames Y= [Yˆ(t+1), Yˆ(t+2), ……., Yˆ(t+n)] depending on the input frames F. Therefore, the

VP can be defined as arranging a sequence of video frames as context and temporal information to predict the next frame [18].

5. Architectures of VP

This section generally introduces the most common DL architectures used as basic building blocks in the VP models, as shown in Fig. 2.

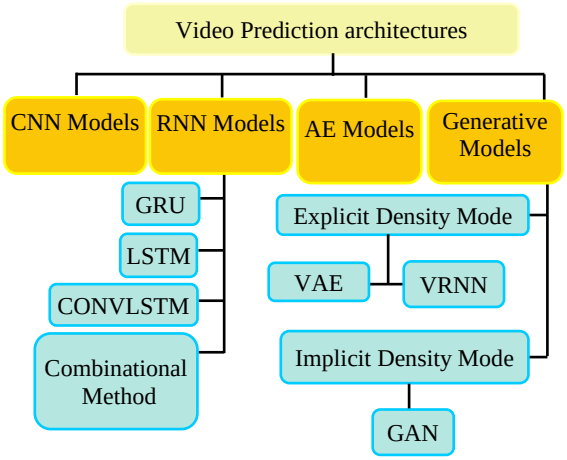

Figure. 2. Backbone video Prediction architectures.

5.1. Convolutional Neural Network Models (CNN)

The 2D CNN layers are the basic component of VP architectures used to extract the spatial framework of visual data. Unfortunately, the implementation of classical 2D CNNs on video suffers from intra- and inter-frame relationships, which are represented as spatial–temporal dependencies. Many works enhance the classical structure of CNNs by adding more layers to the basic models [2],[3], increasing the kernel size [4],[5], combining multiple levels linearly [6], and enhancing pooling operations to maintain feature resolution [6]. Other works proposed a dual 2D CNN model to improve prediction performance. The first model is applied to the sequence of frames to extract spatial features, and the second is applied to other types of data, such as optical flow [7] or segmented data [8], to extract temporal features. However, these updates of classical 2D CNN models cannot overcome the issues of ST feature extraction. As a result, many researchers have proposed new CNN models to improve prediction performance.

5.2. Recurrent Neural Networks (RNN)

RNNs are widely used in VP techniques [9], [10], [11]. Classical RNNs have several substantial constraints. These constraints appear in long-term statements due to exploding and vanishing gradients, which make backpropagation through time (BPTT) very slow. Classical RNNs were enhanced by the gated recurrent unit (GRU) [12], long short-term Memory (LSTM), and convolutional long-term short memory (CONVLSTM) models.

The GRU consists of two gates, while the LSTM and CONVLSTM consist of three gates, as shown in Fig.3 (a-c). These gates regulate the flow of gradients and information to capture or update the contents of memory cells, depending on the input and output state. Thus, these gates prevent the accumulation of irrelevant data or the fading of important data in the cells. They can control how much the gradients are affected by current and prior inputs and outputs. This helps prevent exploding or vanishing gradients. [13], [14].

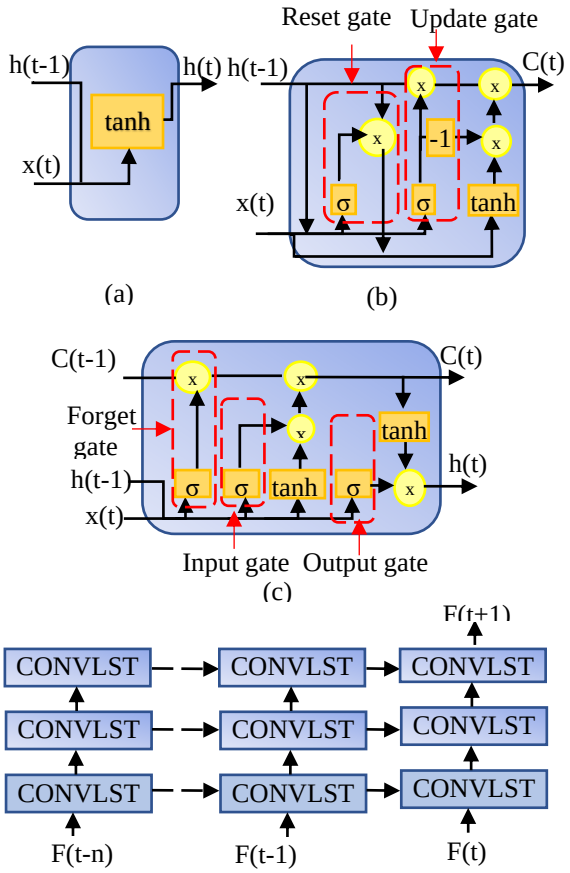

Figure 3. Internal structure of recurrent layers (a) RNN, (b) GRU, (c) CONVLSTM, and (d) CONVLSTM VP model.

The structure of CONVLSTM layers is similar to the LSTM, but internal matrix multiplications are exchanged with convolution operations. However, the 2D-CONVLSTM layers were designed solely to represent temporal relationships in sequential information. The information in the memory cell is horizontally distributed along the time dimension to capture temporal features, as shown in Fig. 3(b). Only hidden states (Ht) are transferred [15] in the vertical dimension.

Many studies applied CONVLSTM layers in VP models to enhance prediction performance. Wang et al. [2] combined the classical LSTM with 3DCNN models that presented an eidetic 3D LSTM (E3D-LSTM). Fan et al. [13] proposed a cubic long short-term memory (cubic-LSTM) unit to build the VP model. This model comprises three components: a spatial unit, a temporal unit, and an output unit.

5.3. Auto-Encoder Model

The autoencoder (AE) network is applied to the training data in a self-supervised task that consists of an encoder, a bottleneck, and a decoder, as shown in Fig. 4. The goal of using the AE

architecture is to learn low-dimensional representations from high-dimensional input data. The key to the AE structure is to reduce data dimensionality and train the model to extract the most important features.

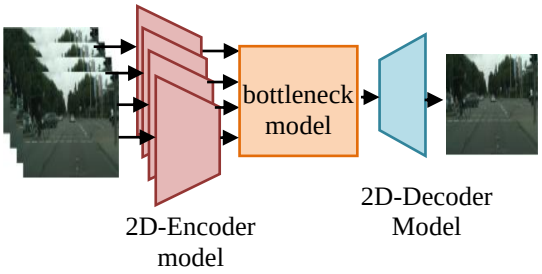

Figure 4. General architecture of the autoencoder model

Beibei Jin et al. [16] presented a transformation AE that predicts variations between adjacent frames. The current frame is fused with the prediction results to generate the future frame. Shayanfar et al. [17] presented a VP architecture based on the multiscale transformation pre-trained VGG.

5.4. Generative Models

The generative VP models are divided into two basic techniques according to how they learn: the first group is trained on the joint probability Pj(x; y) and defines the Pmodel(x) explicitly, as in variational autoencoder (VAE) models [18]. The second group depends on the conditional probability Pc (y|x) and defines Pmodel(x) implicitly, like generative adversarial network (GAN) models [19].

5.4.1. Variational Autoencoders

Variational autoencoders (VAEs) are considered an extension of AE models. The low-dimensional resolutions of X data are formulated as Z sampling, which determines the most significant variation features, as shown in Fig.5. The VAE model extracts the probability values based on its input data. It uses distributional values (mean and variance) to compare the latent variables with the normal distribution, thereby providing a better regularization effect [18].

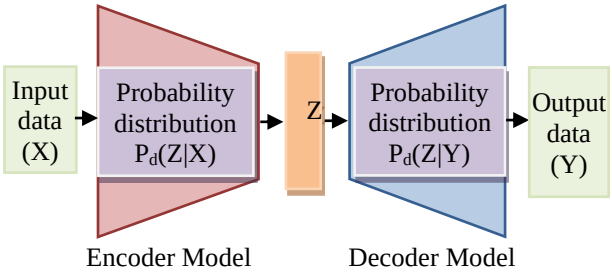

Figure 5. General design of VAE models.

Optimizing the probability function in the direct form is very difficult. Therefore, many works improved the lower bound of the probability function [9], [20], [21]. The predicted outcomes are of high quality and exhibit less blurring than those from classical AE models. The VAE is described as two models that interact to improve the final result.

5.4.2 Generative Adversarial Network

Generative adversarial network (GAN) architectures were inspired by game theory [19],[22]. They consist of two models that are trained in a combination pattern as a minimax game. The generator model creates newly generated samples that are similar to the real samples. On the other hand, the discriminator model classifies these samples to distinguish between real and generated data, as shown in Fig. 6.

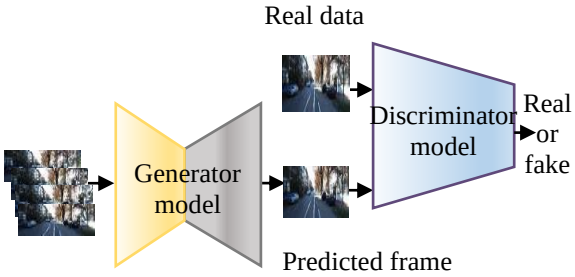

Figure 6. Basic form of GAN architecture.

The generator always creates new samples that fool the discriminator. The basic formulation of GANs is considered an unconditioned model that can generate new fabricated data using random input samples, yet many researchers still use GANs as conditional models. Mirza and Osindero [23] proposed a conditional GAN (cGAN) that leverages additional information, such as class labels, previous predictions, and multimodal samples.

Many researchers suggested using additional types of input data. Frames are applied to the basic generator while the optical flow maps [7], or motion labels [24], are applied to the second generator [19]-[22]. The discriminator model assesses the predicted results [25]. More efficient models were proposed by applying multiple discriminators to improve decision-making. Bhattacharjee and Das.[3] applied two discriminators: the first one is used to enhance the intermediate information, while the second one is implemented to make a true decision about the final results. Fan [26] presented a GAN model with a single generator and dual discriminator. This model predicts the next frame and then performs anomaly detection based on the prediction frame.

6. Datasets

Most VP models are generative, self-supervised, and require only a sequence of video frames as input. This section introduces the most widely used datasets for VP models and provides details, as shown in Table 1. These datasets can be organized by application, as shown in Fig. 7, which lists the most critical datasets for each application, ranked on a scale of [0-10]. Many datasets used in VP training are collected for a specific application. The KTH dataset is used for action recognition, but also for action and video prediction.

Table 1. The most popular video prediction datasets. Information in datasets

	Year	# videos	# frames	Resolution	Information in datasets	Type	Applications				
							FP	AP	TP	MP	O
KTH [27]	2004	2391	250000	160 × 120	RGB frames	R	*	*			
Weizmann [28]	2007	90	9000	180 × 144	RGB frames, SS (701l)	R		*			
Camvid [29]	2008	5	18202	960 × 720	RGB frames	R		*	*		
Bouncing balls[30]	2008	4000	20000	150x150	RGB frames	S	*				
Caltech[31]	2009	137	1000000	----------	RGB frames, Bb	R	*		*		
ViSOR [32]	2010	1529	1360000	----------	RGB frames	R					*
PROST [33]	2010	4 (10)	4936 (9296)	----------	RGB frames	R					*
HMDB-51 [34]	2011	6766	639300	var × 240	RGB frames	R		*	*		
Van Hateren [35]	2012	56	3584	128 × 128	RGB frames	R	*				
UCF101 [36]	2012	13320	2000000	320 × 240	RGB frames	R	*	*		*	*
NORBvideo [37]	2013	110560	552800	640 × 480	RGB frames	R	*				
Penn Action [38]	2013	2326	163841	480 × 270	RGB frames	R	*	*			*
KITTI [39]	2013	151	48791	1392 × 512	RGB frames/L/SS(200)/ Bb	R	*		*	*	
Arcade [40]	2013	-----	--------	210 × 160	RGB frames	S					*
Sports1M [41]	2014	1133 158	--------	640 × 360	RGB frames	R		*		*	*
Human3.6M [42]	2014	4000	3600000	1000x1000	RGB frames	R	*	*			
Moving MNIST [43]	2015	-------	--------	64 × 64	RGB frames	R/S	*		*		
Robotic [44]	2016	57	1500000	640 × 512	RGB frames	R	*				
Cityscapes[45]	2016	50	7000000	2048 × 1024	RGB frames/S/SS2.5K	R	*		*		*
Comma.ai [10]	2016	11	522000	160 × 320	RGB frames	R			*		*
YouTube8M [46]	2016	8200000	---------	---------	RGB frames	R/S	*			*	
YFCC100M [47]	2016	8000	-------	---------	RGB frames	R/S				*	
Inria-3DMovie [48]	2016	27	2476	960 × 540	RGBframes/IS (235)	R		*			*
THUMOS-15 [49]	2017	18404	3000000	320 × 240	RGB frames	R		*			*
BAIR Robot [50]	2017	45000	------	----------	RGB frames	R	*	*			
Apolloscape [51]	2018	4	200000	3384 ×2710	RGB frames/L/SS(147 K)	R		*			*
Robotrix [52]	2018	67	3039252	1920× 1080	RGB frames/D/SS/ IS	S		*			*
Kinetics 600 [53]	2018	500	6000000	64x64	RGB frames	X	*	*			
RoboNet [54]	2019	161000	15000000	------	RGB frames	R	*				
UASOL [55]	2019	33	165365	2280 × 1282	RGB frames/D	R					*
SynPick [56]	2021	21	503232	1920×1080	RGB frames	S	*	*		*	

(SS/IS/PS: Semantic Segmentation/Instant Segmentation Panoptic Segmentation, Bb: Bounding box. (R/SY: Real/ Synthetic, S: Stero, L: LIDAR, D: depth Application: F: Frame prediction, AP: Action Prediction, TP: Trajectory prediction, MP: Motion prediction, O: Other applications)

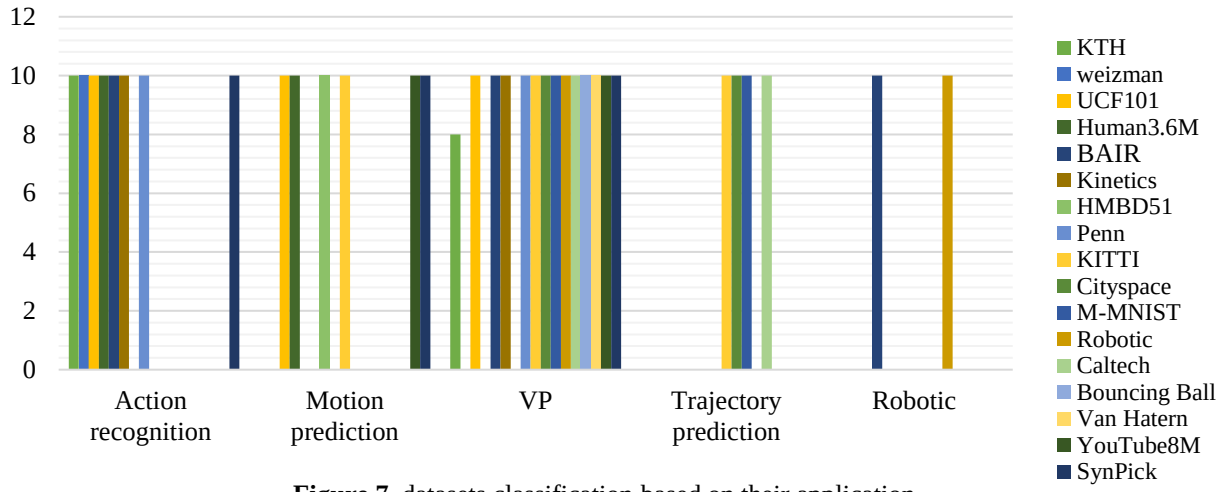

Figure 7. datasets classification based on their application.

7. Metrics and Evaluation Protocol

Most researchers evaluate VP models using criteria that assess the similarity or diversity between the results and the GT data, as described in Table 2.

8. VP Techniques

VP techniques are described in Table 3 based on 51 reviewed VP articles. In 2014, the most popular models depended on classical 2D CNN layers. Many VP models are built on multiscale AE models to extract spatial and temporal features. However, it isn't easy to extract these features simultaneously. Therefore, many researchers began developing their models by combining recurrent models with AE techniques. Ling et al. [57] introduced a multiscale predictive VP model that combined LSTM layers with an AE model.

Table 2. The formulas of the most important quantitative criteria

The Metric	Equation	Symbols
Cross Entropy (CE) [58]	$H(P,Q) = -\sum_{x}^{X} P(a) * \ln(Q(a))$	P(a): the probability of the event a in P Q(a): the probability of event a in Q
Mean Squared Error (MSE) [59]	$MSE = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}(F(x,y) - \hat{F}(x,y))^2$	F(x,y) is the GT frame $\hat{F}(x,y)$ is the predicted frame
Peak Signal to Noise Ratio (PSNR) [60]	$PSNR = 20\log(\frac{MAX}{\sqrt{MSE}})$	Max: is the maximum value of pixels
Similarity Structured Index (SSIM)[61]	$SSIM(x,y) = \left[\frac{2\mu_x\mu_y + C1}{\mu_x^2 + \mu_y^2 + C1}\right]^{\alpha}\left[\frac{2\sigma_x\sigma_y + C2}{\sigma_x^2 + \sigma_y^2 + C2}\right]^{\beta}\left[\frac{2\sigma_{xy} + C3}{\sigma_x\ \sigma_y\ + C3}\right]^{\gamma}$ $\mu_x = \frac{1}{N}\sum_{x=1}^{N} F(x)\ ,\sigma_x = \frac{1}{N-1}\sum_{x=1}^{N}(F(x) - \mu_x)^2$	L: is the dynamic range of each pixel K1, K2 and K3 <<1 $C1 = (K_1L)^2$ $C2 = (K_2L)^2$ $C3 = (K_3L)^2$
Learned Perceptual Image Patch Similarity (LPIPS) [62]	$d(p,p_0) = \sum_{l}\frac{1}{HW}\sum_{w,l}\left\|w_l * (\hat{y}^l_{hw}, \hat{y}^l_{0hw})\right\|_2^2$	The features are extracted from each layer L. The unit is normalized for layer l w_l is the scaling vector H, W: are the spatial dimensions
Frechet Video Distance (FVD)	$d(dis_R, dis_G) = \|\mu_R - \mu_G\|^2 + Tr(\sum R + \sum G - 2(\sum R \sum G)^{1/2})$	μ_R, μ_G: The mean value of $\sum$ R $\sum$ G

		$\sum \mathrm{R} \sum G$ the co-variance of dis_R, dis_G
L1 Loss function [19]	$L1(GT\ ,P\ ) = \sum_{i=1}^{N} \lvert GT\ - P\ \rvert$	GT : the ground truth data P : the predicted data
L2 Loss function [21]	$L2(GT,P) = \sum_{i=1}^{N} (GT-P)^2$	GT : the ground truth data P : the predicted data
Adversarial loss (AL) [23]	$AL = \gamma_{adv} L_{adv}^{G} + \ L_{p2p}$ $L_{adv}^{G} = L_{BCE}(\hat{P}, P)$ $L_{BCE}(\hat{P}, P) = -p\log(\hat{P}) - (1-P)\log(1-\hat{P})$	γ_{adv} a constant value 0.1 $\hat{P}$ is the predicted value of the discriminator P is the GT values
Instruction Over Union (IOU) [63]	$IOU = \frac{TP}{TP+FP+FN}$	TP: number of True positions FP: the number of false positions FN: the number of False Negatives

Patraucean [29] suggested a 2D CONVLSTM AE to extract temporal features. Straka et al. [64] suggested predictive coding CONVLSTM AE with an estimator block as a bottleneck. These techniques suffer from blurry results and increase the number of parameters. In response, researchers began using additional data, such as motion or flow maps, segmentation labels, and so on, to improve performance and reduce blurry predictions.

Table 3. The most important VP models AC: action-conditional extra input dataset. S: State of input data, Po: high-level addition Pose information, SS Semantic Segmentation data, P: Percepts, M: Motion information, IS: Instance Segmentation, De: Depth information, OF: Optical Flow.

Reference	Year	dataset	Input data	Output data	Loss function
			CNN models		
[65]	2014	[36]	RGB frame	RGB frame	CE
[29]	2015	[43], [29]	RGB frame	RGB frame	Ac
[4]	2016	[43]	RGB frame	RGB frame	--------
[66]	2017	[45]	P	SS	PSNR, SSIM, IOU
[67]	2017	[43], [36]	RGB frame	RGB frame	PSNR, SSIM
[68]	2018	[31], [46]	RGB frame	RGB frame	L2, SSIM, PSNR
[69]	2018	[39], [31]	RGB frame	RGB frame	MSE, SSIM
[8]	2019	[45]	RGB frame	SS	IOU, AC
[70]	2021	[29]	RGB frame	RGB frame	Ac
[71]	2023	[43], [29]	RGB frame	RGB frame	Accuracy
			RNN models		
[72]	2014	[30], [37]	RGB frame	RGB frame	Ac
[43]	2015	[43], [34], [34], [29]	RGB frame/ P	RGB frame	CE, L2
[73]	2018	[45]	P	SS	IOU
[74]	2018	[45]	M	M	IOU
[25]	2018	[43], [29]	RGB frame	RGB frame	MSE
[75]	2018	[43], [72], [36],	RGB frame	RGB frame	MSE, PSNR.SSIM
[2]	2019	[43], [27]	RGB frame	RGB frame	SSIM, MSE
[76]	2019	[45]	P	P, IS	L2
[59]	2020	[43], [31], [39]	RGB frame	RGB frame	MSE, SSIM
[77]	2020	[27], [39]	RGB frame	RGB frame	MSE, SSIM
[78]	2021	[42]	RGB frame	RGB frame	SSIM, MSE, MAE, FVD
[79]	2022	--------	RGB frame	RGB frame	SSIM
[80]	2022	[39],[45]	RGB frame/ M	RGB frame	PSNR, SSIM, LPIPS
[62]	2022	[43], [27], [56]	RGB frame	RGB frame	PSNR, LPIPS, SSIM
[81]	2017	[29], [38]	RGB frame Po	RGB frame	CE, IS
[20]	2018	[42], [50]	RGB frame	RGB frame	PSNR, SSIM
[9]	2019	[43], [45], [50]	RGB frame	RGB frame	FVD, LPIPS, SSIM
[82]	2020	[45], [50]	RGB frame	SS, De, Mo	IOU

[83]	2021	[39], [42], [45], [54]	RGB frame	RGB frame	PSNR, FVD, SSIM, LPIPS
			GAN models		
[84]	2017	[47]	RGB frame	RGB frame	CE, AL
[3]	2017	[29], [39], [41]	RGB frame	RGB frame	PSNR, SSIM
[85]	2019	[27]	RGB frame	RGB frame	PSNR, SSIM, LPIPS
[86]	2019	[27], [50]	RGB frame motion	RGB frame	SSIM, PSNR
[22]	2019	[29], [31], [39]	RGB frame	RGB frame	MSE, PSNR, SSIM
[87]	2020	[27], [31], [50]	RGB frame	RGB frame	PSNR, SSIM, LPIPS
[88]	2020	[31], [39]	RGB frame	RGB frame	SSIM, LPIPS
[89]	2020	[39],[45]	RGB frame/OF/SS	RGB frame	SSIM, LPIPS
[58]	2021	[53]	RGB frame	RGB frame	PSNR, SSIM
[90]	2022	[53]	RGB frame	RGB frame	PSNR, FVD, LPIPS, SSIM
[91]	2022	[29], [42]	RGB frame	RGB frame	PSNR, LPIPS
			Combined Model		
[24]	2019	[43], [27]	RGB frame	RGB frame	SSIM, PSNR
[60]	2019	[29], [31]	RGB frame	RGB frame	LPIPS, SSIM, PSNR
[92]	2021	[29], [53]	RGB frame	RGB frame	FVD
[17]	2022	[43], [39]	RGB frame	RGB frame	SSIM, MSE
[93]	2022	[43], [27], [50]	RGB frame	RGB frame	PSNR, MSE, SSIM, LPIPS
[94]	2022	[31], [50], [53], [54]	RGB frame	RGB frame	PSNR, SSIM, LPIPS
[95]	2022	------	RGB frame	RGB frame	MSE
[96]	2022	[29], [42]	RGB frame	RGB frame	PSNR, LPIPS
[97]	2023	[27], [39], [42]	RGB frame	RGB frame	PSNR, LPIPS, SSIM
[64]	2023	[31], [39]	RGB frames	RGB frames	PSNR, MES, SSIM
[98]	2023	[27][29], [39], [50], [54]	RGB frame	RGB frame	PSNR, FVD, LPIPS, SSIM

However, the additional information cannot be found in any standard dataset. Thus, the researchers updated many standard datasets to serve their models. Sun et al. [98] proposed a Motion, Scene, and Object (MOSO) framework for VP based on three AE models. The MOSO encoders are introduced based on inputs that are extracted from a sequence of frames.

In 2017, generative VP models like GANs and VAEs were introduced, based on probability density functions used to extract the most appropriate latent variable from the encoder. This procedure improves prediction performance while increasing model complexity.

In 2018, VP models were built by combining the multi-dimensional theories, which allowed the application of 3D LSTM [34] and 3D-CONV layers [46], as shown in Fig.8. After that, the researchers modified classical models like U-net, VGG, and GAN techniques by combining them with 3D CNN, RNN layers to enhance the prediction performance. Fig. 8 shows the most important methods used over the last 10 years.

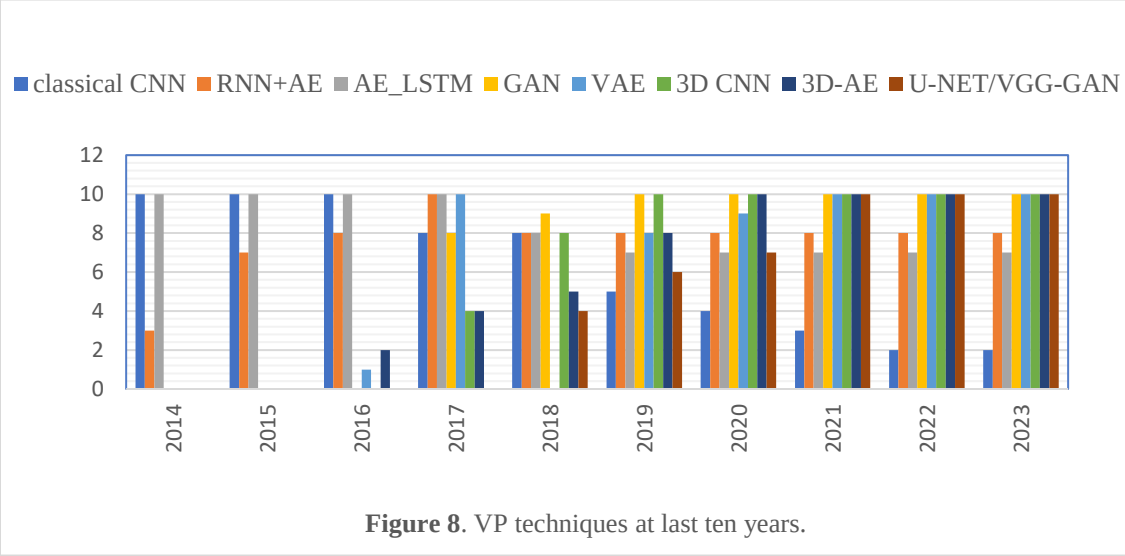

Figure 8. VP techniques at last ten years.

9. Results

A comparison of VP techniques must unify quantitative analysis and datasets across all research to enable scientific analysis of results. Many authors assessed their models using the Moving MNIST synthetic dataset, as shown in Table 4. Angel V. et al. [62] developed a Multiscale Hierarchical Prediction (MS-H Pred) model based on the M-MNIST dataset to extract spatial-temporal features at multiple scales. This model achieved first place with SSIM = 0.97 and second place with PSNR = 25.99. It is also applied to the KTH dataset. Marc Oliu [75] created an AE model with a GRU state using the M-MNIST, KTH, and UCF datasets. This model achieved the highest PSNR on the KTH dataset, as shown in Fig. 9.

Most models were trained on the KITTI, Caltech, Cityscapes, KTH, and UCF101 datasets. Prateep B. [48] proposed a multi-stage GAN to predict long-short information. It achieved the highest PSNR on the KITTI dataset. It ranked second in PSNR on the UCF101 dataset, as shown in Fig. 9. The main reason for using different datasets with the same model during training is to make VP models more accurate and better suited for practical, real-time applications. All models are implemented as pixel synthesis models that use a sequence of input frames. PSNR, MSE, and SSIM are useful metrics for measuring the difference between the GT and predicted frames.

Table 4. The summary of VP criteria and their datasets.

Papers	Model	Year	MSE *10^{-3}	SSIM	PSNR
		M-MNIST			
[67] $^{++1}$	2D-AE	2017	-----	0.93	23.
[75] $^{+++2}$	CNN+GRU	2018	9.47	0.81	21.39
[2] $^{++3}$	3D LSTM	2019	41.3	0.91	-----
[14] $^{++4}$	AE-LSTM	2019	50.5	0.87	-----
[24]$^{++5}$	MultiscaleMultiscale AE-LSTM	2019	--	0.87	27.08
[59] $^{+6}$	two-way AE	2020	22.3	0.94	-----
[17] $^{++7}$	3D-AE	2022	9	0.91	-----
[93]$^{++8}$	2D-AE with VPTR	2022	107	0.84	-----
[61]$^{+++9}$	3D-CNN	2022	23.8	0.95	-----
[62] $^{++10}$	Hierarchical 2D AE-LSTM	2022	-----	0.97	25.99
		UCF 101			
[67] $^{++1}$	2D-AE	2017	-----	0.91	31.79
[3] $^{++11}$	GAN	2017	-----	0.95	38.2
[7]$^{++12}$	GAN	2017	---	0.94	30.5
[11]$^{+++12}$	Parallel MD-LSTM	2018	-----	0.92	34.9
[69] $^{++13}$	2D-AE	2018	-----	0.96	34.26
[75]$^{+++2}$	CNN+GRU	2018	9.08	-----	23.87
[22] $^{++14}$	Cycle GAN	2019	1.37	0.94	35
[60]$^{++15}$	sparse motion 2D-CNN	2019	-----	0.91	30.8
[19] $^{++16}$	GAN	2021	2.39	0.91	30.9
[57]$^{+++++17}$	AE-LSTM	2022	39.5	0.93	------
[96] $^{++18}$	STIPAE-STGRU	2022	----	-----	30.75
[99]$^{+19}$	3D-GAN	2022	0.11	0.26	26.46
		KTH			
[75]$^{+++2}$	CNN+GRU	2018	1.75	------	29.29
[2]$^{++3}$	3D LSTM	2019	-----	0.88	29.31
[24]$^{++5}$	MultiscaleMultiscale AE-LSTM	2019	-----	0.87	27.00
[14]$^{++4}$	AE-LSTM	2019	----	0.81	27.58
[77]$^{++20}$	Inception 2D-LSTM	2020	0.463	0.96	----
[100] $^{++21}$	2D CNN-AE	2021	-----	0.83	27.11
[57]$^{+++++17}$	AE-LSTM	2022	-----	0.88	31.87
[93]$^{++8}$	2D-AE with VPTR	2022	-----	0.85	26.13
[61]$^{+++9}$	3D-CNN	2022	-----	0.90	33.72

[62]++10	Hierarchical 2D AE-LSTM	2022	-----	0.93	28.93
[97]++++22	Pyramidal CNN-LSTM	2023	----	0.89	32.05
[98]++23	Multi-stage-VAE	2023	----	0.82	29.8
	KITTI				
[3]++11	GAN	2017	-----	0.94	40.2
[87] +24	AE-DWT	2018	----	0.91	27.98
[11]+++12	Parallel MD-LSTM	2018	1.94	0.92	28.7
[77]++20	Inception 2D-LSTM	2020	7.191	0.87	-----
[89]++25	CNN-AE	2020	-----	0.61	
[57]++++++17	AE-LSTM	2022	-----	0.45	15.69
[17]++7	3D-AE	2022	25	0.53	-----
[80] ++26	2D-VAE	2022	-----	0.38	14.32
[97]++++22	Pyramidal CNN-LSTM	2023	------	0.86	25.44
[98] ++23	Multi-stage-VAE	2023	------	0.59	21.1
	Human 3.6M				
[11]+++12	Parallel 3D-LSTM	2018	-----	0.99	45.2
[101]++27	Multi-stage AE-LSTM	2018	-----	0.97	39.7
[100]++21	2D CNN-AE	2021	-----	0.91	26.1
[57] +++++17	AE-LSTM	2022	-----	0.96	31.9
[96]++18	STIPAE-STGRU	2022	-----	-----	30.9
[97]++++22	Pyramidal CNN-LSTM	2023	----	0.94	27.5
[64]++28	Predictive AE-LSTM	2023	2.05	0.92	28.4
	Caltech Pedestrian Dataset				
[7]++12	GAN	2017	2.41	0.89	-----
[69] ++13	2D-AE	2018	0.87	0.95	-----
[60]++15	sparse motion 2D-CNN	2019	2.65	0.87	-----
[22]++14	Cycle GAN	2019	1.61	0.92	29.9
[19]++19	GAN	2021	1.88	0.91	28.7
[57]++++++17	AE-LSTM	2022	-----	0.68	19.87
[61] +++9	3D-CNN	2022	1.56	0.94	33.1
[97] ++++22	Pyramidal LSTM	2023	------	0.86	25.4
[64]++28	Predictive AE-LSTM	2023	2.02	0.93	28.5
	Cityscapes				
[101]++27	Multi-stage AE-LSTM	2018	-----	0.77	26.6
[89] ++25	CNN-AE	2020	-----	0.67	----
[80]++26	2D-VAE	2022	-----	0.64	21.4

The ++++n indicates the number of datasets used in each paper, and (n) indicates the order of the paper.

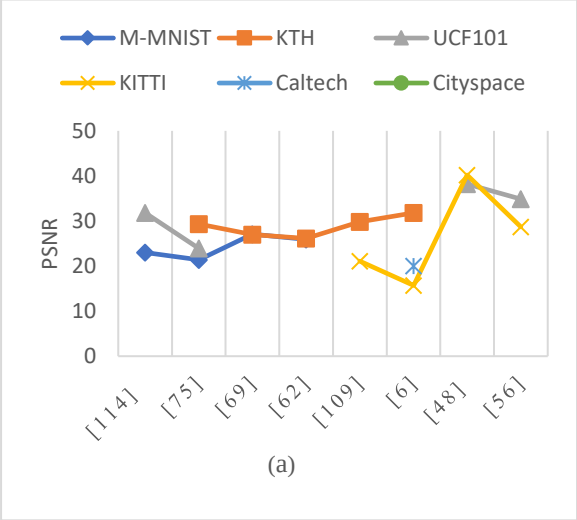

Figure 9. (a) psnr comparison (b) ssim comparison.

At the same time, many researchers began using high-level data semantics, instance segmentation, and panoptic segmentation to enhance VP models, as shown in Table 5. The Instruction over Union (IoU) is the most popular metric for measuring VP performance at the segmentation level. The datasets contain segmentation labels that must be used, as in the Cityscapes dataset.

Table 5. A comparative study based on the Cityscapes dataset. Higher IOU values indicate a better outcome.

papers	Year	#frames (Input-output)	IOU ↑
[66]	2017	3-3	55.5
[73]	2018	4-3	60.06
[74]	2018	4-3	**61.47**
[102]	2018	4-3	**67.1**
[8]	2019	4-1	**65.08**
[82]	2020	5-3	46.4

10. Discussion

VP architectures range from straightforward, direct models to very complex probabilistic models. Many factors, such as dataset type, number of layers, and model structure, play an important role in improving prediction performance and reducing parameter count. VP design faces many challenges. The main challenges are the inherent pixel-level contrast and the curse of dimensionality, which significantly complicate the development of robust prediction models. This often leads to the application of data transformation strategies based on estimation theories. In this case, the prediction performance depends on the accuracy of the estimated block, such as vector-based pertained models.

The second challenge involves balancing the predicted performance and model complexity. This trade-off affects the ability to implement the VP model in a real-time system. This challenge depends on the VP model's structure. Therefore, it is important to choose a suitable CNN structure to build reliable VP models.

Although significant progress has been achieved in video prediction (VP) through deep learning, several promising research paths remain open.

- Integration with Diffusion and Transformer Models: Recent advances in diffusion-based generative models and vision transformer architectures (ViTs) have demonstrated remarkable capacity for capturing global spatial–temporal dependencies. Future VP frameworks could integrate these models with existing CNN or RNN blocks to enhance prediction fidelity and long-term temporal consistency while reducing blurriness in generated frames.
- Benchmark and Dataset Unification: The current VP literature relies on diverse datasets (e.g., KTH, UCF101, KITTI, Cityscapes) with inconsistent resolutions and label formats. Establishing unified benchmarking protocols, shared preprocessing pipelines, and standardized evaluation metrics (e.g., SSIM, PSNR, FVD) would allow more reliable cross-model comparison and accelerate reproducibility in the field.
- Real-Time and Resource-Efficient Deployment: Many high-performing architectures, particularly 3D-CNN and GAN-based models, remain computationally expensive for embedded or real-time applications such as autonomous driving or surveillance. Future research should focus on lightweight architectures, knowledge distillation, and hardware-aware optimization to achieve real-time inference with minimal energy consumption.
- Hybrid Learning and Self-Supervision: Combining supervised, unsupervised, and self-supervised learning can reduce dependency on annotated video data. Integrating multimodal cues such as optical flow, depth, and motion segmentation will further enhance predictive robustness in complex dynamic environments.

Overall, future efforts should emphasize efficient, explainable, and transferable VP architectures that balance prediction accuracy, computational cost, and generalization across diverse visual domains.

11. Conclusion

VP models are the most important part of vision prediction architectures. These models process the sequence of frames to forecast future frame(s), using various architectures and additional data. The VP structures depend on employing the third dimension (temporal component) of videos, which significantly complicates the prediction processing.

This review provides a comprehensive overview of the basic steps for designing VP models based on dataset type, layer type, and overall model structure. The article helps readers understand the fundamental principles of state-of-the-art VP deep learning models and their datasets. These models are executed based on many types of datasets. The most widely used datasets are KTH, UCF101, KITTI, Caltech, and Cityscapes, which are also used to train VP models.

The prediction performance demonstrates that the 2D GAN models outperformed others in terms of PSNR and SSIM using the UCF101 and KITTI datasets. Nevertheless, these models still suffer from instability in system prediction. The 3D CNN models are applied to solve this problem by extracting spatial and temporal features in real time and thereby increasing system stability.

The 3D CNN models achieve satisfactory predictive performance, but they increase model complexity. Future work is required to reduce this complexity.

Conflict of interest

The authors affirm that they have no conflicts of interest regarding the publication of this manuscript.

Author Contribution Statement

Zahraa Talal Al-Mokhtar: Conceived and designed the study, performed the comprehensive literature review, organized the classification of architectures and datasets, and prepared the initial draft of the manuscript.

Shefa A. Dawwd: Contributed to data analysis, interpretation of results, comparative evaluation of models, and critical revision of the manuscript for important intellectual content.

Both authors read and approved the final version of the manuscript and agreed to its submission to the Journal.

References

[1]. S. Oprea *et al.*, “A Review on Deep Learning Techniques for Video Prediction,” *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 16, pp. 2806–2832, Apr. 2020, doi: https://doi.org/10.1109/TPAMI.2020.3045007

[2]. Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, “Eidetic 3d Lstm: A Model For Video Prediction And Beyond,” in *International conference on learning representations (ICLR)*, 2019.

[3]. P. Bhattacharjee and S. Das, "Temporal Coherency-based Criteria for

Predicting Video Frames using Deep Multi-stage Generative Adversarial Networks," in *31st Conference on Neural Information Processing Systems (NIPS2017)*, 2017, pp. 30–40. doi: https://doi.org/10.1201/9781003203964-4.

[4]. B. D. Brabandere, X. Jia, T. Tuytelaars, and L. V. Gool, "Dynamic Filter Networks," *arXiv (Cornell University)*, Jan. 2016, doi: https://doi.org/10.48550/arxiv.1605.09673

[5]. Y. Camgözlü and Y.Kutlu, "Analysis of Filter Size Effect in Deep Learning," *arXiv (Cornell University)*, Jan. 2021, doi: https://doi.org/10.48550/arXiv.2101.01115.

[6]. G. Interdisciplinary, "Enhancing the accuracies by performing pooling decisions adjacent to the output layer," *Sci. Reports.*, vol. 13, no. 1, pp. 13385–13414, 2023, doi: https://doi.org/10.1038/s41598-023-40566-y.

[7]. L. Lin, Lisa M.J. Lee, W. Dai, and E. P. Xing, "Dual Motion GAN for Future-Flow Embedded Video Prediction," arXiv (Cornell University), Aug. 2017, doi: https://doi.org/10.1109/iccv.2017.194

[8]. H. Chiu, E. Adeli, and J. C. Niebles, "Segmenting the Future," *IEEE Robot Autom Lett*, vol. 5, no. 3, pp. 4202–4211, Apr. 2019, doi: https://doi.org/10.1109/LRA.2020.2992184.

[9]. L. Castrejon, N. Ballas, and A. Courville, "Improved Conditional VRNNs for Video Prediction," in *IEEE/CVF international conference on computer vision*, 2019, pp. 7608–7617. doi: https://doi.org/10.1109/iccv.2019.00770.

[10]. W. Lotter, G. Kreiman, and D. Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning," in *5th International Conference on Learning Representations(ICLR)*, May 2017, pp. 08104–08122. doi: https://doi.org/10.48550/arXiv.1605.08104.

[11]. W. Byeon, Q. Wang, R. K. Srivastava, and P. Koumoutsakos, "ContextVP: Fully Context-Aware Video Prediction," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 753–769. doi: https://doi.org/10.1007/978-3-030-01270-0_46.

[12]. D. Linsley, J. Kim, V. Veerabadran, C. Windolf, and T. Serre, "Learning long-range spatial dependencies with horizontal gated recurrent units," in *Advances in neural information processing systems 31*, 2018, pp. 8315–8336. doi: https://doi.org/10.32470/ccn.2018.1116-0.

[13]. H. Fan, L. Zhu, and Y. Yang, "Cubic LSTMs for Video Prediction," in *AAAI conference on artificial intelligence*, 2019, pp. 8263–8270. doi: https://doi.org/10.1609/aaai.v33i01.33018263.

[14]. J. Zhang, Y. Wang, M. Long, J. Wang, and P. S. Yu, "Z-Order Recurrent Neural Networks For Video Prediction," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 230–235. doi: https://doi.org/10.1109/icme.2019.00048.

[15]. C. Luo, X. Li, and Y. Ye, "PFST-LSTM: A SpatioTemporal LSTM Model with Pseudoflow Prediction for Precipitation Nowcasting," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 843–857, 2021, doi: https://doi.org/10.1109/JSTARS.2020.3040648.

[16]. B. Jin, Y. Hu, Y. Zeng, Q. Tang, S. Liu, and J. Ye, "VarNet: Exploring Variations for Unsupervised Video Prediction," *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2018, doi: https://doi.org/10.1109/iros.2018.8594264

[17]. N. Shayanfar, V. Derhami, and M. Rezaeian, "Video Prediction Using Multiscale Deep Neural Networks," *Technol. J. Artif. Intell. Data Min.*, vol. 10, no. 3, pp. 423–431, 2022, doi: https://doi.org/10.22044/jadm.2022.11415.2305 .

[18]. Y. Yang, K. Zheng, C. Wu, and Y. Yang, "Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network," *Sensors (Switzerland)*, vol. 19, no. 11, 2019, doi: https://doi.org/10.3390/s19112528.

[19]. W. Lu, J. Cui, Y. Chang, and L. Zhang, "A Video Prediction Method Based on Optical Flow Estimation and Pixel Generation," *IEEE Access*, vol. 9, pp. 100395–100406, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3096788.

[20]. M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic Variational Video Prediction," in *conference paper at ICLR*, Oct. 2018, pp. 11252–11267. doi: https://doi.org/10.48550/arXiv.1710.11252

[21]. Y. Ye, M. Singh, A. Gupta, and S. Tulsiani, "Compositional Video Prediction," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10353–10362. doi: https://doi.org/10.1109/iccv.2019.01045.

[22]. Y.-H. Kwon and M.-G. Park, "Predicting Future Frames using Retrospective Cycle GAN," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1811–1821. doi: https://doi.org/10.1109/cvpr.2019.00191.

[23]. M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv Prepr. arXiv1411.1784*, Nov. 2014, doi: https://doi.org/10.48550/arXiv.1411.1784.

[24]. J. Lee, J. Lee, S. Lee, and S. Yoon, "Mutual Suppression Network for Video Prediction using Disentangled Features," in *Mutual Suppression Network for Video Prediction using Disentangled Features*, Apr. 2018, pp. 3174–3180. doi: https://doi.org/10.48550/arXiv.1804.04810.

[25]. T. Yu, L. Wang, H. Gu, S. Xiang, and C. Pan, "Deep generative video prediction," *Pattern Recognit. Lett.*, vol. 110, pp. 58–65, Jul. 2018, doi: https://doi.org/10.1016/j.patrec.2018.03.027

[26]. S. Fan, "Video Prediction and Anomaly Detection Algorithm Based On Dual Discriminator," pp. 123–127, 2020, doi: https://doi.org/10.1109/ICCIA49625.2020.00031.

[27]. C. Schüldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," in *17th International Conference on Pattern Recognition. ICPR 2004*, 2004, pp. 32–36. doi: https://doi.org/10.1109/icpr.2004.1334462.

[28]. J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comput. Vis. Image Underst.*, vol. 117, no. 6, pp. 633–659, 2013, doi: https://doi.org/10.1016/j.cviu.2013.01.013.

[29]. V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," in *4th International Conference on Learning Representations, ICLR 2016*, Nov. 2015, pp. 1–13. doi: https://doi.org/10.5220/0007409400002108.

[30]. I. Sutskever, G. Hinton, and G. Taylor, "The Recurrent Temporal Restricted Boltzmann Machine," in *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, 2008.

[31]. P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: A Benchmark," in *IEEE conference on computer vision and pattern recognition*, Jun. 2009, pp. 304–311. doi: https://doi.org/10.1109/cvpr.2009.5206631.

[32]. R. Vezzani and R. Cucchiara, "ViSOR: Video Surveillance Online Repository," *an Integr. Fram. Multimed. Tools Appl.*, vol. 50, no. 2, pp. 359–439, 2010, doi: https://doi.org/10.1145/2483977.2483987.

[33]. J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, "PROST: Parallel Robust Online Simple Tracking," in *IEEE computer society conference on computer vision and pattern recognition*, 2010, pp. 723–730. doi: https://doi.org/10.1109/cvpr.2010.5540145.

[34]. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2556–2563. doi: https://doi.org/10.1109/ICCV.2011.6126543.

[35]. V. Jain *et al.*, "Supervised learning of image restoration with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007, pp. 1–8. doi: https://doi.org/10.1109/ICCV.2007.4408909.

[36]. K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," Dec. 2012, doi: https://doi.org/10.48550/arXiv.1212.0402.

[37]. R. Memisevic and G. Exarchakis, "Learning invariant features by harnessing the aperture problem," in *International Conference on*

Machine Learning, 2010, pp. 100–108.

[38]. W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., 2013, pp. 2248–2255. doi: https://doi.org/10.1109/ICCV.2013.280.

[39]. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013, doi: https://doi.org/10.1177/0278364913491297.

[40]. M. G. Bellemare, J. Vteness, and M. Bowling, "The Arcade Learning Environment: An Evaluation Platform for General Agents," *J. Artif. Intell. Res.*, vol. 47, no. 2013, pp. 253–279, 2013, doi: https://doi.org/10.1613/jair.3912.

[41]. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," in *IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732. doi: https://doi.org/10.1109/cvpr.2014.223.

[42]. C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Trans. PATTERN Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1364, 2014, doi: https://doi.org/10.1109/tpami.2013.248.

[43]. N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised Learning of Video Representations using LSTMs," in *International Conference on machine learning*, 2015, pp. 843–852.

[44]. C. Finn, I. G. Openai, S. Levine, and G. Brain, "Unsupervised Learning for Physical Interaction through Video Prediction," in *Advances in neural information processing systems 29*, 2016, pp. 67–72.

[45]. M. Cordts *et al.*, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223. doi: https://doi.org/10.1109/cvpr.2016.350.

[46]. S. Abu-El-Haija *et al.*, "YouTube-8M: A Large-Scale Video Classification Benchmark," in *arXiv preprint arXiv:1609.08675*, Sep. 2016, pp. 8675–8685. doi: https://doi.org/10.48550/arXiv.1609.08675.

[47]. B. Thomee *et al.*, "YFCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2. Association for Computing Machinery, pp. 64–73, Feb. 01, 2016. doi: https://doi.org/10.1145/2812802

[48]. G. Seguin, P. Bojanowski, R. Lajugie, and I. Laptev, "Instance-level video segmentation from object tracks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3678–3687. doi: https://doi.org/10.1109/cvpr.2016.400.

[49]. H. Idrees *et al.*, "The THUMOS Challenge on Action Recognition for Videos 'in the Wild,'" vol. 155, p. 3, Apr. 2016, doi: https://doi.org/10.1016/j.cviu.2016.10.018.

[50]. F. Ebert, C. Finn, A. X. Lee, and S. Levine, "Self-Supervised Visual Planning with Temporal Skip Connections," in *Conference on Robot Learning*, 2017, pp. 12–16.

[51]. X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape Open Dataset for Autonomous Driving and its Application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2721, Mar. 2018, doi: https://doi.org/10.1109/TPAMI.2019.2926463.

[52]. A. Garcia-Garcia *et al.*, "The RobotriX: An eXtremely Photorealistic and Very-Large-Scale Indoor Dataset of Sequences with Robot Trajectories and Interactions," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Jan. 2019, pp. 6790–6797. doi: https://doi.org/10.1109/iros.2018.8594495.

[53]. J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A Short Note about Kinetics-600," in *arXiv preprint arXiv:1808.01340*, Aug. 2018, p. 01340. doi: https://doi.org/10.48550/arXiv.1808.01340.

[54]. S. Dasari *et al.*, "RoboNet: Large-Scale Multi-Robot Learning," in *RoboNet: Large-Scale Multi-Robot Learning*, Oct. 2019, pp. 885–897. [Online]. Available: http://arxiv.org/abs/1910.11215

[55]. Z. Bauer, F. Gomez-Donoso, E. Cruz, S. Orts-Escolano, and M. Cazorla, "UASOL, a large-scale high-resolution outdoor stereo dataset," *Sci. Data*, vol. 6, no. 1, Dec. 2019, doi: https://doi.org/10.1038/s41597-019-0168-5.

[56]. A. S. Periyasamy, M. Schwarz, and S. Behnke, "SynPick: A Dataset for Dynamic Bin Picking Scene Understanding," in *IEEE 17th International Conference on Automation Science and Engineering (CASE), IEEE*, 2021, pp. 488–493. [Online]. Available: http://www.hdrlabs.com/sibl/archive.html

[57]. C. Ling, J. Zhong, and W. Li, "Predictive Coding Based Multiscale Network with Encoder-Decoder LSTM for Video Prediction," in *arxiv*, Dec. 2022, pp. 11642–11654. doi: https://doi.org/10.48550/arXiv.2212.11642.

[58]. B. Liu, Y. Chen, S. Liu, and H.-S. Kim, "Deep Learning in Latent Space for Video Prediction and Compression," in *IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 701–710. doi: https://doi.org/10.1109/cvpr46437.2021.00076.

[59]. W. Yu, Y. Lu, S. Easterbrook, and S. Fidler, "Efficient and Information-Preserving Future Frame Prediction and Beyond," in *conference paper at ICLR*, 2020.

[60]. Y.-H. Ho, C.-Y. Cho, W.-H. Peng, and G.-L. Jin, "SME-Net: Sparse Motion Estimation for Parametric Video Prediction through Reinforcement Learning," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10460–10470. doi: https://doi.org/10.1109/iccv.2019.01056.

[61]. Z. Gao, C. Tan, L. Wu, and S. Z. Li, "SimVP: Simpler yet Better Video Prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3170–3180. doi: https://doi.org/10.1109/cvpr52688.2022.00317.

[62]. A. Villar-Corrales, A. Karapetyan, A. Boltres, and S. Behnke, "MSPred: Video Prediction at Multiple Spatio-Temporal Scales with Hierarchical Recurrent Networks," *arXiv:2203.09303*, Mar. 2022, doi: https://doi.org/10.48550/arXiv.2203.09303.

[63]. W. Lee *et al.*, "Revisiting Hierarchical Approach for Persistent Long-Term Video Prediction," in *International Conference on Learning Representations ICLR*, Apr. 2021, pp. 06697–06704. doi: https://doi.org/10.48550/arXiv.2104.06697.

[64]. Z. Straka, T. Svoboda, and M. Hoffmann, "PreCNet: Next-Frame Video Prediction Based on Predictive Coding," *IEEE Trans. Neural Networks Learn. Syst.*, vol.. 14, no. 22, pp. 1467–1486, 2023, doi: https://doi.org/10.1109/TNNLS.2023.3240857.

[65]. M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: a baseline for generative models of natural videos," *arXiv Prepr. arXiv1412.6604*, Dec. 2014, doi: https://doi.org/10.48550/arXiv.1412.6604.

[66]. P. Luc, N. Neverova, C. Couprie, J. Verbeek, Y. Lecun, and F. A. Research, "Predicting Deeper into the Future of Semantic Segmentation," in *IEEE International Conference on Computer Vision*, 2017, pp. 648–657. doi: https://doi.org/10.1109/iccv.2017.77.

[67]. X. Chen, W. Wang, J. Wang, and W. Li, "Learning object-centric transformation for video prediction," in *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, Association for Computing Machinery, Inc., Oct. 2017, pp. 1503–1512. doi: https://doi.org/10.1145/3123266.3123349.

[68]. F. A. Reda *et al.*, "SDC-Net: Video prediction using spatially-displaced convolution," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 718–733. doi: https://doi.org/10.1007/978-3-030-01234-2_44.

[69]. W. Liu, "DYAN: A Dynamical Atoms-Based Network For Video Prediction," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 170–185. doi: https://doi.org/10.17760/d20385573.

[70]. V. Kumar, V. Tripathi, and B. Pant, "Unsupervised Learning of Visual Representations via Rotation and Future Frame Prediction for Video

Retrieval," in *Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 701–710. doi: https://doi.org/10.1007/978-3-030-81462-5_61.

[71]. J. van Amersfoort, A. Kannan, M. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-Based Models of Video Sequences," in *CORR*, Jan. 2017, pp. 8435–8446. doi: https://doi.org/10.48550/arXiv.1701.08435.

[72]. V. Michalski, R. Memisevic, and K. Konda, "Modeling Deep Temporal Dependencies with Recurrent 'Grammar Cells,'" in *Advances in neural information processing systems*, 2014, pp. 27–36.

[73]. S. Shahabeddin Nabavi, M. Rochan, Yang, and Wang, "Future Semantic Segmentation with Convolutional LSTM," *BMVC*, vol. 1, no. 2, pp. 3–15, Jul. 2018, doi: https://doi.org/10.48550/arXiv.1807.07946.

[74]. S. Vora, R. Mahjourian, S. Pirk, and A. Angelova, "Future Segmentation Using 3D Structure," *arXiv:1811.11358v1*, pp. 11358–11372, Nov. 2018, doi: https://doi.org/10.48550/arXiv.1811.11358.

[75]. M. Oliu, J. Selva, and S. Escalera, "Folded Recurrent Neural Networks for Future Video Prediction," in *the European Conference on Computer Vision*, 2018, pp. 716–731. doi: https://doi.org/10.1007/978-3-030-01264-9_44.

[76]. J. Sun *et al.*, "Predicting future instance segmentation with contextual pyramid convlTMs," in *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc., Oct. 2019, pp. 2043–2051. doi: https://doi.org/10.1145/3343031.3350949.

[77]. M. Hosseini, A. S. Maida, M. Hosseini, and G. Raju, "Inception-inspired LSTM for Next-frame Video Prediction," in *arXiv preprint arXiv:1909.05622*, Aug. 2019, pp. 05622–05629. doi: https://doi.org/10.48550/arXiv.1909.05622.

[78]. H. Wu, Z. Yao, J. Wang, and M. Long, "MotionRNN: A Flexible Model for Video Prediction with Spacetime-Varying Motions," in *IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15435–15444. doi: https://doi.org/10.1109/cvpr46437.2021.01518.

[79]. P. Desai, C. Sujatha, S. Chakraborty, S. Ansuman, S. Bhandari, and S. Kardiguddi, "Next frame prediction using ConvLSTM," *J. Phys. Conf. Ser.*, vol. 2161, no. 1, pp. 012024–012039, Jan. 2022, doi: https://doi.org/10.1088/1742-6596/2161/1/012024 .

[80]. A. K. Akan, S. Safadoust, and F. Güney, "Stochastic Video Prediction with Structure and Motion," pp. 1–26, 2022, [Online]. Available: http://arxiv.org/abs/2203.10528

[81]. J. Walker, K. Marino, A. Gupta, and M. Hebert, "The Pose Knows: Video Forecasting by Generating Pose Futures," in *IEEE International Conference on Computer Vision*, 2017, pp. 3352–3361. doi: https://doi.org/10.1109/iccv.2017.361.

[82]. A. Hu, F. Cotter, N. Mohan, C. Gurau, and A. Kendall, "Probabilistic Future Prediction for Video Scene Understanding," in *16th European Conference*, Mar. 2020, pp. 767–785. doi: https://doi.org/10.1007/978-3-030-58517-4_45.

[83]. B. Wu, S. Nair, R. Martín-Martín, L. Fei-Fei, and C. Finn, "Greedy Hierarchical Variational Autoencoders for Large-Scale Video Prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2318–2328. doi: https://doi.org/10.1109/cvpr46437.2021.00235.

[84]. C. Vondrick and A. Torralba, "Generating the Future with Adversarial Transformers," in *IEEE Conference on Computer Vision and Pattern Recognition 2017*, 2017, pp. 1020–1028. doi: https://doi.org/10.1109/cvpr.2017.319.

[85]. X. Chen, C. Xu, X. Yang, and D. Tao, "Long-term Video Prediction via Criticization and Retrospection," in *IEEE Transactions on Image Processing*, 2021, pp. 7093–7107. doi: https://doi.org/10.1109/tip.2020.2998297.

[86]. Z. Hu and J. T. L. Wang, "A Novel Adversarial Inference Framework for Video Prediction with Action Control," in *IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[87]. B. Jin *et al.*, "Exploring Spatial-Temporal Multi-Frequency Analysis for High-Fidelity and Temporal-Consistency Video Prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4545–4563. doi: https://doi.org/10.1109/cvpr42600.2020.00461.

[88]. [O. Shouno, "Photo-Realistic Video Prediction on Natural Videos of Largely Changing Frames," *arXiv:2003.08635*, pp. 8635–8641, Mar. 2020, doi: https://doi.org/10.48550/arXiv.2003.08635.

[89]. Y. Wu, H. R. Gao, J. Park, P. Qifeng, and C. Hkust, "Future Video Synthesis with Object Motion Prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5539–5548. doi: https://doi.org/10.1109/cvpr42600.2020.00558.

[90]. W. Yan, D. Hafner, S. James, and P. Abbeel, "Temporally Consistent Video Transformer for Long-Term Video Prediction," Oct. 2022, doi: https://doi.org/10.48550/arXiv.2210.02396.

[91]. Z. Chang, X. Zhang, S. Wang, S. Ma, and W. Gao, "STRPM: A Spatiotemporal Residual Predictive Model for High-Resolution Video Prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13946–13955. doi: https://doi.org/10.1109/cvpr52688.2022.01356.

[92]. P. Luc *et al.*, "Transformation-based Adversarial Video Prediction on Large-Scale Data," in *arXiv*, Mar. 2020, pp. 04035–04044. doi: https://doi.org/10.48550/arXiv.2003.04035.

[93]. X. Ye and G.-A. Bilodeau, "Video Prediction by Efficient Transformers," *Image Vis. Comput.*, vol. 130, pp. 104612–104624, Dec. 2022, doi: https://doi.org/10.1109/icpr56361.2022.9956707.

[94]. Y. Seo, K. Lee, F. Liu, S. James, and P. Abbeel, "HARP: Autoregressive Latent Video Prediction with High-Fidelity Image Generator," in *IEEE International Conference on Image Processing (ICIP)*, Sep. 2022, pp. 3943–3947. doi: https://doi.org/10.1109/ICIP46576.2022.9897982.

[95]. H. Geng, T. Wang, X. Zhuang, D. Xi, Z. Hu, and L. Geng, "GAN-rcLSTM: A Deep Learning Model for Radar Echo Extrapolation," *Atmosphere (Basel)*, vol. 13, no. 5, May 2022, doi: https://doi.org/10.3390/atmos13050684.

[96]. Z. Chang, X. Zhang, S. Wang, S. Ma, and W. Gao, "STIP: A SpatioTemporal Information-Preserving and Perception-Augmented Model for High-Resolution Video Prediction," *arXiv*, Jun. 2022, doi: https://doi.org/10.48550/arXiv.2206.04381.

[97]. C. Ling, W. Li, and J. Zhong, "Analyzing and Improving the Pyramidal Predictive Network for Future Video Frame Prediction," in *arxiv*, Jan. 2023, pp. 1–12. doi: https://doi.org/10.48550/arXiv.2301.05421.

[98]. M. Sun, W. Wang, X. Zhu, and J. Liu, "MOSO: Decomposing Motion, Scene, and Object for Video Prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Mar. 2023. doi: https://doi.org/10.1109/cvpr52729.2023.01796.

[99]. M. Backus, Y. Jiang, and D. Murphy, "Video Frame Prediction with Deep Learning," in *IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 701–710. doi: https://doi.org/10.1109/siu49456.2020.9302047.

[100]. X. Gao, Y. Jin, Q. Dou, C.-W. Fu, and P.-A. Heng, "Accurate Grid Keypoint Learning for Efficient Video Prediction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Jul. 2021, pp. 5908–5915. doi: https://doi.org/10.1109/iros51168.2021.9636874.

[101]. J. Xu, B. Ni, Z. Li, S. Cheng, and X. Yang, "Structure Preserving Video Prediction," in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 1460–1469. doi: https://doi.org/10.1109/cvpr.2018.00158.

[102]. A. M. Terwilliger, G. Brazil, and X. Liu, "Recurrent Flow-Guided Semantic Forecasting," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Sep. 2018, pp. 1703–1712. doi: https://doi.org/10.1109/wacv.2019.00186.