All Predictions Matter: an Online Video Prediction Approach

Melan Vijayaratnam, Marco Cagnazzo, Giuseppe Valenzise, Enzo Tartaglione

All Predictions Matter: an Online Video Prediction Approach

Melan Vijayaratnam*, Marco Cagnazzo*†, Giuseppe Valenzise‡ and Enzo Tartaglione*

*LTCI, Télécom Paris, Institut Polytechnique de Paris, France
†University of Padua, Department of Information Engineering, Italy
‡Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, France

***Abstract*—To effectively manage and utilize the massive amount of visual data generated by the surging number of videos, decision-making systems must predict and reason about future outcomes. This paper proposes a novel online approach for video prediction that enables continual learning in the presence of new data, as periodic training of neural networks may not be practical. We utilize all predictions, including intermediate computations obtained during the inference process, to improve the performance of video prediction. To achieve this, we incorporate a weighting scheme in the loss that accounts for all the predictions during the learning process. Additionally, we leverage semantic segmentation to assess the performance of extrapolated frames by focusing on the position of the objects in the scene. Our approach stands out from state-of-the-art methods as it uses intermediate predictions, which are available due to the iterative nature of forecasting future frames. Our method improves the offline counterpart for the same network by 1.45 dB for predicting five steps in the future.**

***Index Terms*—Extrapolation, video prediction, online learning, metric, segmentation**

I. Introduction

The human capacity to forecast future events and adapt present behavior accordingly is a well-established phenomenon in cognitive psychology and behavioral sciences [1]. As such, expecting the same for systems is key for understanding about the world that surrounds us. The applications of video prediction range from assisting in medical diagnosis [2], for autonomous driving to help the car to anticipate and react to potential hazards on the road [3] to low-latency video transmission [4]. Being a self-supervised task, the understanding only comes from the data itself, preventing the need for data labeling efforts.

Online deep learning methods have been presented as a way to scale with the stream of data [5]. It has been studied more specifically in various fields of computed vision, from classification [6] to semantic segmentation [7]. In the presented work of Zhang *et al.* [8], the authors apply online learning to video depth estimation that would normally require labeled data for the network to be updated but they devise a technique to be able to do so in a self-supervised way. Video prediction networks also benefit from the online learning paradigm, which encourage to present a novel methodology that can apply to all such networks.

To enhance the accuracy of video prediction for distant future sequences, we utilize intermediate frames in the prediction process. These intermediate frames are saved and combined with new ground truth images as they are received to update the model based on a weighting of all predictions in the loss computation. Overall, our method improves the performance of video prediction, particularly for longer temporal horizons, resulting in more accurate predictions of future frames in a video sequence.

Furthermore, we present a method for evaluating video prediction algorithms at the object level. We accomplish this by borrowing from the field of semantic segmentation and creating a pseudo label segmented image from the ground truth, which we then compare to extrapolated frames. As a result, we focus on the objects in the scenes and their locations rather than the entire scene.

II. Related Work

In the typical setup of evaluation of deep learning architectures, the model weights are typically learned on the training set, the hyperparameters are fine-tuned on the validation set, and the pre-trained weights are used on the test set. This is the batch-learning strategy, where the system learns the model only once. Before being deployed, the model is pre-trained offline, and afterward, it is frozen. Online learning techniques differ in that they continuously update and improve a model's performance as new data becomes available. The model is trained on a stream of data, with each new observation providing an opportunity for the model to learn and adapt in real-time. Interest in online learning has emerged for classification tasks [9], formally introduced in [10]. Existing video extrapolation methods [11] only considered the batch learning paradigm. Our target use case, on the other hand, has a critical distinction that allows us to progress toward a more effective framework. More precisely, for any image predicted by the extrapolator, its ground truth (the actual image) will arrive and allow a refinement of the neural network. This idea naturally leads us toward the on-line learning paradigm. The approach we propose in this paper is based on online learning [12] and allows the system to learn the model on the fly which means keeping learning even after being deployed as new data arrives.

Since video prediction is a self-supervised task [13], there is no need for human annotation as the information is already

present in the data. Zhang *et al.* [8] apply online adaptation to to consider the task of depth estimation as a self-supervised task in a self-supervised manner not to require depth data explicitly and adapt to evolving data streams. Later, the concept was developed for online monocular depth estimation [14]. Online learning has been shown to improve streaming policies [15]. Our work is connected to these studies as they employ video depth estimation in an online environment, similar to our objective of developing video extrapolation networks that function with online streams of video sequences.

III. Online video prediction scheme

In certain applications, such as compensating for latency through extrapolation [4], it is essential to have the ability to make predictions at a specific horizon in the future. The horizon h is defined as the number of frames we want to predict in the future. As it is well known in the literature, the larger h, the more difficult it is to get a reliable prediction. To address the decrease in prediction accuracy when dealing with large values of h, a frequently employed approach involves the iterative application of the prediction network. This entails making predictions for future frames within a shorter time horizon, and subsequently using these predictions as input to the prediction network to extrapolate frames farther away in time [11]. This iteration process can be exploited in online learning by defining a loss function that employs a weighted mean of the errors of each intermediate prediction. In an online setting, it means that as soon as a new frame from the sequence arrives, multiple forward passes coming from all approximations of the new images will occur.

Figure 1 presents the proposed scheme for online video prediction. To predict the sequence stream ahead of h frames, we start from the pre-trained weights resulting from the training process. By storing past predictions of $\hat{I}_n^h$, i.e., the predicted frames of the ground truth frame I at time step n using horizon h, we use them later when the ground truth arrives to update the prediction network. The input frames from the video prediction network, namely the context frames, can be either true (available) frames, or predicted frames from the iterative process. At each time step, the video extrapolation network $\mathcal{F}$, which would be fixed in an offline learning scenario, is updated. We denote as $\mathcal{F}_n$ the updated model at time n. By following the depicted process, the extrapolated frames for I_n can be obtained as follows (assuming as an example that $\mathcal{F}$ takes 2 context frames as input):

$$\hat{I}_n^1 = \mathcal{F}_{n-1}(I_{n-1}, I_{n-2}) \tag{1}$$

$$\hat{I}_n^2 = \mathcal{F}_{n-2}\left(\mathcal{F}_{n-2}(I_{n-2}, I_{n-3}), I_{n-2}\right) \tag{2}$$

$$\hat{I}_n^3 = \mathcal{F}_{n-3}\left(\mathcal{F}_{n-3}\left(\mathcal{F}_{n-3}(I_{n-3}, I_{n-4}), I_{n-3}\right), \hat{I}_{n-2}^1\right) \tag{3}$$

More in general, when we recursively re-circulate the last predicted output back as input h times in order to predict h steps in the future, frame n is predicted h times: at time $n-1$, $n-2, \ldots, n-h$. We can define a new overall loss $\mathcal{L}^*$ that takes advantage on one hand of all these intermediate predictions, and on the other of the availability ground truth frames:

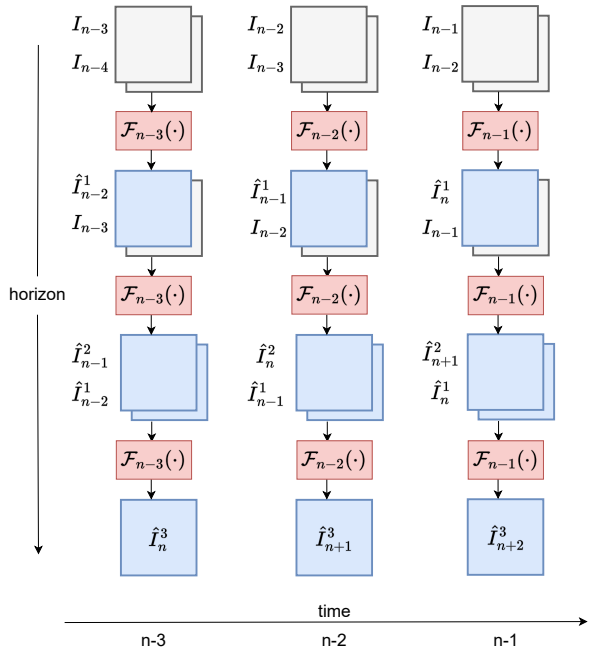


Fig. 1: Prediction of 3 steps in the future. The vertical axis represents how far we want to predict in the future and the horizontal one represents the stream of data arriving. Ground truth frames are depicted in gray and predicted frames in blue. The ground truth frames allow to get the predicted sequence 3 steps in the future $\hat{I}_n^3$, $\hat{I}_{n+1}^3$, $\hat{I}_{n+2}^3$. All the intermediate computed frames will be used to update the network as ground truth arrives.

$$\mathcal{L}^* = \sum_{i=1}^{h} \lambda_i \, \mathcal{L}\left(\hat{I}_n^i; I_n\right), \tag{4}$$

where λ refers to the weight assigned to each of the different predictions. The loss $\mathcal{L}^*$ is a weighted sum of all the per-frame losses $\mathcal{L}\left(\hat{I}_n^i; I_n\right)$ over the horizon h where $\mathcal{L}\left(\hat{I}_n^i; I_n\right)$ is often a mean squared error, but other relevant loss metrics can be used. At the arrival of new ground truth frame, the network will update itself with the loss with the formulation in Equation 4.

IV. Semantic segmentation based metric

PSNR has been criticized for not being a good objective fidelity metric [16]. Regardless, it is still widely popular and used to compare different frames from videos. It relies on every pixel of the reference frame and compares it to a target frame. Using methods from the semantic segmentation field [17], we may further verify the accuracy of the pixels at the object level in the scene. Semantic segmentation involves assigning per-pixel predictions of object categories to an image, providing a comprehensive description of the scene that includes information about the object category, location,

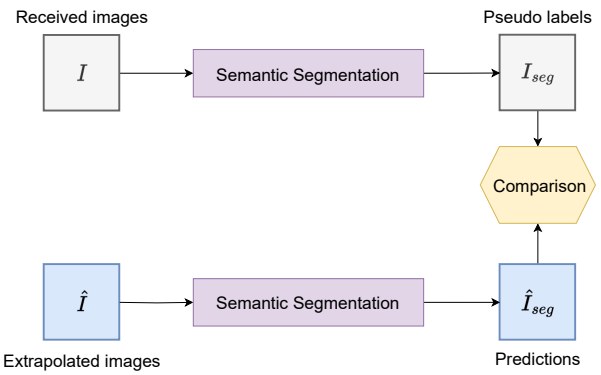


Fig. 2: Semantic segmentation based metric for video prediction.

and shape. By applying semantic segmentation to the images in question, we can observe the positions of the objects and confirm the observations made earlier. We elaborate on the method to evaluate the extrapolation methods using semantic segmentation. We demonstrate a method to evaluate the extrapolation methods using a semantic segmentation on Figure 2 scheme. The received frames I are fed to a semantic segmentation network to generate pseudo labels (since the ground truth is not provided) for segmentation $\tilde{I}_{seg}$, and compared to the segmentation maps $\hat{I}_{seg}$ computed on the extrapolated images $\hat{I}$. For this evaluation, we choose DeepLabv3+ [18] pre-trained on Cityscapes, having Resnet-101 as backbone. The adopted evaluation metric is the Intersection-Over-Union, denoted as IoU: this is a method to quantify the overlap between the target segmentation mask and our prediction segmentation output over the union of both quantities.

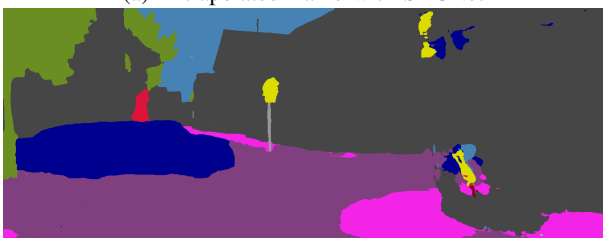
V. Experiments and Discussion

In the following experiments, we analyze the effect of the proposed online video prediction technique and evaluate them using common metrics and the segmentation-based metric introduced in this work.

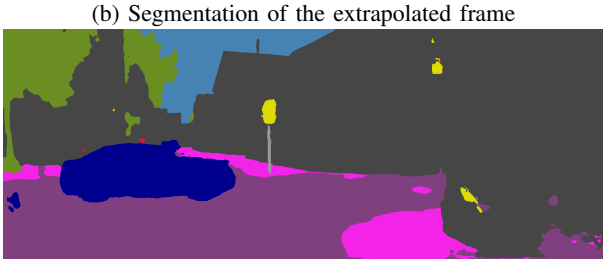
A. Datasets

We train the learning-based extrapolation methods, MCNet [19] and SDCNet [20] on the Caltech Pedestrian dataset [21], collected from a vehicle driving through regular traffic in an urban environment. The dataset consists of around 10 hours of dashcam footage with 65 different video sequences captured at 30 fps. We additionally use sequences from the Kitti [22] and DriveSeg [23] manual scene for evaluation purposes which are both datasets taken the same way as Caltech pedestrian from a moving vehicle. We use the sequence #14 from Kitti consisting of 320 frames and the first 500 frames of DriveSeg. Regarding the optical flow-based method FlowNet2 [24], we only make use of the pre-trained weights on MPI-Sintel which is an optical flow data set derived from the film Sintel [25].

(a) Extrapolated frame with SDCNet



(b) Segmentation of the extrapolated frame



(c) Segmentation of the true image

Fig. 3: Segmentation outputs for predicting one step in the future. Image taken from the Kitti dataset.

B. Choice of video prediction networks

As discussed in [26], video prediction methods can be motion-based, pixel-based, or fusion-based. Motion-based methods focus on the motion in the image which could be done with the optical flow information. Pixel-based methods generate the entirety of the pixels from scratch and finally, fusion-based methods combine both motion and pixel-based methods. We choose a technique from each class, starting with FlowNet2 [24] for predicting optical flow. Combined with a warping that moves the pixels according to the optical flow, an estimate of the next image can be obtained. MCNet [19] uses long short-term memory modules from image differences to generate a new frame. SDCNet [20] uses both optical flows and convolutional kernels from the pixels to generate the extrapolated frame. We perform offline experiments that correspond to having the weights of the neural network being frozen at validation as well as online experiments on SDCNet, with weights learning during validation.

We also include a simple frame-copy extrapolation, dubbed CopyLast. This method just copies the last available frame. Although it is not a real extrapolation method, it is often used as a reference. In particular, for understanding the visual quality of the prediction: if the predicted image is not better than CopyLast, it means that we are introducing large artifacts.

Approach	PSNR ↑			SSIM ↑			VMAF ↑		
	h=1	h=3	h=5	h=1	h=3	h=5	h=1	h=3	h=5
CopyLast	21.25	18.87	17.96	0.50	0.42	0.40	16.12	9.33	8.05
MCNet	23.19	20.66	19.36	0.60	0.52	0.49	19.84	8.91	6.47
FlowNet2 + warp	24.92	21.44	20.03	0.73	0.53	0.48	32.55	10.89	7.04
SDCNet offline	25.38	23.18	22.06	0.76	0.68	0.65	39.59	24.51	18.37
SDCNet online (ours)	**26.53**	**24.07**	**22.73**	**0.83**	**0.75**	**0.71**	**51.27**	**32.86**	**24.55**

(a) Quantitative results on Kitti scene 014

Approach	PSNR ↑			SSIM ↑			VMAF ↑		
	h=1	h=3	h=5	h=1	h=3	h=5	h=1	h=3	h=5
CopyLast	27.65	23.64	22.21	0.72	0.54	0.45	47.34	29.22	22.49
MCNet	28.84	25.20	22.68	0.89	0.74	0.61	61.05	40.78	27.70
FlowNet2 + warp	31.82	27.00	24.72	0.92	0.79	0.65	71.77	42.26	26.86
SDCNet offline	34.23	29.93	28.21	0.95	0.88	0.83	80.44	56.91	45.23
SDCNet online (ours)	**35.89**	**31.71**	**29.66**	**0.98**	**0.93**	**0.89**	**87.58**	**69.48**	**57.64**

(b) Quantitative results on DriveSeg

TABLE I: Comparison of the proposed online method with other extrapolation methods

C. Experimental results

In Table I we observe the PSNR in the YCbCr color space [27], SSIM [28], and VMAF [29], as they are widely used objective metrics. The reference extrapolated video is compared to the original input sequences. CopyLast serves as a simple baseline that uses the last available frame and corresponds to not anticipating the future while FlowNet2 combined with a warping allows predicting the future frames. For every extrapolation horizon, the weights are reinitialized from the pre-trained weights. The weights assigned to the λ are chosen so that $\lambda_i = 1 \, \forall i$, signifying that each of the parts of the sum given in the equation 4 has equal importance. The online proposed method applied to SDCNet outperforms the same network in offline mode by 0.89 dB in Kitti and 1.78 dB in DriveSeg at horizon $h = 3$, meaning predicting three steps in the future, which results in a latency compensation of 100 ms.

D. Ablation study

We perform multiple experiments to validate our proposed online approach for video prediction. To do so, we compare our proposed method, which we call "Uniform" due to the equal importance to every predictions. "First only" corresponds to considering the first prediction only and "Last only" only the last prediction. Table II shows that our approach outperforms the competing approaches, and proves the proposed approach of considering every prediction is beneficial to the network. At $h = 1$, the methods behave the same due to having a single weighting term, therefore we do not report these results as these can be found in Table I.

E. Discussion about segmentation

In Table III, we report the intersection over union (IoU) of the class "car", which is predominant in the chosen sequences. In the Kitti scene, the IoU seems to follow the same trend as the PSNR and demonstrates that the online adaptation brings an increase in performance. Concerning DriveSeg, the IoU from both methods are very close, which contradicts the PSNR results of the online outperforming CopyLast. Upon further examination, it was discovered that in the Kitti dataset, the moving cars are spaced further apart from each other compared to the DriveSeg dataset where the cars are closely grouped together. The image in Figure 3 displays an issue caused by extrapolation at the back of the car, resulting in the segmentation network incorrectly categorizing this artifact as a car.

Weighting λ_i	PSNR ↑			
	h=2	h=3	h=4	h=5
First only	24.95	23.99	23.27	22.66
Last Only	24.91	23.89	23.09	22.46
Uniform	**25.05**	**24.07**	**23.37**	**22.73**

(a) Kitti scene

Weighting λ_i	PSNR ↑			
	h=2	h=3	h=4	h=5
First only	33.24	31.59	30.46	29.57
Last Only	33.20	31.38	30.12	29.21
Uniform	**33.32**	**31.71**	**30.58**	**29.66**

(b) DriveSeg

TABLE II: Ablation study on the weighting in the online scheme

VI. Conclusion

This paper introduces an online learning algorithm for video prediction. We exploit every prediction to improve the video extrapolation network and not just the resulting frames of the desired horizon. This comes at the price of additional complexity by making use of intermediate and unused pre-

Approach	IoU car ↑ h=1	h=2	h=3	h=4	h=5
CopyLast	0.50	0.29	0.19	0.16	0.17
MCNet	0.38	0.20	0.14	0.09	0.18
FlowNet2 + warp	0.70	0.53	0.40	0.30	0.22
SDCNet offline	0.69	0.56	0.45	0.34	0.23
Ours	**0.72**	**0.58**	**0.55**	**0.48**	**0.29**

(a) IoU for Kitti scene 14

Approach	IoU car ↑ h=1	h=2	h=3	h=4	h=5
CopyLast	0.87	0.83	0.80	0.78	0.75
MCNet	0.80	0.72	0.67	0.64	0.58
FlowNet2 + warp	0.86	0.83	0.79	0.76	0.73
SDCNet offline	0.86	0.80	0.73	0.73	0.69
Ours	**0.88**	**0.84**	**0.81**	**0.78**	**0.76**

(b) IoU for DriveSeg

TABLE III: Intersection over Union comparison between CopyLast and extrapolation methods over Kitti and DriveSeg for the car class.

dicted frames but with an increase in quality as demonstrated by the experiments. The segmentation-oriented quality metric focusing on the object rather than every pixel also seems promising and may stimulate further work towards enforcing shape consistency of objects in difficult environments.

VII. Acknowledgments

This work was funded by the ANR AAPG2020 national fund (ANR-20-CE25-0014).

References

[1] T. Suddendorf and J. Redshaw, "Anticipation of future events," *Encyclopedia of animal cognition and behavior*, pp. 1–9, 2017.

[2] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, et al., "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[3] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 33–47, 2020.

[4] M. Vijayaratnam, M. Cagnazzo, G. Valenzise, A. Trioux, and M. Kieffer, "Towards zero-latency video transmission through frame extrapolation," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2122–2126.

[5] D. Sahoo, Q. Pham, J. Lu, and S. C. Hoi, "Online deep learning: Learning deep neural networks on the fly," *arXiv preprint arXiv:1711.03705*, 2017.

[6] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, 2022.

[7] R. Volpi, P. De Jorge, D. Larlus, and G. Csurka, "On the road to online adaptation for semantic image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19184–19195.

[8] Z. Zhang, S. Lathuiliere, A. Pilzer, N. Sebe, E. Ricci, and J. Yang, "Online adaptation through meta-learning for stereo depth estimation," *arXiv preprint arXiv:1904.08462*, 2019.

[9] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3366–3385, 2021.

[10] S. Shalev-Shwartz et al., "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[11] 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.

[12] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao, "Online learning: A comprehensive survey," *Neurocomputing*, vol. 459, pp. 249–289, 2021.

[13] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1476–1485.

[14] Z. Zhang, S. Lathuiliere, E. Ricci, N. Sebe, Y. Yan, and J. Yang, "Online depth learning against forgetting in monocular videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4494–4503.

[15] T. Karagkioules, G. S. Paschos, N. Liakopoulos, A. Fiandrotti, D. Tsilimantos, and M. Cagnazzo, "Online learning for adaptive video streaming in mobile networks," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 1, pp. 1–22, 2022.

[16] K. Navas, D. K. Gayathri, M. Athulya, and A. Vasudev, "Mwpsnr: A new image fidelity metric," in *2011 IEEE Recent Advances in Intelligent Computational Systems*. IEEE, 2011, pp. 627–632.

[17] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[19] R. Villegas, J. Yang, S. Hong, et al., "Decomposing motion and content for natural video sequence prediction," *arXiv preprint arXiv:1706.08033*, 2017.

[20] F. A. Reda, G. Liu, K. J. Shih, R. Kirby, J. Barker, D. Tarjan, A. Tao, and B. Catanzaro, "Sdc-net: Video prediction using spatially-displaced convolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 718–733.

[21] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 304–311.

[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[23] L. Ding, J. Terwilliger, R. Sherony, et al., "MIT driveseg (manual) dataset for dynamic driving scene segmentation," Tech. Rep., Technical report, Massachusetts Institute of Technology, 2020.

[24] E. Ilg, N. Mayer, T. Saikia, et al., "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2462–2470.

[25] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer, 2012, pp. 611–625.

[26] H. Gao, H. Xu, Q.-Z. Cai, R. Wang, F. Yu, and T. Darrell, "Disentangling propagation and generation for video prediction," in *IEEE International Conf. on Computer Vision (ICCV)*, 2019, pp. 9006–9015.

[27] G. Sullivan and K. Minoo, "Objective quality metric and alternative methods for measuring coding efficiency," in *document JCTVC-H0012, ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC), 8th Meeting: San Jose, CA, USA*, 2012, pp. 1–10.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[29] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal Feature Integration and Model Fusion for Full Reference Video Quality Assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Aug. 2019.