Received: 25 June 2023 | *Revised:* 27 August 2023 | *Accepted:* 27 August 2023
DOI: 10.37917/ijeee.20.1.19 *Vol. 20 | Issue 1 | June 2024*

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Open Access
IJEEE University of Basrah College of Engineering

Deep Learning Video Prediction Based on Enhanced Skip Connection

Zahraa T. Al Mokhtar*, Shefa A. Dawwd
Department of Computer Engineering, College of Engineering, University of Mosul, Mosul, Iraq

Correspondance
*Zahraa Talal Abed
Department of Computer Engineering,
College of Engineering,
University of Mosul, Mosul, Iraq
Email: zahraatalal84@gmail.com

Abstract
Video prediction theories have quickly progressed especially after a great revolution of deep learning methods. The prediction architectures based on pixel generation produced a blurry forecast, but it is preferred in many applications because this model is applied on frames only and does not need other support information like segmentation or flow mapping information making getting a suitable dataset very difficult. In this approach, we presented a novel end-to-end video forecasting framework to predict the dynamic relationship between pixels in time and space. The 3D CNN encoder is used for estimating the dynamic motion, while the decoder part is used to reconstruct the next frame based on adding 3DCNN CONVLSTM2D in skip connection. This novel representation of skip connection plays an important role in reducing the blur predicted and preserved the spatial and dynamic information. This leads to an increase in the accuracy of the whole model. The KITTI and Cityscapes are used in training and Caltech is applied in inference. The proposed framework has achieved a better quality in PSNR=33.14, MES=0.00101, SSIM=0.924, and a small number of parameters (2.3 M).

Keywords
3DCNN, 3DCNN-CONLSTM2D skip connection, pixel generation, and video prediction.

I. Introduction

Video prediction (VP) is the most interesting approach in computer vision and object trajectories. It utilizes the sequence of consecutive data as input to appreciate what will occur in the next frames. This procedure is considered valuable in the scope of object segmentation [1], [2], anomaly detection [3], [4], motion prediction [5], [6], autonomous driving applications [7], human pose estimation and recognition [8], pedestrian detection and tracking [9], [10], weather forecasting [11]and many other applications depend on predicting future frames of a video sequence. There are many concepts in video prediction; such as motion recognition and object detection. These applications are discriminator methods and need to extract principle information helping for recognition and/or detection without preserving the whole dynamic information in each frame, then the excessive or irrelevant information is discarded [12]. However, another concept of video prediction used in many applications is a generative method that imposes simulation of the whole environment. These models extract different levels of dynamic information from the frames to create the value of each pixel that belongs to the next frames. Generally, generative models are more exciting than discriminator models [13]. Recently, VP approaches have depended on enhancing the method of extracting spatial and dynamic information by applying many patterns of models. For example, many studies suggested decomposing video frames into spatial and dynamic parts at first and using CNN/RNN models to sample the dynamics component [14]. Other recent works proposed evolving high-level spatial features from input frames, predicting the dynamic features based on high-level features, and applying decoder models to reconstruct future frames [15].

This method is applied in our study to extract the dynamic and spatial features. In this approach, we suggested an end-to-end generative model that merges the 2D and 3D CNN convolution layers to build the Autoencoder model and added a 3D CNN-CONVLSTM2D in skip connection to extract the dynamic features from the intermediate output of each encoder level. Firstly, our approach extracted the dynamic features based on the high-level spatial features which are extracted depending on 2D CNN models. The 2D CNN model is applied on each frame to produce the high-level features in each frame in a separate manner. The Encoder part consists of five input frames which are applied to 2D CNN models; the intermediate outputs of each layer are concatenated and applied to 3D CNN high-level block and 3D CNN-2D CONVLSTM2D skip connection to extract the dynamic features from the low and mid-level dimensional features. This novel type of skip connection preserves the Spatiotemporal distribution consistency of our approach and captures the dynamic information on heretical features without any additional information on the input video. Finally, our approach applied the Cityscapes and KITTI datasets for training the proposed model and the CalTech dataset for testing the model. This study is organized as follows: Section II. displays an overview of the related video prediction model. Section III. produces a complete description of all blocks applied in our deep model. Section IV. describes the evaluation matrices used to measure the performance of the models. Section V. compares the experimental results with different methods based on single-frame and multi-frame prediction. Finally, the conclusion of this study is explained in Section VI. , in addition to the future works are suggested.

II. Related Works

The expectation of the future frames itself is a difficult issue. Intuitively, the techniques should be able to capture the pixel-wise modification and the estimated dynamic motion to allow pixel values based on past or current frames transformed into the next frames [16]. based on Predictive techniques [17] like 2D Auto-Encoders (AE) [18] and recurrent neural networks (RNN) [19]. But, most 2D AE models suffer from blurry prediction results due to the inconsistency between spatial and temporal features. So, many algorithms are proposed to enhance the prediction performance by adding many layers in the Encoder and Decoder parts [20], building the intermediate block between the encoder and the decoder [14], as shown in TableI.

Padmashree Desai et.al. [21] proposed a VP model based on the CONVLSTM encoder-CONVLSTM decoder. YunboWang et.al [22] proposed a VP model that combined the 3D-LSTM with 3D CONV based on the AE. W. Lotter et. al. [23] suggested a predictive neural network model (PredNet) which combined the ConvLSTM with predictive coding concept to predict next frames by creating a local prediction in each layer. Ruben Villegas et.al. [14] introduced a 2D AE model with a CONVLSTM as the bottleneck stage to predict the next frame at the pixel level. While. Zdenek Straka et.al. [24] suggested AE with Predictive Coding Net (PreCNet) which is applied as an estimator block between the encoder and decoder parts. Xi Ye [25] presented a AE model with an intermediate neural process (NP) block that maps spatiotemporal input coordinates to produce each pixel value of the output. Zhangyang Gao et.al. [26] proposed simple spatial-temporal features translator between the encoder and the decoder part to enhance the blurry prediction. In contrast, Denton, E.L. et al. [13] proposed decomposition approaches to analyze each frame's content and motion, and then fed into separate encoders. Henglai Wei et al. [27] proposed the 2D CNN spatial model and 2D CNN flow model to capture the spatial and dynamic information separately based on Auto-encoder models. while Zheng Chang et al. [6] used a motion model to predict the next frame based on extracting Spatial-Temporal features from a series of frames that are encoded in 2D CNN models. On the other side, J. CHO et al. [29] suggested an Auto-encoder model based on a direct skip connection with a special stage of global context propagation networks (GCPN) between the encoder and decoder model. R. Zhang [28] suggested Skip Attention Encoder–Decoder (SAED) to preserve the attention of human motion features in Spatial-Temporal features based on Gated Recurrent Unite(GRU). All the above approaches still suffer from blurry prediction and complexity of the design. So, a novel lightweight VP method is suggested which applies a novel 3D CNN skip connection and high level

TABLE I.
Video Prediction Techniques

Add RNN Layers	
Papers	Techniques
P.Desai [21]	2D CONVLSTM AE.
Y.Wang [22]	3D LSTM+3D CONV AE
Intermediate Block	
W. Lotter [23]	PredNet+the ConvLSTM
R. Villegas [14]	CONVLSTM+2D AE
Z. Straka [24]	Estimator Block+2D AE.
Xi Ye [25]	NP block+2D AE
Z. Gao [26]	ST translator+2D AE
Multi model AE	
Denton [13]	content and motion AE model.
H. Wei [27]	spatial+flow models
Z. Chang [6]	Motion+spatial models
Add Skip connection into 2D AE	
R. Zhang [28]	Skip Attention AE model

3D CNN as an intermediate block to decrease the inconsistency between the spatial and temporal features in a small number of parameters. The proposed model uses five 2D CONV models to capture the spatial features from five consecutive input frames. The cubic 3D CNN models are used to capture a dynamic feature without any other supported information or additional models to decrease the blurry prediction. The dimensions of x and y of each frame extract the spatial features in the SPN model (Spatial Prediction Network) while the t-axis or z-axis of the multi-input frame extracts the temporal features from the 3D CNN Network which represents a novel bottleneck between the encoder and the decoder part. Instead of using a direct skip connection between the encoder and decoder part; the 3D CNN-CONVLSTM 2D skip connection is modified to enhance the prediction performance and decrease the blurry prediction by capture the most proper motion features,as shown in Fig. 1. The main added by this paper are explained as follows: 1) Our model applied a sequence of previous and current frames as inputs of SPN and then used three models' 3D CNN Networks; The first 3D block is to compress the dynamic information from low-resolution features. The second and the third model are used to determine the mid and high-resolution features as described in Fig. 1. The CONVLSTM layers are added to increase the accuracy of the model and preserve the dynamic information in the execrated low and high-resolution features from 3D CNN models. 2) We present a novel lightweight model of a small number of parameters and record a good performance i.e., PSNR and MSE.

III. The Proposed Method

A. The Overview of the Model

Let X_i is the ith frame in the inputs of video frames X = [X_{t-i}, .., X_t]. The basic target of our model is to predict the next future m frames O = [O_{t+1},..., O_{t+m}] depending on the sequence of input X. The difficulty of the proposed model is managing the complicated evolution of each pixel by combining two essential scene elements, namely the scene's content context and motion dynamics [30]. The framework takes n consecutive input frames X_i € R H, W, C where i € 0, n. In this paper, the number of input frames is 5 consecutive frames [31]. Our eventual goal is to get accurate forecasting results of future frames by applying complex RGB frames at time steps t_0, t_1, ..., t_n. The encoder is built by using spatial prediction networks (SPN). The 3D CNN is used to extract the dynamic sense of motion in each pixel of 5 frames and features. The addition of a 3D CNN-CONVLSTM 2D skip connection can increase the accuracy of the features, Fig. 1.

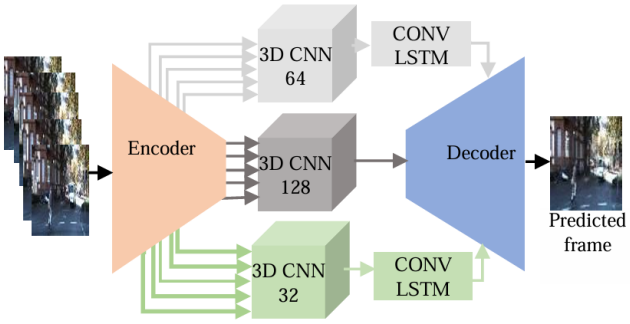

Fig. 1. The proposed model

B. The Spatial Prediction Networks (SPN)

The Encoder part of our model consists of 5 blocks of SPN, as shown in Fig. 2(a). The SPN can generate the low, mid, and high-resolution features for each frame. This model consists of three layers of CONV 2D and batch normalization (B.N.), as shown in Fig. 2(b). A good performance can be achieved by employing multi-layers of convolutions at multi-frame video prediction because it processes the spatial invariance of the frame. In this model, the spatial features are separately taken from each frame by applying SPN models on each input. The ground truth of each block is the next frame depending on the input of each block. The detailed specifications of each layer are described in Table II. The five models of SPN represent the encoder part.

C. The 3D CNN Network

The output of each model in the encoder part is passed to the next stage (3D CNN Models). The layers in this model are 3D CONV and B.N. to extract the dynamic sense from input features as explained in Fig. 3. This part of our method contains three models; each model consists of 4 Blocks, as shown in Fig. 4. The Model has five groups of input features of [H, W, F_n] described as [5, H, W, F_n]. H and W are the height and width of the input. The 3D model calculates the dynamic sense of motion depending on time (t) and space.The dimension of time is used to extract the dynamic features and another dimension is not changed to preserve the intensity and colors of pixels. The outputs of the 3D CNN Model are [1, H, W, F_n]. The design of the three models is the same but the difference is the filter numbers. The F_n in the first model is 128 and 64, and 32 in the second and third models respectively. In each model, the F_n is still constant in all layers of models. the aim of applying the 3D CNN layers is to find the proper dynamic features in consecutive frames. Table III describes the details of each layer in the 3D CNN Model.

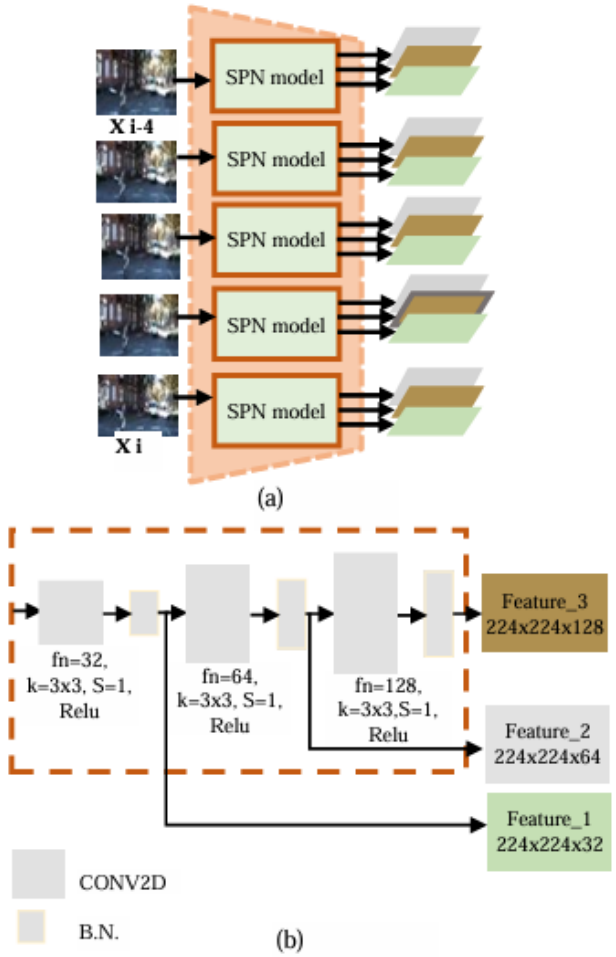

Fig. 2. The SPN architecture (a) The encoder model. (b) The SPN architecture.

D. The Decoder Model

The proposed decoder depends on the skip connection [32], As seen in Table IV which describes all the details of the Decoder part and the second stage of the skip connection. The main purpose of skip connections in a classical AE Model is to reconstruct the details of the target data or object which enhances the performance of the models. From this point, we suggest modifying the connection method between the compatible layers on both sides of AE networks. In our network, these connections are modified by adding 3D CNN-CONVLSTM 2D. The proposed skip connection contains two stages, the first stage is represented by the 3D-CNN model at f_n =32, and 64, and the second stage is the 2D CONVLSTM layer, as shown in Fig. 1. This stage is described in the decoder part as shown in Fig. 5. These connections mitigated the trouble of the information bottleneck [16]. The extracted features from the SPN layers are united with the resulting

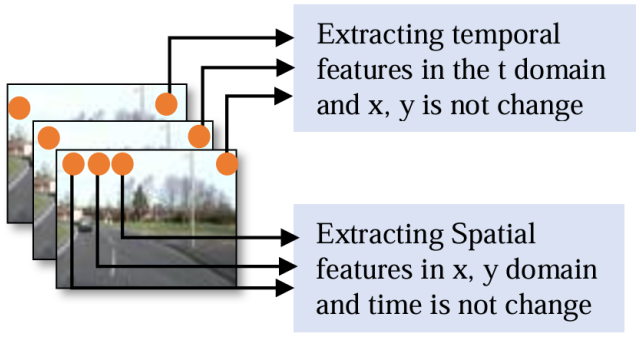

Fig. 3. Spatial and temporal relationship over adjacent Frame

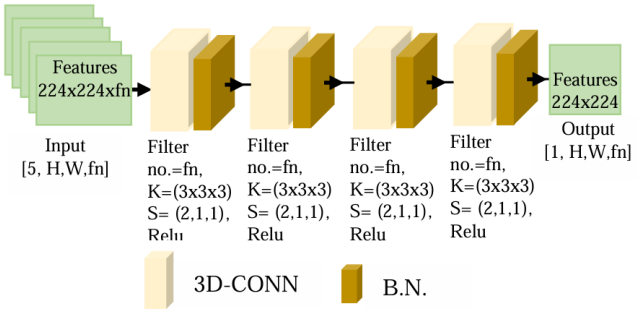

Fig. 4. The 3D CNN model

features of the current and past frames. So, the 3D CNN-CONVLSTM 2D connection generates new data which is an

TABLE II.
SPN MODEL LAYERS

Layer type	Kernelsize	no.Filter(F_n)	Stride
2D CONV	(3,3)	32	(1,1)
B.N.	– – –	– – –	– – –
2D CONV	(3,3)	64	(1,1)
B.N.	– – –	– – –	– – –
2D CONV	(3,3)	128	(1,1)
B.N.	– – –	– – –	– – –

TABLE III.
3D CNN MODEL LAYERS

Layer type	Kernel size	no.Filter(F_n)	Stride
3D CONV	(3,3,3)	F_n	(1,1,1)
B.N.	– – –	– – –	– – –
3D CONV	(3,3,3)	F_n	(2,1,1)
B.N.	– – –	– – –	– – –
3D CONV	(3,3,3)	F_n	(2,1,1)
B.N.	– – –	– – –	– – –
3D CONV	(3,3,3)	F_n	(2,1,1)
B.N.	– – –	– – –	– – –

TABLE IV.
DETAILED SPECIFICATIONS OF EACH LAYER IN DECODER MODEL.

Block type	Input layer	Layer type	Kernel size	No.Filter(F_n)	Stride	output
	$Feature_{128}$	3D CONV	(3,3,3)	128	(1,1,1)	$Frecons_{128}$
	– – –	B.N.	– – –	– – –	– – –	– – –
	$Frecons_{128}$	3D CONV	(3,3,3)	64	(1,1,1)	– – –
	– – –	B.N.	– – –	– – –	– – –	$Frecons_{64_1}$
Skip connection	$Feature_{64}$	2D CONVLSTM	(3,3)	64	(1,1)	$Frecons_{64_2}$
	– – –	Concatenate [$Frecons_{64_1}$,$Frecons_{64_2}$]			– – –	$Fconcat_{128}$
	$Fconcat_{128}$	3D CONV	(3,3,3)	64	(1,1,1)	– – –
	– – –	B.N.	– – –	– – –	– – –	– – –
	– – –	3D CONV	(3,3,3)	32	(1,1,1)	– – –
	– – –	B.N.	– – –	– – –	– – –	$Frecons_{32_1}$
Skip connection	$Feature_{32}$	2D CONVLSTM	(3,3)	32	(1,1)	$Frecons_{32_2}$
	– – –	Concatenate [$Frecons_{32_1}$, $Frecons_{32_2}$]			– – –	$Fconcat_{64}$
	$Fconcat_{64}$	3D CONV	(3,3,3)	64	(1,1,1)	——
	– – –	B.N.	– – –	– – –	– – –	– – –
	– – –	3D-CONV	(3,3,3)	32	(1,1,1)	– – –
	– – –	B.N.	– – –	– – –	– – –	– – –
Output layer	– – –	2D CONV	(3,3)	3	(1,1)	Next frame

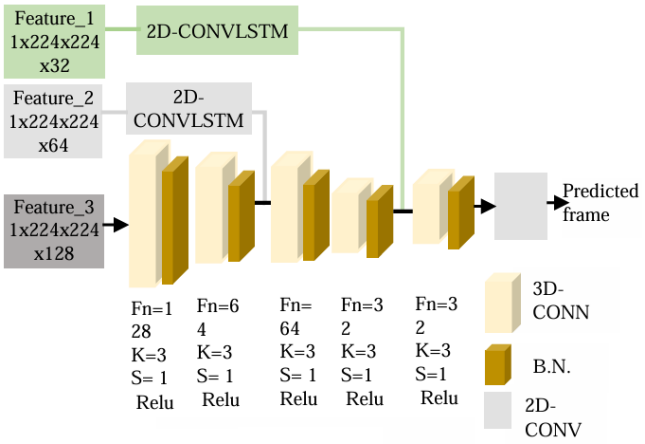

Fig. 5. The decoder architecture

amended feature map to the decoding layer. The enhancement of data connections between the encoder and decoder layers is very useful for forecasting, especially when the next frame is extremely correlated with the previous one. This helps to boost motion continuity. The major purposes of 3D CNN-@D-ConvLSTM skip connections are: (i) Shun the encoder bottleneck to conserve low and mid- levels of information. (ii) Enhance the performance of the model by updating the intermediate connection information which makes the motion flow better.

IV. EVALUATION METRICS

The quality of image is typically measured by the Mean Square Error (MSE), Peak Signal Noise Ratio (PSNR), Structural Similarity Index (SSIM) [33] etc. In this study, we fundamentally applied MSE, PSNR, and SSIM as metrics to estimate model quality. In image compression and other domains, PSNR is frequently employed as a means of signal reconstruction quality monitoring. It is computed using Equation 1 [30].

$$PSNR(GT,PF) = 10\log \frac{255^2}{\sum_{i=0}^{N} (GT^i - PF^i)^2} \quad (1)$$

Where:
PF: represents the predicted frame.
GT: is the ground truth of the frame.
The SSIM is one of the best metrics that deal with the mean of the data to gauge brightness, variance, and covariance factors. The variance is applied to evaluate the contrast, while the covariance is used to assess data structure. These three factors are more closely related to human perception to compare the similarity of images. SSIM is typically employed in jobs requiring the evaluation of image quality such as image super-resolution, image compression, and others, which is determined as Equation 2 [30].

$$SSIM(Y_1,Y_2) = \frac{(2\,\mu_{Y_1}\mu_{Y_2} + c_1)(2\sigma_{Y_2Y_1} + c_2)}{\left(\mu_{Y_1}^2 + \mu_{Y_2}^2 + c_1\right)\left(\sigma_{Y_1}^2 + \sigma_{Y_2}^2 + c_2\right)} \quad (2)$$

Where:
μ_{Y_1}: is the mean of ground truth frame Y_1.
μ_{Y_2}: is the mean of the expected frame.
σ_{Y_1} : is the variance of an image of ground truth frame Y_1.
σ_{Y_2}: is the variance of the predicted frame.
$\sigma_{Y_2Y_1}$:is the covariance of Y_1,Y_2 .
C_1 and C_2 are described in Equation 3 [30].

$$C_1 = (K_1L)^2 \qquad \text{and} \qquad C_2 = (K_2L)^2 \tag{3}$$

C_1 and C_2: are utilized to preserve the stability of the computational procedure.
L: determines the dynamic range of each pixel value. In our approach, we choose k_1 =0.01 and k_2 =0.03 [34]. The values of SSIM are in the range of [-1,1], A high value of SSIM means that more compatibility between the prediction result and the real data or ground truth data. In this model to measure the SSIM in our Python code, the sliding window size is 11 x 11, and the variance of the Gaussian distribution is 1.5.

V. EXPERIMENTS

1) Training Details

Keras and TensorFlow are used to implement the model. To increase the stability of the model, the Batch normalization is applied after each layer except the output layer, and the Adam optimizer is tuning at (learning rate lr = 0.0001; beta1 = 0.9; beta2 = 0.999). The Batch size=5 and epoch=100. The system is executed and applied on an NVIDIA RTX3060 GPU with 12 Giga Bytes memory. The training and testing operation is applied in an end-to-end way, we resize the frames of the Cityscapes and KITTI dataset to 224x224.

A. Datasets

1) KITTI [35]

This is the most public dataset for VP, autonomous driving, and mobile robotics, and is considered a standard set for computer vision models. It consists of 57 videos with 1392x512 RGB pixel resolution based on hours of traffic scripts applied with a different modality of sensors, this dataset included gray-scale stereo cameras, high-resolution RGB, and a 3D-laser scanner [36]. the original dataset did not compose GT for semantic segmentation, because of the popularity of this set, many researchers were encouraged to add parts of the dataset. in 2015, the KITTI dataset was modified by adding a 200-frame for both instance and semantic segmentation in a pixel-level formula [37].

2) Cityscapes [38]

This dataset is very similar to the KITTI dataset and many papers used the KITTI and Cityscapes together. Cityscapes introduces a large-scale database based on 50 videos with 2048x1024 RGB pixel resolution. This dataset contains instance-wise, semantic, and dense pixels for 30 categories grouped into 8 classes of urban street scenes. The dataset is approximately composed of 5K fine-explained images (1 frame in 30 seconds) and 20K annotated coarse ones (one frame every 20-seconds or 20 meters recorded by the car). This set was recorded in 50 different cities spending several months in good conditions weather, and daytime. All frames are produced as stereo pairs. The dataset also consists of extra High-level data like outside temperature, vehicle sensors, and GPS tracks to increase the performance of VP models.

B. Single Frame Prediction

We apply three evaluation metrics to compare our proposed model with prior studies: Peak Signal-to-noise Ratio (PSNR), Mean square error, and Structural Similarity Index Measure (SSIM). Firstly, it must determine the input frames in our model, we choose 5 frame inputs by applying different numbers of frame (3-10) input frames and measure the performance of the system as shown in Table V.

TABLE V.
NUMBER OF FRAMES APPLIED

no.input frames	MSE	PSNR	SSIM
2	0.00159	28.905	0.821
4	0.00139	29.012	0.891
5	0.00101	33.135	0.924
6	0.00101	33.135	0.924
7	0.00101	33.135	0.924
8	0.00101	33.135	0.924
9	0.00101	33.135	0.924
10	0.00125	30.011	0.910

The metric values of PSNR, MSE, and SSIM are generally constant in 5 frames and above. But, when the number of inputs reaches to 10 frames, the prediction performance decreases because the structure of our model cannot extract the suitable features. So, A good next-frame prediction is achieved in the proposed model as shown in Fig. 6, we apply 5-frame inputs. The results are displayed in Table VI compared with other state-of-the-art models applied to predict the next frame and utilized the same data (all state-of-the-art models use the KITTI data set for training and Caltech for testing).

Our proposed system is outperforming in MSE, PSNR, and many parameters in comparison to others. However, PreC-Net [24] outperformed our system in SSIM. That is due to the reduced number of parameters required for training. Our model is qualitatively applied to 10 training repetitions in many frames of the KITTI and Cityscape dataset. The results of different approaches [23], [39], [24] applied the KITTI

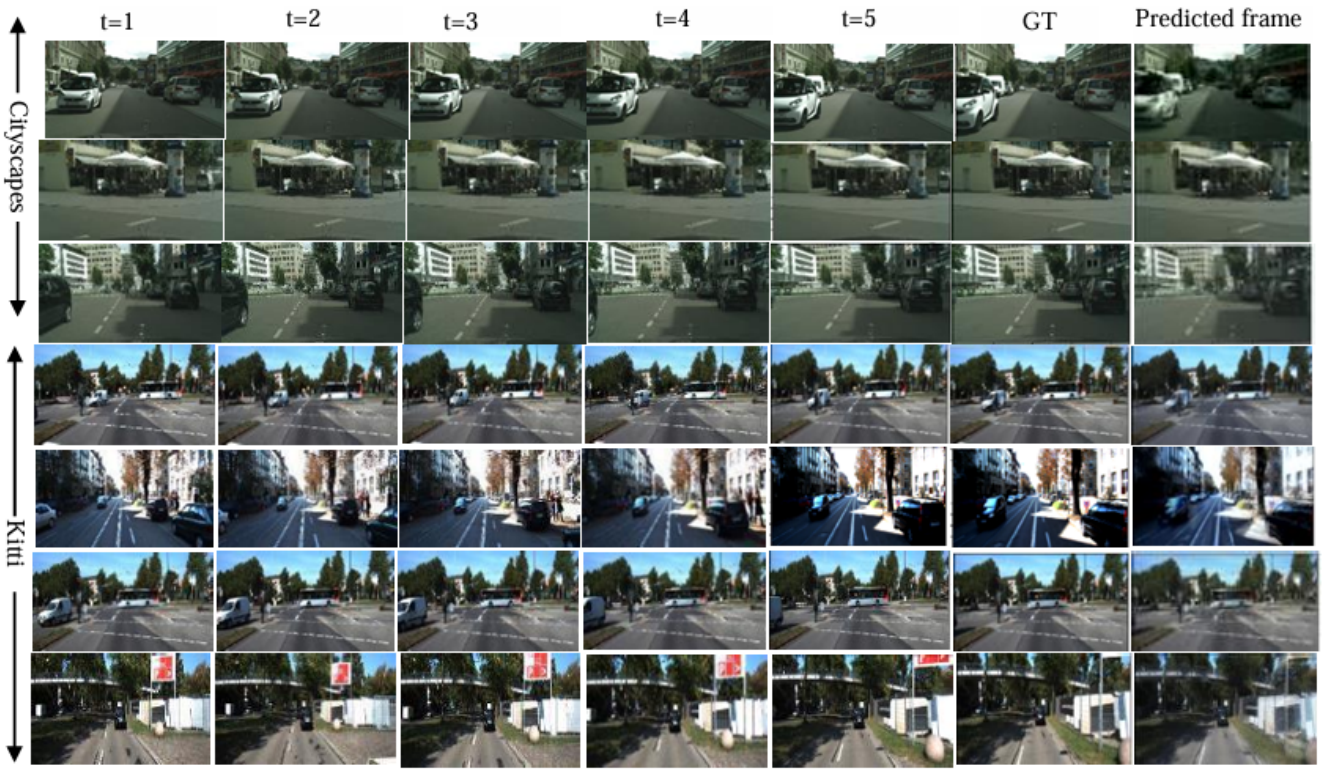

Fig. 6. The qualitative analysis from Cityscapes and KITTI training datasets. The first five columns contain actual frames at t=i-4, i-3, i-2, i-1, and i. The sixth column contains the GT. The final column describes the results of our approach.

TABLE VI.
THE PERFORMANCE ON CALTECH PEDESTRIAN DATASET AFTER TRAINING ON KITTI DATASET

Method	MSE	PSNR	SSIM	no.Par.
last feame [39]	0.0079	23.2	0.779	– – –
Beyond MSE [36]	0.00326	– – –	0.881	– – –
DM-GAN [40]	0.00241	– – –	0.899	113M
CtrlGen [41]	– – –	26.5	0.900	– – –
PredNet [23]	0.00242	27.6	0.905	6.9M
DVF [42]	0.0022	27.9	0.904	3.8M
PreCNet [24]	0.00209	28.3	0.926	7.0M
ContextVP [39]	0.00194	28.7	0.921	8.6M
RC-GAN [43]	0.00161	29.2	0.919	– – –
OURs	0.00101	33.135	0.924	2.3M

and Cityscape dataset as the training set are compared with our study, as shown in Fig. 7. All models recorded a good result but we tend to decrease the blurry prediction as small as possible by preserving on the details of frames. Our results are mostly better than other methods because the modified of skip connection can preserve the details and edges of the frame which tends to be subtler and sharper compared with the GT frame. We observe that our proposed model outperforms PredRNN and PreCNet by extracting motion features and protecting detailed structures for spatial and time steps, while the prediction results of PredRNN, PreCNet, and ContextVP become blurry especially in the edges of objects in the frame as shown in the red and yellow area in Fig. 7.

C. Multi Frame Prediction

To evaluate our approach in forecasting multiple frames, we employed the same single-frame proposed prediction model Fig. reffig:fig2. A next-frame prediction approach is used to access the first 5 frames in each step, i.e., The first five frames at (t=1,2,3,4,5) are applied to predict frame 6 at t=6. Then, the predicted frame is concatenated with another previous frame to predict the seventh frame at t=7. The eighth frame is predicted at t=8 depending on the seventh predicted frame another previous frame, as shown in Fig. 8. A quantitative analysis of the PreCNET model, PredNet model, and RC-GAN in multiple frame prediction is presented in Table VII.

In our approach, the numbers of input frames are 5 sequence frames and the sequence of predicted frames is 5 frames too. The PredNet applied 2 input frames [23], the

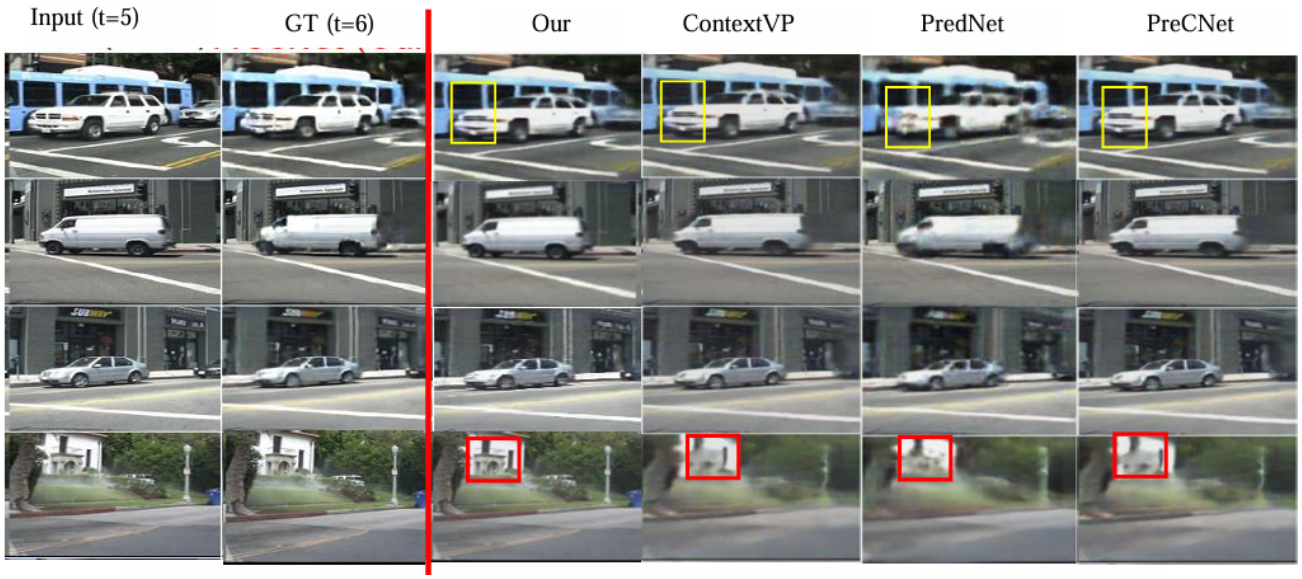

Fig. 7. A qualitative comparison of PreCNet, PredNet, and Context VP models. Based on Caltech Pedestrian Dataset by rows: set07-v011, set10-v010, set10-v010, set06-v009.

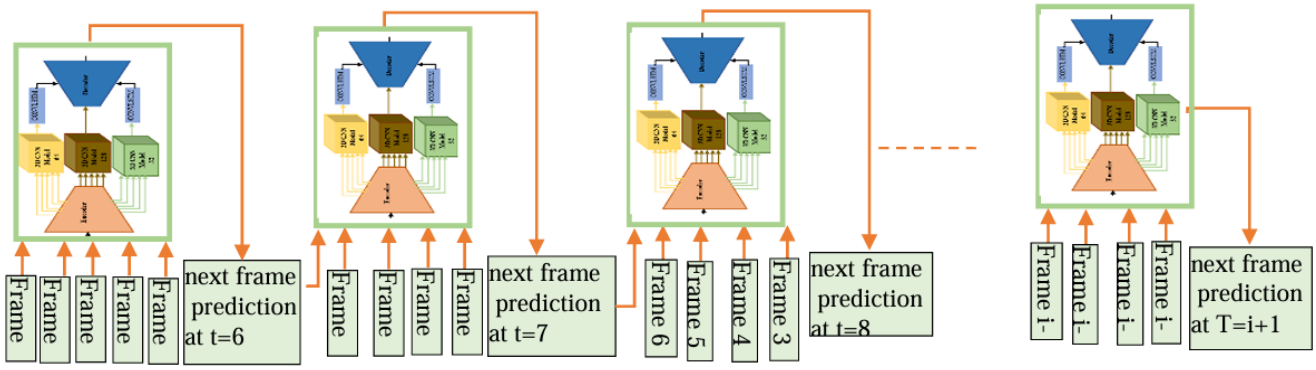

Fig. 8. Our proposed approach in multi-frame prediction

PreCNet recorded the best results in 7 input frames [24], and the RC-GAN [37] utilized 4 frame input. Our model outperforms other state-of-the-art models and recorded the best result in PSNR, and SSIM. Although our model recorded the second rank in SSIM when predicting the first frame prediction, the degradation of results when predicting other frames are a little smaller and the best results compared with other models, the modified skip connection preserved the details of features and this makes the multi-frame prediction is pretty good, as described in Fig. 9.

The results look good until t=10, unfortunately, the quantitative measurements exposed that our approach suffered from a blurry prediction after t=10 because of the accumulated error of predicted frames which were used as input to the next stage of multi-frame predicted models. So, we can say that our model works in 5-frame input and 5-frame output.

VI. Conclusion

This study presents a lightweight method of connection between the analysis and reconstruction sides of the AE model by the cubic 3D CNN-ConvLSTM 2D network. our model depends on the resolution of the spatial-temporal learning by adding a multi-frame as a group of input data to estimate the deep-in-time structure. This strengthens and enhances the dynamics feature which is very important in video prediction applications. The modified of skip connection by adding a cubic 3D CNN-ConvLSTM 2D alleviates and decreases the vanishing gradient problem. The proposed 3D CNN mid-level features play as a bottleneck between the encoder and the decoder part to capture the proper and suitable dynamic features.

TABLE VII.
THE QUANTITATIVE ANALYSIS FOR MULTI FRAME MODELS

Method	metric	$t=6$	$t=7$	$t=8$
PredNET [23]	PSNR	27.6	21.7	20.3
	SSIM	0.905	0.72	0.66
RC-GAN [37]	PSNR	29.2	25.9	22.3
	SSIM	0.91	0.83	0.73
PreCNet [24]	PSNR	28.5	23.4	20.2
	SSIM	0.93	0.82	0.69
our	PSNR	33.135	26.108	23.802
	SSIM	0.924	0.818	0.792

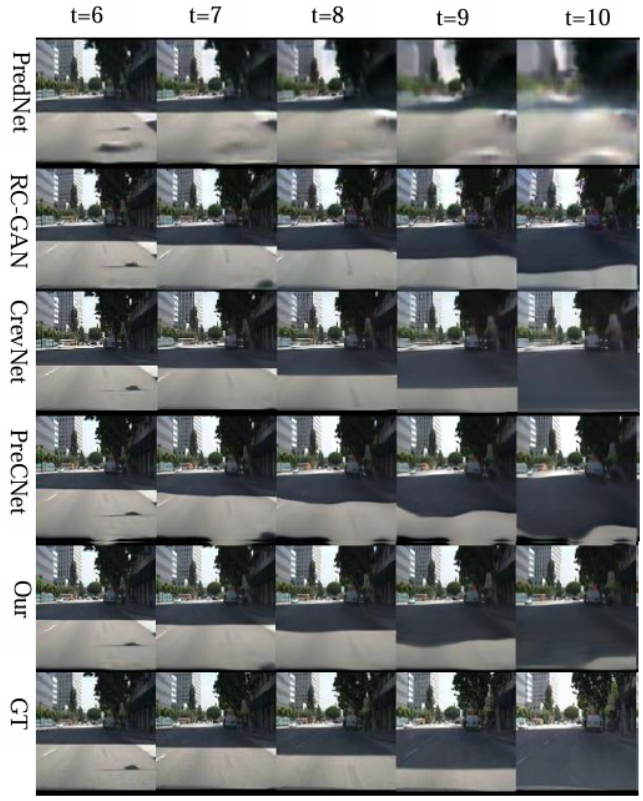

Fig. 9. A qualitative evaluation of multi frame prediction algorithms that were selected. The inputs of PredNet and PreCNet are 10 frames, and RC-GAN used 4 input sequences. Our approach is applied 5 input frames.

These two additional blocks increase the prediction performance. The basic idea of our model is to define different levels of spatial features by 2DCONV layers representing in the SPN model, and then use 3DCONV layers to estimate the temporal dynamics features at each hierarchical level. The low and mid features are passed into 3DCNN-CONVLSTM2D skip connection which represents the novel step of our model and reinforces the dynamics features between the encoder and decoder. The high-level 3DCNN represents the bottleneck between the encoder and the decoder parts. These additional blocks enhance the experimental results and get the best performance in PSNR and MSE when compared with other state-of-the-art models. In single-frame prediction, the model ranked first in PSNR and second with SSIM=0.924. The proposed approach was training on a widely used benchmark dataset. i.e., KITTI and Cityscapes for training, Caltech Pedestrian Dataset for testing, which contains videos from complex environments listed from a car-mounted camera. n the other side, in multi-frame prediction, the results of our approach are suitable at 5 frame prediction and more accurate and better than some of the competitors. A quantitative comparison describes that our approach records high PSNR with the small number of parameters in five output predictions and the multi-frame prediction results are slowly degrading with less blurred frame prediction. We plan to enhance the results of our model by adding a 3DCNN-CONVLSTM Skip connection in U-net architecture instead of a direct skip connection and show how much this idea can increase the SSIM and PNSR. In multi-frame prediction, we need to increase the number of frames that can be predicted (more than $t = 10$) and keep high resolution in the output frame by enhancing the PSNR and SSIM.

CONFLICT OF INTEREST

The author have no conflict of relevant interest to this article.

REFERENCES

[1] K. Xu, L. Wen, G. Li, L. Bo, and Q. Huang, "Spatiotemporal cnn for video object segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1379–1388, 2019.

[2] M. S. Pavel, H. Schulz, and S. Behnke, "Object class segmentation of rgb-d video using recurrent convolutional neural networks," *Neural Networks*, vol. 88, pp. 105–113, 2017.

[3] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

[4] Q. M. Rahman, N. Sünderhauf, P. Corke, and F. Dayoub, "Fsnet: A failure detection framework for semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3030–3037, 2022.

[5] L.-Y. Gui, Y.-X. Wang, D. Ramanan, and J. M. Moura, "Few-shot human motion prediction via meta-learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 432–450, 2018.

[6] Z. Chang, X. Zhang, S. Wang, S. Ma, Y. Ye, X. Xinguang, and W. Gao, "Mau: A motion-aware unit for video prediction and beyond," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 26950–26962, Curran Associates, Inc., 2021.

[7] Z. Zou, R. Zhang, S. Shen, G. Pandey, P. Chakravarty, A. Parchami, and H. X. Liu, "Real-time full-stack traffic scene perception for autonomous driving with roadside cameras," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 890–896, IEEE, 2022.

[8] S. Tulsiani, A. A. Efros, and J. Malik, "Multi-view consistency as supervisory signal for learning shape and pose prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2897–2905, 2018.

[9] X. Chen, Y. Jia, X. Tong, and Z. Li, "Research on pedestrian detection and deepsort tracking in front of intelligent vehicle based on deep learning," *Sustainability*, vol. 14, no. 15, p. 9281, 2022.

[10] L. Chen, I. Grimstead, D. Bell, J. Karanka, L. Dimond, P. James, L. Smith, and A. Edwardes, "Estimating vehicle and pedestrian activity from town and city traffic cameras," *Sensors*, vol. 21, no. 13, p. 4564, 2021.

[11] P. Hewage, M. Trovati, E. Pereira, and A. Behera, "Deep learning-based effective fine-grained weather forecasting model," *Pattern Analysis and Applications*, vol. 24, no. 1, pp. 343–366, 2021.

[12] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine, "Stochastic adversarial video prediction," *arXiv preprint arXiv:1804.01523*, 2018.

[13] E. L. Denton *et al.*, "Unsupervised learning of disentangled representations from video," *Advances in neural information processing systems*, vol. 30, 2017.

[14] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," *arXiv preprint arXiv:1706.08033*, 2017.

[15] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Argyros, "A review on deep learning techniques for video prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2806–2826, 2020.

[16] K. Fan, C. Joung, and S. Baek, "Sequence-to-sequence video prediction by learning hierarchical representations," *Applied Sciences*, vol. 10, no. 22, p. 8288, 2020.

[17] B. Liu, Y. Chen, S. Liu, and H.-S. Kim, "Deep learning in latent space for video prediction and compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 701–710, 2021.

[18] V.-T. Le and Y.-G. Kim, "Attention-based residual autoencoder for video anomaly detection," *Applied Intelligence*, vol. 53, no. 3, pp. 3240–3254, 2023.

[19] Y. Lu, K. M. Kumar, S. shahabeddin Nabavi, and Y. Wang, "Future frame prediction using convolutional vrnn for anomaly detection," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–8, IEEE, 2019.

[20] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Deep convolutional autoencoder-based lossy image compression," in *2018 Picture Coding Symposium (PCS)*, pp. 253–257, IEEE, 2018.

[21] P. Desai, C. Sujatha, S. Chakraborty, S. Ansuman, S. Bhandari, and S. Kardiguddi, "Next frame prediction using convlstm," in *Journal of Physics: Conference Series*, vol. 2161, p. 012024, IOP Publishing, 2022.

[22] 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*, 2018.

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

[24] Z. Straka, T. Svoboda, and M. Hoffmann, "Precnet: Next-frame video prediction based on predictive coding," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[25] X. Ye and G.-A. Bilodeau, "A unified model for continuous conditional video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3603–3612, 2023.

[26] Z. Gao, C. Tan, L. Wu, and S. Z. Li, "Simvp: Simpler yet better video prediction," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3170–3180, 2022.

[27] H. Wei, X. Yin, and P. Lin, "Novel video prediction for large-scale scene using optical flow," *arXiv preprint arXiv:1805.12243*, 2018.

[28] R. Zhang, X. Shu, R. Yan, J. Zhang, and Y. Song, "Skip-attention encoder–decoder framework for human motion prediction," *Multimedia Systems*, pp. 1–10, 2022.

[29] J. Cho, J. Lee, C. Oh, W. Song, and K. Sohn, "Wide and narrow: Video prediction from context and motion," *arXiv preprint arXiv:2110.11586*, 2021.

[30] 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.

[31] X. Ye and G.-A. Bilodeau, "Video prediction by efficient transformers," *Image and Vision Computing*, vol. 130, p. 104612, 2023.

[32] J. Santokhi, P. Daga, J. Sarwar, A. Jordan, and E. Hewage, "Temporal autoencoder with u-net style skip-connections for frame prediction," *arXiv preprint arXiv:2011.12661*, 2020.

[33] M. A. Yilmaz and A. M. Tekalp, "Effect of architectures and training methods on the performance of learned video frame prediction," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4210–4214, IEEE, 2019.

[34] N. Shayanfar, V. Derhami, and M. Rezaeian, "Video prediction using multi-scale deep neural networks," *Journal of AI and Data Mining*, vol. 10, no. 3, pp. 423–431, 2022.

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

[36] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," 2016.

[37] Z. Chang, X. Zhang, S. Wang, S. Ma, Y. Ye, and W. Gao, "Stae: A spatiotemporal auto-encoder for high-resolution video prediction," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2021.

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

[39] W. Byeon, Q. Wang, R. K. Srivastava, and P. Koumoutsakos, "Contextvp: Fully context-aware video prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[40] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion gan for future-flow embedded video prediction," in *proceedings of the IEEE international conference on computer vision*, pp. 1744–1752, 2017.

[41] Z. Hao, X. Huang, and S. Belongie, "Controllable video generation with sparse trajectories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7854–7863, 2018.

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

[43] Y.-H. Kwon and M.-G. Park, "Predicting future frames using retrospective cycle gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1811–1820, 2019.