A Review on Deep Learning Techniques for Video Prediction

Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros

Abstract—The ability to predict, anticipate and reason about future outcomes is a key component of intelligent decision-making systems. In light of the success of deep learning in computer vision, deep-learning-based video prediction emerged as a promising research direction. Defined as a self-supervised learning task, video prediction represents a suitable framework for representation learning, as it demonstrated potential capabilities for extracting meaningful representations of the underlying patterns in natural videos. Motivated by the increasing interest in this task, we provide a review on the deep learning methods for prediction in video sequences. We firstly define the video prediction fundamentals, as well as mandatory background concepts and the most used datasets. Next, we carefully analyze existing video prediction models organized according to a proposed taxonomy, highlighting their contributions and their significance in the field. The summary of the datasets and methods is accompanied with experimental results that facilitate the assessment of the state of the art on a quantitative basis. The paper is summarized by drawing some general conclusions, identifying open research challenges and by pointing out future research directions.

Index Terms—Video prediction, future frame prediction, deep learning, representation learning, self-supervised learning

1 Introduction

Will the car hit the pedestrian? That might be one of the questions that comes to our minds when we observe Figure 1. Answering this question might be in principle a hard task; however, if we take a careful look into the image sequence we may notice subtle clues that can help us predicting into the future, e.g., the person's body indicates that he is running fast enough so he will be able to escape the car's trajectory. This example is just one situation among many others in which predicting future frames in video is useful. In general terms, the prediction and anticipation of future events is a key component of intelligent decision-making systems. Despite the fact that we, humans, solve this problem quite easily and effortlessly, it is extremely challenging from a machine's point of view. Some of the factors that contribute to such complexity are occlusions, camera movement, lighting conditions, clutter, or object deformations. Even so, video prediction models are able to extract rich spatio-temporal features from natural videos in a self-supervised fashion. This was fostered by the great strides deep learning has made in different research fields such as human action recognition and prediction [1], semantic segmentation [2], and registration [3], to name a few. Because of their ability to learn adequate representations from high-dimensional data [4], deep learning-based models fit perfectly into the learning by prediction paradigm.

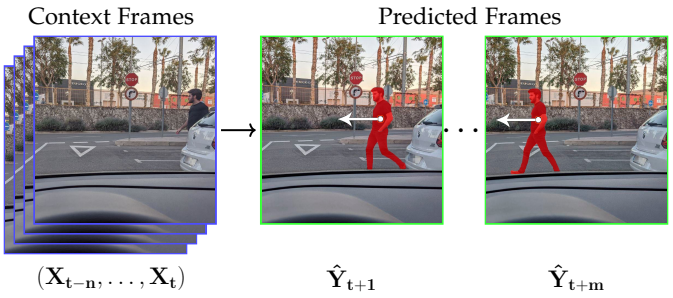


Fig. 1. A pedestrian appeared from behind the white car with the intention of crossing the street. The autonomous car must make a call: hit the emergency braking routine or not. This all comes down to predict the next frames $(\hat{Y}_{t+1}, \ldots, \hat{Y}_{t+m})$ given a sequence of context frames $(X_{t-n}, \ldots, X_t)$, where n and m denote the number of context and predicted frames, respectively. From these predictions at a representation level (RGB, high-level semantics, etc.) a decision-making system would make the car avoid the collision.

1.1 Application Domains

Video prediction methods have been successfully applied in a broad range of application domains such as robotics, autonomous driving, action anticipation, and more. Relying on action-conditioned video prediction, robots were able to successfully manipulate previously unseen objects [5]. In the same domain, video prediction has facilitated decision making in vision-based robotic control [6] and motion planning [7], [8] and has provided accurate world models, of high-dimensional environments, to model-based Reinforcement Learning (RL) approaches. For instance, video prediction has enabled planning in unknown environments [9], [10]. This has helped model-based RL to achieve similar

- *S. Oprea, P. M.-Gonzalez, J.A. C.-Vargas, and J. G.-Rodriguez are with the Department of Computer Technology. S. O.-Escolano is with the Department of Computer Science and Artificial Intelligence. University of Alicante, San Vicente del Raspeig, E-03690, Spain. E-mail:{soprea,pmartinez,jacastro,jgarcia}@dtic.ua.es, sorts@dccia.ua.es.*
- *A. G.-Garcia is with the Institute of Space Sciences (ICE-CSIC), Campus UAB, Barcelona, E-08193, Spain. E-mail: garciagarcia@ice.csic.es.*
- *A. Argyros is with the Institute of Computer Science, FORTH, Heraklion GR-70013, Greece and with the Computer Science Department, University of Crete, Heraklion, Greece. E-mail: argyros@ics.forth.gr.*

Manuscript received April 19, 2005; revised August 26, 2015.

or better results compared to model-free approaches, with fewer interactions and improved generalization capabilities.

Regarding self-driving cars, the trajectory prediction in traffic of pedestrians [11] or generic agents [12] is extremely useful to anticipate future events. Furthermore, the probabilistic prediction of multi-modal futures [13] demonstrated great success when it comes to traffic uncertainty. Likewise, the synergy between video prediction and action anticipation was successfully proven with the prediction of visual embeddings [14] and motion representations [15]. Some other tasks in which video prediction has been applied successfully are: prediction of instance/semantic segmentation maps [16], [17], [18], anomaly detection [19], precipitation nowcasting [20], [21], and video interpolation [22].

1.2 Review Scope and Terminology

In this review, we put our focus on deep learning techniques and how they have been extended or applied to video prediction. We limit this review to the future video prediction given the context of a sequence of previous frames, leaving aside methods that predict future from a static image. In this context, the terms video prediction, future frame prediction, next video frame prediction, future frame forecasting, and future frame generation are used interchangeably. To the best of our knowledge, this is the first review in the literature that focuses on video prediction using deep learning techniques.

2 Video Prediction

Besides its biological roots, video prediction draws inspiration from computational models of the predictive coding paradigm [23], [24], [25], [26]. Predictive coding states that human brain builds complex mental representations of the physical and causal rules that govern the world. This arises from the conceptual acquisition and the accumulation of background knowledge from early ages, primarily through observation and interaction [27], [28], [29]. From a brain processing perspective, these mental representations are continuously updated through the prediction of raw sensory inputs. The brain refines the already understood world models from the mismatch between its predictions and the actual sensory input [30].

2.1 Problem Definition

Video prediction closely captures the essence of the predictive coding paradigm. On this basis, video prediction is defined as the task of inferring the subsequent frames in a video, based on a sequence of previous frames used as a context. Let $\mathbf{X}_t \in \mathrm{R}^{w \times h \times c}$ be the t-th frame in the video sequence $\mathbf{X} = (X_{t-n}, \ldots, X_{t-1}, X_t)$ with n frames, where w, h, and c denote width, height, and number of channels, respectively. The target is to predict the next m frames $\mathbf{Y} = (\hat{Y}_{t+1}, \hat{Y}_{t+2}, \ldots, \hat{Y}_{t+m})$ from the input $\mathbf{X}$.

Different from video generation that is mostly unconditioned, video prediction is conditioned on a previously learned representation from a sequence of input frames. At a first glance, and in the context of learning paradigms, one can think about the video prediction task as a supervised learning approach because the target frame acts as a label. However, as this information is already available in the input video sequence, no extra labels or human supervision is needed. Therefore, learning by prediction is a self-supervised task, filling the gap between supervised and unsupervised learning.

Under the assumption that good predictions can only be the result of accurate representations, learning by prediction is a feasible approach to verify how accurately the system has learned the underlying patterns in the input data. In other words, it represents a suitable framework for representation learning [31], [32]. Furthermore, because of its potential to extract meaningful representations from video sequences, video prediction is an excellent intermediate step between natural videos and decision-making.

2.2 Exploiting the Time Dimension of Videos

Unlike static images, videos provide complex transformations and motion patterns ordered in the time dimension. Focusing on a small image patch in the same spatial location through consecutive time steps, a wide range of visually similar local deformations are identified due to the temporal coherence. In contrast, when looking at the big picture, the consecutive frames are visually different but semantically coherent. The variability in the visual appearance of a video at different scales, is mainly due to occlusions, changes in the lighting conditions, and camera motion, among other factors. From this source of temporally ordered visual cues, predictive models are able to extract representative spatio-temporal correlations depicting the dynamics in a video sequence. For instance, Agrawal *et al.* [33] established a direct link between vision and motion, attempting to reduce supervision efforts when training deep predictive models.

The importance of the time dimension in video understanding models has been well studied [34]. The implicit temporal ordering in videos, also known as the arrow of time, indicates whether a video sequence is playing forward or backward. Using this temporal direction as a supervisory signal [35], [36], [37] further encouraged predictive models to implicitly or explicitly model spatio-temporal correlations of a video sequence to understand the dynamics of a scene. The time dimension of a video reduces the supervision effort and makes the prediction task self-supervised.

2.3 Dealing with Stochasticity

Predicting how a square is moving, could be extremely challenging even in a deterministic environment such as the one represented in Figure 2. The lack of contextual information and the multiple equally probable outcomes hinder the prediction task. But, what if we use two consecutive frames as context? Under this configuration and assuming a physically perfect environment, the square will be indefinitely moving in the same direction. This represents a deterministic outcome, an assumption that many authors made in order to deal with future uncertainty. Assuming a deterministic outcome would narrow the prediction space to a unique solution. However, this assumption is not suitable for natural videos. The future is by nature multimodal, since the probability distribution defining all the possible future outcomes in a context has multiple modes, i.e. there are multiple equally probable and valid outcomes. Furthermore, on

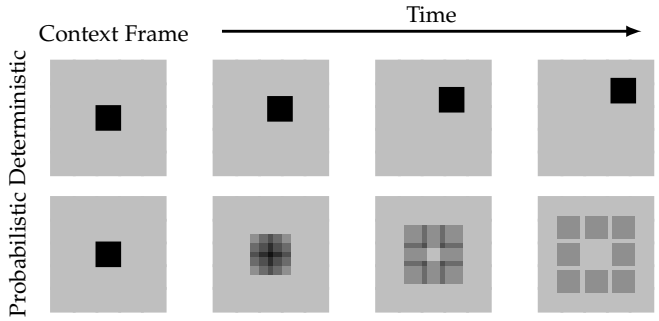


Fig. 2. At top, a deterministic environment where a geometric object, e.g. a black square, starts moving following a random direction. At bottom, probabilistic outcome. Darker areas correspond to higher probability outcomes. As uncertainty is introduced, probabilities get blurry and averaged. Figure inspired by [38].

the basis of a deterministic universe, we indirectly assume that all possible outcomes are reflected in the input data. These assumptions make the prediction under uncertainty an extremely challenging task.

Most of the existing deep learning-based models in the literature are deterministic. Although the future is uncertain, a deterministic prediction would suffice some easily predictable situations. For instance, most of the movement of a car is largely deterministic, while only a small part is uncertain. However, when multiple predictions are equally probable, a deterministic model will learn to average between all the possible outcomes. This averaging effect depends on the loss function and is visually represented in predictions as blurriness, specially on long time horizons. However, it can be mitigated by constructing a loss function that does not lead to averaging. As deterministic models are unable to handle real-world settings characterized by chaotic dynamics, authors considered that incorporating uncertainty to the model is a crucial aspect. Probabilistic approaches dealing with these issues are discussed in Section 4.6.

2.4 The Devil is in the Loss Function

The design and selection of the loss function for the video prediction task is of utmost importance. Pixel-wise losses, e.g. ℓ_2, ℓ_1 and Mean-Squared Error (MSE), are widely used in both unstructured and structured predictions. Although leading to plausible predictions in deterministic scenarios, such as synthetic datasets and video games, they struggle with the inherent uncertainty of natural videos. In a probabilistic environment, with different equally probable outcomes, pixel-wise losses aim to accommodate uncertainty by blurring the prediction, as we can observe in Figure 2. In other words, the deterministic loss functions average out multiple equally plausible outcomes in a single, blurred prediction. In the pixel space, these losses are unstable to slight deformations and fail to capture discriminative representations to efficiently regress the broad range of possible outcomes. This makes difficult to draw predictions maintaining the consistency with our visual similarity notion. A recent study [39] performed an in-depth analysis of the generalization capabilities of different loss functions for the video prediction task. Besides video prediciton, the impact of different loss functions was analyzed in image restoration restoration [40], classification [41], camera pose regression [42] and structured prediction [43], among others. This fosters reasoning about the importance of the loss function, particularly when making long-term predictions in high-dimensional and multimodal natural videos.

Most of distance-based loss functions, such as based on ℓ_p norm, come from the assumption that data is drawn from a Gaussian distribution. But, how these loss functions address multimodal distributions? Assuming that a pixel is drawn from a bimodal distribution with two equally likely modes Mo_1 and Mo_2, the mean value $\overline{Mo} = (Mo_1 + Mo_2)/2$ would minimize the ℓ_p-based losses over the data, even if $\overline{Mo}$ has very low probability [44]. This suggests that the average of two equally probable outcomes would minimize distance-based losses such as, the MSE loss. However, this applies to a lesser extent when using ℓ_1 norm as the pixel values would be the median of the two equally likely modes in the distribution. In contrast to the ℓ_2 norm that emphasizes outliers with the squaring term, the ℓ_1 promotes sparsity thus making it more suitable for prediction in high-dimensional data [44]. Based on the ℓ_2 norm, the MSE is also commonly used in the training of video prediction models. However, it produces low reconstruction errors by merely averaging all the possible outcomes in a blurry prediction as uncertainty is introduced. In other words, the mean image would minimize the MSE error as it is the global optimum, thus avoiding finer details such as facial features and subtle movements as they are noise for the model. Most of the video prediction approaches rely on pixel-wise loss functions, obtaining roughly accurate predictions in easily predictable datasets.

One of the ultimate goals of many video prediction approaches is to palliate the blurry predictions when it comes to uncertainty. For this purpose, authors broadly focused on: directly improving the loss functions; exploring adversarial training; alleviating the training process by reformulating the problem in a higher-level space; or exploring probabilistic alternatives. Some promising results were reported by combining the loss functions with sophisticated regularization terms, e.g. the Gradient Difference Loss (GDL) to enhance prediction sharpness [44] and the Total Variation (TV) regularization to reduce visual artifacts and enforce coherence [22]. Perceptual losses were also used to further improve the visual quality of the predictions [45], [46], [47], [48], [49]. However, in light of the success of the Generative Adversarial Networks (GANs), adversarial training emerged as a promising alternative to disambiguate between multiple equally probable modes. It was widely used in conjunction with different distance-based losses such as: MSE [50], ℓ_2 [51], [52], [53], or a combination of them [44], [54], [55], [56], [57], [58]. To alleviate the training process, many authors reformulated the optimization process in a higher-level space (see Section 4.5). While great strides have been made to mitigate blurriness, most of the existing approaches still rely on distance-based loss functions. As a consequence, the regress-to-the-mean problem remains an open issue. This has further encouraged authors to reformulate existing deterministic models in a probabilistic fashion.

TABLE 1
Summary of the most widely used datasets for video prediction (**S**/**R**: **S**ynthetic/**R**eal, **st**: **st**ereo, **de**: **de**pth, **ss**: **s**emantic **s**egmentation, **is**: **i**nstance **s**egmentation, **sem**: **sem**antic, **I**/**O**: **I**ndoor/**O**utdoor environment, **bb**: **b**ounding **b**ox, **Act**: **Act**ion label, **ann**: **ann**otated, **env**: **env**ironment, **ToF**: **T**ime **o**f **F**light, **vp**: camera **v**iew**p**oints respect human).

								provided data and ground-truth						
name[1]	**year**	**S/R**	**#videos**	**#frames**	**#ann. frames**	**resolution**	**#classes**	**RGB**	**st**	**de**	**ss**	**is**	**other annotations**	**env.**
Action and human pose recognition datasets														
KTH [59]	2004	R	2391	250 000[2]	0	160 × 120	6 (action)	✓	✗	✗	✗	✗	Act.	O
Weizmann [60]	2007	R	90	9000[2]	0	180 × 144	10 (action)	✓	✗	✗	✗	✗	Act.	O
HMDB-51 [61]	2011	R	6766	639 300	0	*var* × 240	51 (action)	✓	✗	✗	✗	✗	Act., vp	I/O
UCF101 [62]	2012	R	13 320	2 000 000[2]	0	320 × 240	101 (action)	✓	✗	✗	✗	✗	Act.	I/O
Penn Action D. [63]	2013	R	2326	163 841	0	480 × 270	15 (action)	✓	✗	✗	✗	✗	Act., Human poses, vp	I/O
Human3.6M [64]	2014	SR	4000[2]	3 600 000	0	1000x1000	15 (action)	✓	✗	ToF	✗	✗	Act., Human poses & meshes	I/O
THUMOS-15 [65]	2017	R	18 404	3 000 000[2]	0	320 × 240	101 (action)	✓	✗	✗	✗	✗	Act., Time span	I/O
Driving and urban scene understanding datasets														
Camvid [66]	2008	R	5	18 202	701 (ss)	960 × 720	32 (sem)	✓	✗	✗	✓	✗	✗	O
CalTech Pedest. [67]	2009	R	137	1 000 000[2]	250 000 (bb)	640 × 480	-	✓	✗	✗	✗	✗	Pedestrian bb & occlusions	O
Kitti [68]	2013	R	151	48 791	200 (ss)	1392 × 512	30 (sem)	✓	✓	LiDAR	✓	✓	Odometry	O
Cityscapes [69]	2016	R	50	7 000 000[2]	25 000 (ss)	2048 × 1024	30 (sem)	✓	✓	stereo	✓	✓	Odometry, temp, GPS	O
Comma.ai [70]	2016	R	11	522 000[2]	0	160 × 320	-	✓	✗	✗	✗	✗	Steering angles & speed	O
Apolloscape [71]	2018	R	4	200 000	146 997 (ss)	3384 × 2710	25 (sem)	✓	✓	LiDAR	✓	✓	Odometry, GPS	O
nuScenes [72]	2019	R	1000	1 400 000	40 000 (bb, ss)	1600 × 900	32 (sem)	✓	✗	LiDAR	✓	✗	Radar, Odometry, GPS	O
Waymo Open D. [73]	2020	R	1950	200 000	200 000 (bb)	1920 × 1280	4 (sem)	✓	✗	LiDAR	✗	✗	Odometry, 2D/3D bb	O
Object and video classification datasets														
Sports1m [74]	2014	R	1 133 158	n/a	0	640 × 360 (*var.*)	487 (sport)	✓	✗	✗	✗	✗	Sport label	I/O
YouTube8M [75]	2016	R	8 200 000	n/a	0	*variable*	1000 (topic)	✓	✗	✗	✗	✗	Topic label, Segment info	I/O
YFCC100M [76]	2016	SR	8000	n/a	0	*variable*	-	✓	✗	✗	✗	✗	User tags, Localization	I/O
Video prediction datasets														
Bouncing balls [77]	2008	S	4000	20 000	0	150 × 150	-	✓	✗	✗	✗	✗	✗	-
Van Hateren [78]	2012	R	56	3584	0	128 × 128	-	✓	✗	✗	✗	✗	✗	I/O
NORBvideos [79]	2013	R	110 560	552 800	All (is)	640 × 480	5 (object)	✓	✗	✗	✗	✓	✗	I
Moving MNIST [80]	2015	SR	*custom*[3]	*custom*[3]	0	64 × 64	-	✓	✗	✗	✗	✗	✗	-
Robotic Pushing [5]	2016	R	57 000	1 500 000[2]	0	640 × 512	-	✓	✗	✗	✗	✗	Arm pose	I
BAIR Robot [81]	2017	R	45 000	n/a	0	n/a	-	✓	✗	✗	✗	✗	Arm pose	I
RoboNet [82]	2019	R	161 000	15 000 000	0	*variable*	-	✓	✗	✗	✗	✗	Arm pose	I
Other-purpose and multi-purpose datasets														
ViSOR [83]	2010	R	1529	1 360 000[2]	0	*variable*	-	✓	✗	✗	✗	✗	User tags, human bb	I/O
PROST [84]	2010	R	4 (10)	4936 (9296)	All (bb)	*variable*	-	✓	✗	✗	✗	✗	Object bb	I
Arcade Learning [85]	2013	S	*custom*[3]	*custom*[3]	0	210 × 160	-	✓	✗	✗	✗	✗	✗	-
Inria 3DMovie v2 [86]	2016	R	27	2476	235 (is)	960 × 540	-	✓	✓	✗	✗	✓	Human poses, bb	I/O
Robotrix [87]	2018	S	67	3 039 252	All (ss)	1920 × 1080	39 (sem)	✓	✗	✓	✓	✓	Normal maps, 6D poses	I
UASOL [88]	2019	R	33	165 365	0	2280 × 1282	-	✓	✓	stereo	✗	✗	✗	O

[1] some dataset names have been abbreviated to enhance table's readability.
[2] values estimated based on the framerate and the total number of frames or videos, as the original values are not provided by the authors.
[3] *custom* indicates that as many frames as needed can be generated. This is related to datasets generated from a game, algorithm or simulation, involving interaction or randomness.

3 DATASETS

As video prediction models are mostly self-supervised, they need video sequences as input data. However, some video prediction methods rely on extra supervisory signals, e.g. segmentation maps, and human poses. This makes out-of-domain video datasets perfectly suitable for video prediction. Table 1 shows an overview of the most used datasets for video prediction. Detailed descriptions for each one of them can be found in the supplementary material.

4 VIDEO PREDICTION METHODS

In the video prediction literature we find a broad range of different methods and approaches. Early models focused on directly predicting raw pixel intensities, by implicitly modeling scene dynamics and low-level details (Section 4.1). However, extracting a meaningful and robust representation from raw videos is challenging, since the pixel space is highly dimensional and extremely variable. From this point, reducing the supervision effort and the representation dimensionality emerged as a natural evolution. On the one hand, the authors aimed to disentangle the factors of variation from the visual content, i.e. factoring the prediction space. For this purpose, they: (1) formulated the prediction problem into an intermediate transformation space by explicitly modeling the source of variability as transformations between frames (Section 4.2); (2) separated motion from the visual content with a two-stream computation (Section 4.3). On the other hand, some models narrowed the output space by conditioning the predictions on extra variables (Section 4.4), or reformulating the problem in a higher-level space (Section 4.5). High-level representations are increasingly more attractive for intelligent systems to support their decision making. For instance, the semantic segmentation space is easily interpretable as the pixels are categorical, in contrast to unprocessed videos where pixels represent raw intensities. Besides simplifying the prediction task, some other works addressed the future uncertainty in predictions. As the vast majority of video prediction models are deterministic, they are unable to manage probabilistic environments. To address this issue, several authors proposed modeling future uncertainty with probabilistic

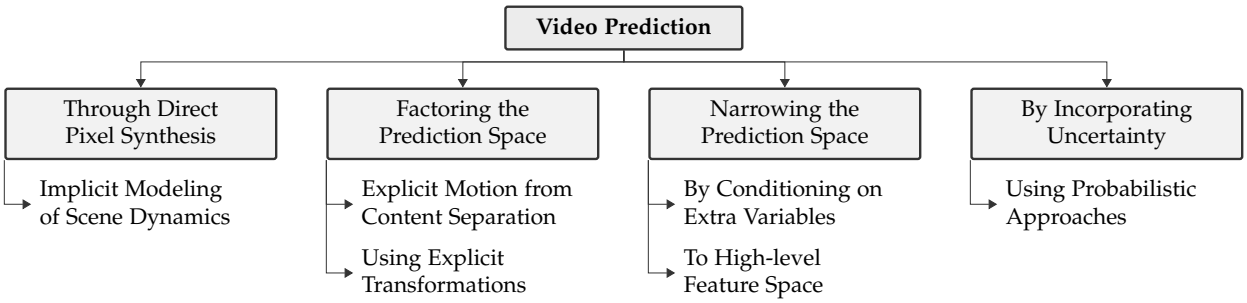


Fig. 3. Classification of video prediction models.

models (Section 4.6).

So far in the literature, there is no specific taxonomy that classifies video prediction models. In this review, we have classified the existing methods according to the video prediction problem they addressed and following the classification illustrated in Figure 3. For simplicity, each subsection extends directly the last level in the taxonomy. The taxonomy is not mutually exclusive, as some methods can be classified into several categories since they address multiple goals. For instance, [17], [55], [89] are probabilistic models making predictions in a high-level space as they address both the future uncertainty and high dimensionality in videos. The category of these models was specified according to their main contribution.

The most relevant methods, ordered in a chronological order, are summarized in Table 2 containing low-level details. From these details such as, the backbone architecture and loss functions, we could easily identify whether a model is probabilistic or deterministic. Furthermore, to better understand the foundations of such methods, we have included a section on backbone deep learning architectures in the supplementary material. Prediction is a widely discussed topic in different fields and at different levels of abstraction. For instance, the future prediction from a static image [90], [91], [92], [93], [94], [95], human action prediction [1], and model-based RL [9], [10], [96] are a different but inspiring research fields.

Although related, the aforementioned topics are outside the scope of this particular review, as it focuses purely on the video prediction methods using a sequence of previous frames as context.

4.1 Direct Pixel Synthesis

Initial video prediction models attempted to directly predict future pixel intensities without any explicit modeling of the scene dynamics. Ranzato *et al.* [97] discretized video frames in patch clusters using k-means. They assumed that non-overlapping patches are equally different in a k-means discretized space, yet similarities can be found between patches. The method is a convolutional extension of a Recurrent Neural Network (RNN)-based model [98] making short-term predictions at the patch-level using Cross Entropy (CE) loss. As the full-resolution frame is a composition of the predicted patches, some tilling effect can be noticed. Predictions of large and fast-moving objects are accurate, however, when it comes to small and slow-moving objects there is still room for improvement. These are common issues for most methods making predictions at the patch-level. Addressing longer-term predictions, Srivastava *et al.* [80] proposed several Autoencoder (AE)-based approaches incorporating Long Short-Term Memory (LSTM) units to model the temporal coherence. Using convolutional [99] and flow [100] percepts alongside RGB image patches, authors tested the models on multi-domain tasks and considered both unconditioned and conditioned decoder versions. The latter only marginally improved the prediction accuracy. Replacing the fully connected LSTMs with convolutional LSTMs, Shi *et al.* proposed an end-to-end model efficiently exploiting spatial correlations [20]. This enhanced prediction accuracy and reduced the number of parameters.

Inspired by adversarial training: Building on the recent success of the Laplacian Generative Adversarial Networks (LAPGANs), Mathieu *et al.* proposed the first multi-scale architecture for video prediction that was trained in an adversarial fashion [44]. Their novel GDL regularization combined with ℓ_1-based reconstruction and adversarial training represented a leap over the previous state-of-the-art models [80], [97] in terms of prediction sharpness. However, it was outperformed by the Predictive Coding Network (PredNet) [70] which stacked several convolutional LSTMs (ConvLSTMs) vertically connected by a bottom-up propagation of the local ℓ_1 error computed at each level. Previously to PredNet, the same authors proposed the Predictive Generative Network (PGN) [50], an end-to-end model trained with a weighted combination of adversarial loss and MSE on synthetic data. Using a similar training strategy as [44], Zhou *et al.* used a convolutional AE to learn long-term dependencies from time-lapse videos [101]. Built on Progressively Growing GANs (PGGANs) [102], Aigner *et al.* proposed the FutureGAN [103], a three-dimensional (3d) convolutional Encoder-decoder (ED)-based model. They used the Wasserstein GAN with gradient penalty (WGAN-GP) loss [104] and conducted experiments on increasingly complex datasets. Extending [20], Zhang *et al.* proposed a novel LSTM-based architecture where hidden states are updated along a z-order curve [105]. Dealing with distortion and temporal inconsistency in predictions and inspired by the Human Visual System (HVS), Jin *et al.* [106] first incorporated multi-frequency analysis into the video prediction task to decompose images into low and high frequency bands. High-fidelity and temporally consistent predictions with the ground truth were reported outperforming state of the art. Distortion and blurriness are further accentuated when it

comes to predict under fast camera motions. To this end, Shouno [107] implemented a hierarchical residual network with top-down connections. Leveraging parallel prediction at multiple scales, authors reported finer details and textures under fast and large camera motion.

Bidirectional flow: Under the assumption that video sequences are symmetric in time, Kwon *et al.* [108] explored a retrospective prediction scheme training a generator also on reversed input sequences. Their cycle GAN-based approach ensures the consistency of bidirectional prediction through retrospective cycle constraints. Similarly, Hu *et al.* [58] proposed a novel cycle-consistency loss used to train a GAN-based approach (VPGAN). Future frames are generated from a sequence of context frames and their variation in time, denoted as Z. Under the assumption that Z is symmetric in the encoding space, it is manipulated to generate desirable moving directions. In the same spirit, other works focused on both, forward and backward predictions [36], [109]. Enabling state sharing between the encoder and decoder, Oliu *et al.* proposed the folded Recurrent Neural Network (fRNN) [110], a recurrent AE architecture featuring Gated Recurrent Units (GRUs) that implement a bidirectional flow of the information. The model demonstrated a stratified representation, which makes the topology more explainable, as well as efficient compared to regular AEs in terms or memory consumption and computational requirements.

Exploiting 3D convolutions: for modeling short-term features, Wang *et al.* [111] integrated them into a recurrent network demonstrating state-of-the-art results on both video prediction and early activity recognition. While 3D convolutions efficiently preserves local dynamics, RNNs enables long-range video reasoning. Their *eidetic* 3d LSTM (E3d-LSTM) network features a gated-controlled self-attention module, i.e. *eidetic* 3D memory, that effectively manages historical memory records across multiple time steps. Outperforming previous works, Yu *et al.* proposed the Conditionally Reversible Network (CrevNet) [112] consisting of two modules, an invertible AE and a Reversible Predictive Model (RPM). While the bijective two-way AE ensures no information loss and reduces the memory consumption, the RPM extends the reversibility from spatial to temporal domain. Some other works used 3D convolutional operations to model the time dimension [103].

Analyzing the previous works, Byeon *et al.* [113] identified a lack of spatial-temporal context in the representations, fact that leads to blurry results when dealing with uncertainty. Although authors addressed this contextual limitation with dilated convolutions and multi-scale architectures, the context representation progressively vanishes in long-term predictions. To address this issue, they proposed a context-aware model that efficiently aggregates per-pixel contextual information at each layer and in multiple directions. The core of their proposal is a context-aware layer consisting of two blocks, one aggregating the information from multiple directions and the other blending them into a unified context.

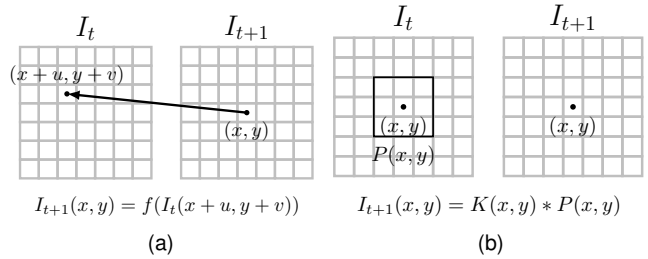


Fig. 4. Representation of transformation-based approaches. (a) Vector-based with a bilinear interpolation. (b) Kernel-based applying transformations as a convolutional operation. Figure inspired by [114].

4.2 Using Explicit Transformations

Let $\mathbf{X} = (X_{t-n}, \ldots, X_{t-1}, X_t)$ be a video sequence of n frames, where t denotes time. Instead of learning the visual appearance, transformation-based approaches assume that visual information is already available in the input sequence. To deal with the strong similarity and pixel redundancy between successive frames, these methods explicitly model the transformations that takes a frame at time t to the frame at $t+1$. These models are formally defined as follows:

$$\mathbf{Y}_{t+1} = \mathcal{T}\left(\mathcal{G}\left(\mathbf{X}_{t-n:t}\right), \mathbf{X}_{t-n:t}\right), \tag{1}$$

where $\mathcal{G}$ is a learned function that outputs future transformation parameters, which applied to the last observed frame $\mathbf{X}_t$ using the function $\mathcal{T}$, generates the future frame prediction $\mathbf{Y}_{t+1}$. According to the classification of Reda *et al.* [114], $\mathcal{T}$ function can be defined as a vector-based resampling such as bilinear sampling, or adaptive kernel-based resampling, e.g. using convolutional operations. For instance, a bilinear sampling operation is defined as:

$$\mathbf{Y}_{t+1}(x, y) = f\left(\mathbf{X}_t(x+u, y+v)\right), \tag{2}$$

where f is a bilinear interpolator such as [22], [115], [116], (u, v) is a motion vector predicted by $\mathcal{G}$, and $X_t(x, y)$ is a pixel value at (x,y) in the last observed frame X_t. Approaches following this formulation are categorized as vector-based resampling operations and are depicted in Figure 4a. On the other side, in the kernel-based resampling, the $\mathcal{G}$ function predicts the kernel $\mathrm{K}(x, y)$ which is applied as a convolution operation using $\mathcal{T}$, as depicted in Figure 4b and is mathematically represented as follows:

$$\mathbf{Y}_{t+1}(x, y) = \mathrm{K}(x, y) * \mathbf{P}_t(x, y), \tag{3}$$

where $\mathrm{K}(x, y) \in \mathbb{R}^{NxN}$ is the 2D kernel predicted by the function $\mathcal{G}$ and $P_t(x, y)$ is an $N \times N$ patch centered at (x, y).

Combining kernel and vector-based resampling into a hybrid solution, Reda *et al.* [114] proposed the Spatially Displaced Convolution (SDC) module that synthesizes high-resolution images applying a learned per-pixel motion vector and kernel at a displaced location in the source image. Their 3D Convolutional Neural Network (CNN) model trained on synthetic data and featuring the SDC modules, reported promising predictions of a high-fidelity.

4.2.1 Vector-based Resampling

Bilinear models use multiplicative interactions to extract transformations from pairs of observations in order to relate

images, such as Gated Autoencoders (GAEs) [117]. Inspired by these models, Michalski *et al.* proposed the Predictive Gating Pyramid (PGP) [118] consisting of a recurrent pyramid of stacked GAEs. To the best of our knowledge, this was the first attempt to predict future frames in the affine transform space. Multiple GAEs are stacked to represent a hierarchy of transformations and capture higher-order dependencies. From the experiments on predicting frequency modulated sin-waves, authors stated that standard RNNs were outperformed in terms of accuracy. However, no performance comparison was conducted on videos.

Based on the Spatial Transformer (ST) module [119]: To provide spatial transformation capabilities to existing CNNs, Jaderberg *et al.* [119] proposed the ST module. It regresses different affine transformation parameters for each input, to be applied as a single transformation to the whole feature map(s) or image(s). Moreover, it can be incorporated at any part of the CNNs and it is fully differentiable. The ST module is the essence of vector-based resampling approaches for video prediction. As an extension, Patraucean *et al.* [66] modified the grid generator to consider per-pixel transformations instead of a single dense transformation map for the entire image. They nested a LSTM-based temporal encoder into a spatial AE, proposing the AE-convLSTM-flow architecture. The prediction is generated by resampling the current frame with the flow-based predicted transformation. Using the components of the AE-convLSTM-flow architecture, Lu *et al.* [120] assembled an extrapolation module which is unfolded in time for multi-step prediction. Their Flexible Spatio-semporal Network (FSTN) features a novel loss function using the DeePSiM perceptual loss [45] in order to mitigate blurriness. An exhaustive experimentation and ablation study was carried out, testing multiple combinations of loss functions. Also inspired by the ST module for the volume sampling layer, Liu *et al.* proposed the Deep Voxel Flow (DVF) architecture [22]. It consists of a multi-scale flow-based ED model originally designed for the video frame interpolation task, but also evaluated on a predictive basis reporting sharp results. Liang *et al.* [56] use a flow-warping layer based on a bilinear interpolation. Finn *et al.* proposed the Spatial Transformer Predictor (STP) motion-based model [5] producing 2D affine transformations for bilinear sampling. Pursuing efficiency, Amersfoort *et al.* [121] proposed a CNN designed to predict local affine transformations of overlapping image patches. Unlike the ST module, authors estimated transformations of input frames off-line and at a patch level. As the model is parameter-efficient, it was unfolded in time for multi-step prediction. This resembles RNNs as the parameters are shared over time and the local affine transforms play the role of recurrent states.

4.2.2 Kernel-based Resampling

As a promising alternative to the vector-based resampling, recent approaches synthesize pixels by convolving input patches with a predicted kernel. However, convolutional operations are limited in learning spatial invariant representations of complex transformations. Moreover, due to their local receptive fields, global spatial information is not fully preserved. Using larger kernels would help to preserve global features, but in exchange for a higher memory consumption. Pooling layers are another alternative, but loosing spatial resolution. Preserving spatial resolution at a low computational cost is still an open challenge for future video frame prediction task. Transformation layers used in vector-based resampling [22], [66], [119] enabled CNNs to be spatially invariant and also inspired kernel-based architectures.

Inspired by the Convolutional Dynamic Neural Advection (CDNA) module [5]: In addition to the STP vector-based model, Finn *et al.* [5] proposed two different kernel-based motion prediction modules outperforming previous approaches [44], [122], (1) the Dynamic Neural Advection (DNA) module predicting different distributions for each pixel and (2) the CDNA module that instead of predicting different distributions for each pixel, it predicts multiple discrete distributions that are convolutionally applied to the input. While, CDNA and STP mask out objects that are moving in consistent directions, the DNA module produces per-pixel motion. Similar to the CDNA module, Klein *et al.* proposed the Dynamic Convolutional Layer (DCL) [123] for short-range weather prediction. Likewise, Brabandere *et al.* [124] proposed the Dynamic Filter Networks (DFN) generating sample (for each image) and position-specific (for each pixel) kernels. This enabled sophisticated and local filtering operations in comparison with the ST module, that is limited to global spatial transformations. Different to the CDNA model, the DFN uses a softmax layer to filter values of greater magnitude, thus obtaining sharper predictions. Moreover, temporal correlations are exploited using a parameter-efficient recurrent layer, much simpler than [20], [80]. Exploiting adversarial training, Vondrick *et al.* proposed a conditional Generative Adversarial Network (cGAN)-based model [125] consisting of a discriminator similar to [126], and a CNN generator featuring a transformer module inspired by the CDNA model. Different from the CDNA model, transformations are not applied recurrently on a per-frame basis. To deal with in-the-wild videos and make predictions invariant to camera motion, authors stabilized the input videos. However, no performance comparison with previous works has been conducted. Improving [127], Luc *et al.* [128] proposed the Transformation-based & TrIple Video Discriminator GAN (TrIVD-GAN-FP) featuring a novel recurrent unit that computes the parameters of a transformation used to warp previous hidden states without any supervision. These Transformation-based Spatial Recurrent Units (TSRUs) are generic modules and can replace any traditional recurrent unit in currently existent video prediction approaches.

Object-centric representation: Instead of focusing on the whole input, Chen *et al.* [51] modeled individual motion of local objects, i.e. object-centered representations. Based on the ST module and a pyramid-like sampling [129], authors implemented an attention mechanism for object selection. Moreover, transformation kernels were generated dynamically as in the DFN, to then apply them to the last patch containing an object. Although object-centered predictions is novel, performance drops when dealing with multiple objects and occlusions as the attention module fails to distinguish them correctly.

4.3 Explicit Motion from Content Separation

Drawing inspiration from two-stream architectures for action recognition [130], and unconditioned video generation [131], authors decided to factorize the video into content and motion to process each on a separate pathway. By factoring videos, the prediction is performed on a lower-dimensional temporal dynamics separately from the spatial layout. Although this makes end-to-end training difficult, splitting the prediction task into more tractable problems demonstrated good results.

The Motion-content Network (MCnet) [132] was the first end-to-end model in disentangling scene dynamics from the visual appearance, i.e. motion-content factorization. It proved better generalization capabilities and stable long-term predictions compared to [44], [80]. In a similar fashion, yet working in a higher-level pose space, Denton *et al.* proposed Disentangled-representation Net (DRNET) [133] using a novel adversarial loss —it isolates the scene dynamics from the visual content, considered as the discriminative component— to completely disentangle motion dynamics from content. Outperforming [44], [132], the DRNET demonstrated a clean motion from content separation by reporting plausible long-term predictions on both synthetic and natural videos. To improve prediction variability, Liang *et al.* [56] fused the future-frame and future-flow prediction into a unified architecture with a shared probabilistic motion encoder. Aiming to mitigate the ghosting effect in disoccluded regions, Gae *et al.* [134] proposed a two-staged approach consisting of a separate computation of flow and pixel predictions. As they focused on inpainting occluded regions of the image using flow information, they improved results on disoccluded areas avoiding undesirable artifacts and enhancing sharpness. Wu *et al.* [135] proposed a two-staged architecture that firstly predicts the static background to then, using this information, predict the moving objects in the foreground. Final output is generated through composition and by means of a video inpainting module. Reported predictions are quite accurate, yet performance was not contrasted with the latest video prediction models.

Although previous approaches disentangled motion from content, they have not performed an explicit decomposition of videos into primitive object representations. Addressing this issue, Hsieh *et al.* proposed the Decompositional Disentangled Predictive Autoencoder (DDPAE) [136] that decomposes videos into components featuring low-dimensional temporal dynamics. For instance, on the Moving MNIST dataset, DDPAE first decomposes images into individual digits. After that, each digit is factorized into its visual appearance and spatial location, being the latter easier to predict. Although experiments were performed only on synthetic data, this model is a promising baseline encouraging predictive models to explore visual representation decomposition [137], [138], [139].

4.4 Conditioned on Extra Variables

Conditioning the prediction on extra variables such as vehicle odometry or robot state, among others, would narrow the prediction space. These variables have a direct influence on the dynamics of the scene, providing valuable information that facilitates the prediction task. For instance, the motion captured by a camera placed on the dashboard of an autonomous vehicle is directly influenced by the wheel-steering and acceleration. Without explicitly exploiting this information, we rely blindly on the model's capabilities to correlate the wheel-steering and acceleration with the perceived motion.

Following this paradigm, Oh *et al.* first performed long-term video predictions conditioned by control inputs from Atari games [122]. Although the proposed ED-based models reported very long-term predictions (+100), performance drops when dealing with small objects (e.g. bullets in Space Invaders) and uncertainty. However, ℓ_2 loss leads to accurate and long-term predictions for deterministic synthetic videos, such as those extracted from Atari video games. Built on [122], Chiappa *et al.* [140] proposed alternative architectures and training schemes alongside an in-depth performance analysis for both short and long-term prediction. Similar to [122], Kaiser et al. [10] recently used a convolutional ED to learn a world model of Atari games in a self-supervised fashion. This is part of their model-based RL approach called Simulated Policy Learning (SimPLe) focused on learning to play Atari games. In the experiments, SimPLe outperforms previous model-free algorithms requiring only 1-2 hours of in-game interactions. Similar model-based control from visual inputs performed well in restricted scenarios [141], but was inadequate for unconstrained environments.

Deterministic approaches are unable to deal with natural videos in the absence of control variables. To address this limitation, the models proposed by Finn et al. [5] successfully made predictions on natural images, conditioned on the robot state and robot-object interactions performed in a controlled scenario. These models predict per-pixel transformations conditioned by the previous frame, to finally combine them using a composition mask. They outperformed [44], [122] on both conditioned and unconditioned predictions, however the quality of long-term predictions degrades over time because of the blurriness caused by the MSE loss function. Furthermore, Dosovitskiy *et al.* [142] proposed a sensorimotor control model which enables interaction in complex and dynamic 3d environments. The approach is a RL-based technique, with the difference that instead of building upon a monolithic state and a scalar reward, the authors consider high-dimensional input streams, such as raw visual input, alongside a stream of measurements or player statistics. Although the outputs are future measurements instead of visual predictions, it was proven that using multivariate data benefits decision-making over conventional scalar reward approaches. The synergy between model-based RL [9], [10], [96] and video prediction is well defined as the latter aims to model an accurate representation of high-dimensional environments, while the former uses the learned world models as a context for decision-making.

4.5 In the High-level Feature Space

Despite the vast work on video prediction models, there is still room for improvement in natural video prediction. To deal with the curse of dimensionality, authors reduced the prediction space to higher-level representations, such

as semantic and instance segmentation, and human pose. Since the pixels are categorical, the semantic space greatly simplifies the prediction task, yet unexpected deformations in semantic maps and disocclusions, i.e. initially occluded scene entities become visible, induce uncertainty. However, high-level prediction spaces are more tractable and constitute good intermediate representations. By bypassing the prediction in the raw pixel space, models reported longer-term and more accurate predictions.

4.5.1 Semantic Segmentation

By decomposing the visual scene into semantic entities, such as pedestrians, vehicles and obstacles, the output space is narrowed to high-level scene properties. This intermediate representation represents a more tractable space as pixel values of a semantic map are categorical. In other words, scene dynamics are modeled at the semantic entity level. This has encouraged authors to (1) leverage future prediction to improve parsing results [52] and (2) directly predict segmentation maps into the future [16], [57], [143].

Exploring the scene parsing in future frames, Jin *et al.* proposed the Parsing with prEdictive feAtuRe Learning (PEARL) framework [52] which was the first to explore the potential of a GAN-based predictive model to improve per-pixel segmentation. Specifically, this framework conducts two complementary predictive learning tasks. Firstly, it captures the temporal context from input data by using a single-frame prediction network. Then, these temporal features are embedded into a frame parsing network through a transform layer for generating per-pixel future segmentations. Although the prediction model was not compared with existing approaches, PEARL outperforms the traditional parsing methods by generating temporally consistent segmentations. In a similar fashion, Luc *et al.* [57] extended the msCNN model of [44] to the novel task of predicting semantic segmentations of future frames, using softmax pre-activations instead of raw pixels as input. The use of intermediate features or higher-level data as input is a common practice in the video prediction performed in the high-level feature space. Some authors refer to this type or input data as percepts. Luc *et al.* explored different combinations of loss functions, inputs (using RGB information alongside percepts), and outputs (autoregressive and batch models). Results on short, medium and long-term predictions are sound, however, the models are not end-to-end and they do not capture explicitly the temporal continuity across frames. To address this limitation and extending [52], Jin *et al.* first proposed a model for jointly predicting motion flow and scene parsing [144]. Flow-based representations implicitly draw temporal correlations from the input data, thus producing temporally consistent segmentations. Per-pixel accuracy improved when segmenting small objects, e.g. pedestrians and traffic signs, which are more likely to vanish in long-term predictions. Similarly, except that time dimension is modeled with a LSTMs instead of motion flow estimation, Nabavi *et al.* proposed a simple bidirectional ED-LSTM [145] using segmentation masks as input. Although the literature on knowledge distillation [146], [147] stated that softmax pre-activations carry more information than class labels, this model outperforms [57], [144] on short-term predictions.

Using motion flow estimation alongside LSTM-based temporal modeling, Terwilliger *et al.* [18] proposed a novel method performing a LSTM-based feature-flow aggregation. Authors further simplify the semantic space by disentangling motion from semantic entities [132], achieving low overhead and efficiency. Therefore, they segment the current frame and perform future optical flow prediction, which are finally combined with a novel end-to-end warp layer. An improvement on short-term predictions was reported over previous works [57], [144], yet performing worse on mid-term predictions. Similar to [18], F2MF model [148] predict semantic segmented frames by wrapping past convolutional features into the future using a regressed dense displacement field. To deal with disocclusions, authors complemented the main model with a classical feature-to-feature forecast module similar to [16], [149]. F2MF outperformed previous works on the CityScapes dataset without using structure information [150] or precomputed optical flow [18].

A different approach was proposed by Vora *et al.* [150] which first incorporated structure information to predict future 3D segmented point clouds. Their geometry-based model consists of several derivable sub-modules: (1) the pixel-wise segmentation and depth estimation modules which are jointly used to generate the 3d segmented point cloud of the current RGB frame; and (2) an LSTM-based module trained to predict future camera ego-motion trajectories. The future 3d segmented point clouds are obtained by transforming the previous point clouds with the predicted ego-motion. Their short-term predictions improved the results of [57], however, the use of structure information for longer-term predictions is not clear.

The main disadvantage of two-staged, i.e. not end-to-end, approaches [18], [57], [144], [145], [150] is that their performance is constrained by external supervisory signals, e.g. optical flow [151], segmentation [152] and intermediate features or percepts [153]. Breaking this trend, Chiu *et al.* [149] first solved both the semantic segmentation and forecasting problems in a single end-to-end trainable model by using raw pixels as input. This ED architecture is based on two networks: the student, performing the forecasting task, and the teacher guiding the student using a novel knowledge distillation loss. An in-depth ablation study was performed, validating the performance of the ED architectures as well as the 3D convolution used for capturing the temporal scale instead of a LSTM or ConvLSTM, as in previous works.

Avoiding the flood of deterministic models, Bhattacharyya *et al.* proposed a Bayesian formulation of the ResNet model in a novel architecture to capture model and observation uncertainty [17]. As a main contribution, their dropout-based Bayesian approach leverages synthetic likelihoods [154] to encourage prediction diversity and deal with multi-modal outcomes. Since Cityscapes sequences have been recorded in the frame of reference of a moving vehicle, authors conditioned the predictions on vehicle odometry.

4.5.2 Instance Segmentation

While great strides have been made in predicting future segmentation maps, the authors attempted to make predictions at a semantically richer level, i.e. future prediction of

semantic instances. Predicting future instance-level segmentations is a challenging and weakly unexplored task. This is because instance labels are inconsistent and variable in number across the frames in a video sequence. Since the representation of semantic segmentation prediction models is of fixed-size, they cannot directly address semantics at the instance level.

To overcome this limitation and introducing the novel task of predicting instance segmentations, Luc *et al.* [16] predict fixed-sized feature pyramids, i.e. features at multiple scales, used by the Mask R-CNN [155] network. The combination of dilated convolutions and multi-scale, efficiently preserve high-resolution details improving the results over previous methods [57]. To further improve predictions, Sun *et al.* [156] focused on modeling not only the spatio-temporal correlations between the pyramids, but also the intrinsic relations among the feature layers inside them. That is, enriching the contextual information using the proposed Context Pyramid ConvLSTMs (CP-ConvLSTMs). Although the authors have not shown any long-term predictions nor compared with semantic segmentation models, their approach is the state of the art in the task of predicting instance segmentations.

4.5.3 Other High-level Spaces

Although semantic and instance segmentation spaces were the most used in video prediction, other high-level spaces such as human pose and keypoints, represent a promising avenue.

Human Pose: As the human pose is a low-dimensional and interpretable structure, it represents a cheap supervisory signal for predictive models. This has fostered pose-guided prediction methods, in which pixel-level predictions are conditioned by intermediate representations of human poses. However, most of these methods are limited to videos with human presence.

From a supervised prediction of human poses, Villegas *et al.* [54] regress future frames through analogy making [157]. Although background is not considered in the prediction, authors compared the model against [20], [44] reporting long-term results. To make the model unsupervised on the human pose, Wichers *et al.* [53] adopted different training strategies: end-to-end prediction minimizing the ℓ_2 loss, and through analogy making, constraining the predicted features to be close to the outputs of the future encoder. Different from [54], predictions are made in the feature space. As a probabilistic alternative, Walker *et al.* [55] fused a conditioned Variational Autoencoder (cVAE)-based probabilistic pose predictor with a GAN. While the probabilistic predictor enhances the diversity in the predicted poses, the adversarial network ensures prediction realism. As this model struggles with long-term predictions, Fushishita *et al.* [158] addressed long-term video prediction of multiple outcomes avoiding the error accumulation and vanishing gradients by using a unidimensional CNN trained in an adversarial fashion. To enable multiple predictions, they have used additional inputs ensuring trajectory and behavior variability at a human pose level. To better preserve the visual appearance in the predictions than [54], [132], [159], Tang *et al.* [160] firstly predict human poses using a LSTM-based model to then synthesize pose-conditioned future frames using a combination of different networks: a global GAN modeling the time-invariant background alongside a coarse human pose, a local GAN refining the coarse-predicted human pose, and a 3D-AE to ensure temporal consistency across frames.

Keypoint-based representations: The keypoint coordinate space is a meaningful, tractable and structured representation for prediction, ensuring stable learning. It enforces model's internal representation to contain object-level information. This leads to better results on tasks requiring object-level understanding such as, trajectory prediction, action recognition and reward prediction. As keypoints are a natural representation of dynamic objects, Minderer *et al.* [89] reformulated the prediction task in the keypoint coordinate space. They proposed an AE architecture with a bottleneck consisting of a Variational Recurrent Neural Network (VRNN) that predicts dynamics in the keypoint space. Although this model qualitatively outperforms the Stochastic Video Generation (SVG) [161], Stochastic Adversarial Video Prediction (SAVP) [159] and Encoder Predictor with Visual Analogy (EPVA) [53] models, the quantitative evaluation reported similar results.

4.6 Incorporating Uncertainty

Although high-level representations significantly reduce the prediction space, the underlying distribution still has multiple modes. In other words, different plausible outcomes would be equally probable for the same input. Addressing multimodal distributions is not straightforward for regression and classification approaches, as they regress to the mean and aim to discretize a continuous high-dimensional space, respectively. To deal with the inherent unpredictability of natural videos, some works introduced latent variables into existing deterministic models or directly relied on generative models such as GANs and Variational Autoencoders (VAEs).

Inspired by DVF, Xue *et al.* [186] proposed a cVAE-based [187], [188] multi-scale model featuring a novel cross convolutional layer trained to regress the difference image or Eulerian motion [189]. Background on natural videos is not uniform, however the model implicitly assumes that the difference image would accurately capture the movement in foreground objects. Introducing latent variables into a convolutional AE, Goroshin *et al.* [175] proposed a probabilistic model for learning linearized feature representations to linearly extrapolate the predicted frame in a feature space. Uncertainty is introduced to the loss by using a cosine distance as an explicit curvature penalty. Authors focused on evaluating the linearization properties, yet the model was not contrasted to previous works. Extending [92], [186], Fragkiadaki *et al.* [177] proposed several architectural changes and training schemes to handle marginalization over stochastic variables, such as sampling from the prior and variational inference. Their stochastic ED architecture predicts future optical flow, i.e., dense pixel motion field, used to spatially transform the current frame into the next frame prediction. To introduce uncertainty in predictions, the authors proposed the k-best-sample-loss (MCbest) that draws K outcomes penalizing those similar to the ground-truth.

TABLE 2
Summary of video prediction models (**c**: **c**onvolutional; **r**: **r**ecurrent; **v**: **v**ariational; **ms**: **m**ulti-**s**cale; **st**: **st**acked; **bi**: **bi**directional; **P**: **P**ercepts; **M**: **M**otion; **PL**: **P**erceptual **L**oss; **AL**: **A**dversarial **L**oss; **S**/**R**: using **S**ynthetic/**R**eal datasets; **SS**: **S**emantic **S**egmentation; **D**: **D**epth; **S**: **S**tate; **Po**: **Po**se; **O**: **O**dometry; **IS**: **I**nstance **S**egmentation; **MS**: **M**ulti-**S**tep prediction; **npf**: num. of **p**redicted **f**rames, ⋆ 1-5, ⋆⋆ 5-10, ⋆⋆⋆ 10-100, ⋆⋆⋆⋆ over 100 frames; **ood**: tested on **o**ut-**o**f-**d**omain tasks).

				details				evaluation			
method	**year**	**architecture**	**datasets (train, valid, test)**	**input**	**output**	**MS**	**loss function**	**S/R**	**npf**	**ood**	**code**
Direct Pixel Synthesis											
Ranzato *et al.* [97]	2014	rCNN	[62], [78]	RGB	RGB	✗	CE	R	⋆	✗	✗
Srivastava *et al.* [80]	2015	LSTM-AE	[61], [62], [74], [80]	RGB,P	RGB	✓	CE, ℓ_2	SR	⋆⋆⋆	✓	✓
PGN [50]	2015	LSTM-cED	[77]	RGB	RGB	✗	MSE, AL	S	⋆	✗	✗
Shi *et al.* [20]	2015	cLSTM	[80]	RGB	RGB	✗	CE	S	⋆⋆⋆	✓	✗
BeyondMSE [44]	2016	msCNN	[62], [74]	RGB	RGB	✓	ℓ_1, GDL, AL	R	⋆⋆	✗	✓
PredNet [70]	2017	stLSTMs	[64], [67], [68], [162]	RGB	RGB	✓	ℓ_1,ℓ_2	SR	⋆⋆	✓	✓
ContextVP [113]	2018	MD-LSTM	[62], [64], [67], [68]	RGB	RGB	✓	ℓ_1, GDL	R	⋆⋆	✗	✗
fRNN [110]	2018	cGRU-AE	[59], [62], [80]	RGB	RGB	✓	ℓ_1	SR	⋆⋆⋆	✗	✓
E3d-LSTM [111]	2019	r3D-CNN	[59], [80], [163], [164]	RGB	RGB	✓	ℓ_1, ℓ_2, CE	SR	⋆⋆⋆	✓	✓
Kwon *et al.* [108]	2019	cycleGAN	[62], [67], [68], [165], [166]	RGB	RGB	✓	ℓ_1, LoG, AL	R	⋆⋆⋆	✗	✗
Znet [105]	2019	cLSTM	[59], [80]	RGB	RGB	✓	ℓ_2, BCE, AL	SR	⋆⋆⋆	✗	✗
VPGAN [58]	2019	GAN	[59], [81]	RGB,Z	RGB	✓	ℓ_1, L_{cycle}, AL	R	⋆⋆⋆	✗	✗
Jin *et al.* [106]	2020	cED-GAN	[59], [67], [68], [81]	RGB	RGB	✓	ℓ_2, GDL, AL	R	⋆⋆⋆	✗	✗
Shouno *et al.* [107]	2020	GAN	[67], [68]	RGB	RGB	✓	L_p, AL, PL	R	⋆⋆⋆	✗	✗
CrevNet [112]	2020	3d-cED	[67], [68], [80], [167]	RGB	RGB	✓	MSE	SR	⋆⋆⋆	✓	✓
Using Explicit Transformations											
PGP [118]	2014	st-rGAEs	[77], [79]	RGB	RGB	✓	ℓ_2	SR	⋆	✗	✗
Patraucean *et al.* [66]	2015	LSTM-cAE	[61], [80], [83], [84]	RGB	RGB	✗	ℓ_2, ℓ_δ	SR	⋆	✓	✓
DFN [124]	2016	r-cED	[62], [80]	RGB	RGB	✓	BCE	SR	⋆⋆⋆	✓	✓
Amersfoort *et al.* [121]	2017	CNN	[62], [80]	RGB	RGB	✓	MSE	SR	⋆⋆	✗	✗
FSTN [120]	2017	LSTM-cED	[62], [74], [80], [83], [84]	RGB	RGB	✓	ℓ_2, ℓ_δ, PL	SR	⋆⋆⋆	✗	✗
Vondrick *et al.* [125]	2017	cGAN	[76]	RGB	RGB	✓	CE, AL	R	⋆⋆⋆	✓	✗
Chen *et al.* [51]	2017	rCNN-ED	[62], [80]	RGB	RGB	✓	CE, ℓ_2, GDL, AL	SR	⋆⋆	✗	✗
DVF [22]	2017	ms-cED	[62], [65]	RGB	RGB	✓	ℓ_1, TV	R	⋆	✓	✓
SDC-Net [114]	2018	CNN	[67], [75]	RGB,M	RGB	✓	ℓ_1, PL	SR	⋆⋆	✓	✗
TrIVD-GAN-FP [128]	2020	DVD-GAN	[62], [81], [168]	RGB	RGB	✓	L_{hinge} [56]	R	⋆⋆⋆	✗	✗
Explicit Motion from Content Separation											
MCnet [132]	2017	LSTM-cED	[59], [60], [62], [74]	RGB	RGB	✓	ℓ_p, GDL, AL	R	⋆⋆⋆	✗	✓
Dual-GAN [56]	2017	VAE-GAN	[62], [65], [67], [68]	RGB	RGB	✓	ℓ_1, KL, AL	R	⋆⋆	✗	✗
DRNET [133]	2017	LSTM-ED	[59], [80], [169], [170]	RGB	RGB	✓	ℓ_2, CE, AL	SR	⋆⋆⋆⋆	✓	✓
DPG [134]	2019	cED	[67], [171], [172]	RGB	RGB	✓	ℓ_p, TV, PL, CE	SR	⋆⋆	✗	✗
Conditioned on Extra Variables											
Oh *et al.* [122]	2015	rED	[85]	RGB,A	RGB	✓	ℓ_2	S	⋆⋆⋆⋆	✓	✓
Finn *et al.* [5]	2016	st-cLSTMs	[5], [64]	RGB,A,S	RGB	✓	ℓ_2	R	⋆⋆⋆	✗	✓
In the High-level Feature Space											
Villegas *et al.* [54]	2017	LSTM-cED	[63], [64]	RGB,Po	RGB,Po	✓	ℓ_2, PL, AL [45]	R	⋆⋆⋆⋆	✓	✗
PEARL [52]	2017	cED	[69], [173]	RGB	SS	✗	ℓ_2, AL	R	⋆	✓	✗
S2S [57]	2017	msCNN	[69], [173]	P	SS	✓	ℓ_1, GDL, AL	R	⋆⋆⋆	✗	✓
Walker *et al.* [55]	2017	cVAE	[62], [63]	RGB,Po	RGB	✓	ℓ_2, CE, KL, AL	R	⋆⋆⋆	✓	✗
Jin et al. [144]	2017	cED	[69], [162]	RGB,P	SS,M	✓	ℓ_1, GDL, CE	R	⋆⋆⋆	✓	✗
EPVA [53]	2018	LSTM-ED	[64]	RGB	RGB	✓	ℓ_2, AL	SR	⋆⋆⋆⋆	✓	✓
Nabavi *et al.* [145]	2018	biLSTM-cED	[69]	P	SS	✓	CE	R	⋆⋆	✗	✗
F2F *et al.* [16]	2018	st-msCNN	[69]	P	P,SS,IS	✓	ℓ_2	R	⋆⋆⋆	✓	✓
Vora *et al.* [150]	2018	LSTM	[69]	ego-M	ego-M	✗	ℓ_1	R	⋆	✓	✗
Chiu *et al.* [149]	2019	3D-cED	[69], [71]	RGB	SS	✗	CE, MSE	R	⋆⋆	✗	✗
Bayes-WD-SL [17]	2019	bayesResNet	[69]	SS,O	SS	✓	KL	SR	⋆⋆⋆	✓	✓
Sun *et al.* [156]	2019	st-ms-cLSTM	[69], [86]	P	P,IS	✓	ℓ_2, [155]	R	⋆⋆	✗	✗
Terwilliger *et al.* [18]	2019	M-cLSTM	[69]	RGB,P	SS	✓	CE, ℓ_1	R	⋆⋆⋆	✗	✓
Struct-VRNN [89]	2019	cVRNN	[64], [174]	RGB	RGB	✓	ℓ_2, KL	SR	⋆⋆	✓	✓
F2MF [148]	2020	[18]	[69]	RGB	RGB	✓	ℓ_2	R	⋆⋆	✗	✗
Incorporating Uncertainty											
Goroshin *et al.* [175]	2015	cAE	[169], [176]	RGB	RGB	✗	$\ell_2, penalty$	SR	⋆	✗	✗
Fragkiadaki *et al.* [177]	2017	vED	[64], [178]	RGB	RGB	✗	$KL, MCbest$	R	⋆	✓	✗
EEN [179]	2017	vED	[180], [181], [182]	RGB	RGB	✓	ℓ_1, ℓ_2	SR	⋆⋆	✗	✓
SV2P [38]	2018	CDNA	[5], [64], [81]	RGB	RGB	✓	ℓ_p, KL	SR	⋆⋆⋆	✗	✓
SVG [161]	2018	LSTM-cED	[59], [80], [81]	RGB	RGB	✓	ℓ_2, KL	SR	⋆⋆⋆⋆	✗	✓
Castrejon *et al.* [183]	2019	vRNN	[69], [80], [81]	RGB	RGB	✓	KL	SR	⋆⋆⋆	✗	✗
Hu *et al.* [13]	2020	cED	[69], [71], [184], [185]	RGB	SS,D,M	✓	$CE, \ell_\delta, L_d, L_c, L_p$	R	⋆⋆⋆	✓	✗

Incorporating latent variables into the deterministic CDNA architecture for the first time, Babaeizadeh *et al.* proposed the Stochastic Variational Video Prediction (SV2P) [38] model handling natural videos. Their time-invariant posterior distribution is approximated from the entire input video sequence. Moreover, with the explicit modeling of uncertainty using latent variables, the deterministic CDNA model is outperformed. By combining a standard deterministic architecture (LSTM-ED) with stochastic latent variables, Denton *et al.* proposed the SVG network [161]. Different from SV2P, the prior is sampled from a time-varying posterior distribution, i.e. it is a learned-prior instead of fixed-prior sampled from the same distribution. Most of the VAEs use a fixed Gaussian as a prior, sampling randomly at each time step. Exploiting the temporal dependencies, a learned-prior predicts high variance in uncertain situations, and a low variance when a deterministic prediction suffices. The SVG model is easier to train and reported sharper predictions in contrast to [38]. Built upon SVG, Villegas *et al.* [190] implemented a baseline to perform an in-depth empirical study on the importance of the inductive bias, stochasticity, and model's capacity in the video prediction task. Different from previous approaches, Henaff *et al.* proposed the Error Encoding Network (EEN) [179] that incorporates uncertainty by feeding back the residual error —the difference between the ground truth and the deterministic prediction— encoded as a low-dimensional latent variable. In this way, the model implicitly separates the input video into deterministic and stochastic components.

On the one hand, latent variable-based approaches cover the space of possible outcomes, yet predictions lack of realism. On the other hand, GANs struggle with uncertainty, but predictions are more realistic. Searching for a trade-off between VAEs and GANs, Lee *et al.* [159] proposed the SAVP model. It was the first to combine latent variable models with GANs to improve variability in video predictions, while maintaining realism. Under the assumption that blurry predictions of VAEs are a sign of underfitting, Castrejon *et al.* extended the VRNNs to leverage a hierarchy of latent variables and better approximate data likelihood [183]. Although the backpropagation through a hierarchy of conditioned latents is not straightforward, several techniques alleviated this issue such as, KL beta warm-up, dense connectivity pattern between inputs and latents, and Ladder Variational Autoencoders (LVAEs) [191]. As most of the probabilistic approaches fail in approximating the true distribution of future frames, Pottorff *et al.* [192] reformulated the video prediction task without making any assumption about the data distribution. They proposed the Invertible Linear Embedding (ILE) that enables exact maximum likelihood learning of video sequences, by combining an invertible neural network [193], also known as reversible flows, and a linear time-invariant dynamic system. The ILE handles nonlinear motion in the pixel space and scales better to longer-term predictions compared to adversarial models [44]. Also based on Glow model [193] and sharing goals with [192], VideoFlow [194] approaches exact likelihood maximization using normalized flows through invertible transformations. These flow-based architectures present several advantages such as, exact log-likelihood evaluation and faster sampling than autoregressive models, while still producing high-quality long-term and stochastic predictions.

While previous variational approaches [159], [161] focused on predicting a single frame of low resolution in restricted, predictable or simulated datasets, Hu *et al.* [13] jointly predict full-frame ego-motion, static scene, and object dynamics on complex real-world urban driving. Featuring a novel spatio-temporal module, their five-component architecture learns rich representations that incorporate both local and global spatio-temporal context. The model outperformed existing spatio-temporal architectures, by predicting semantic segmentation, depth and optical flow. However, no performance comparison with [159], [161] has been carried out.

5 Performance Evaluation

This section presents the results of the previously analyzed video prediction models on the most popular datasets on the basis of the metrics described below.

5.1 Metrics and Evaluation Protocols

For a fair evaluation of video prediction systems, multiple aspects of prediction need to be addressed such as whether the predicted sequences look realistic, are plausible and cover all possible outcomes. To the best of our knowledge, there are no evaluation protocols and metrics that evaluate predictions by fulfilling all these aspects simultaneously.

The most widely used evaluation protocols for video prediction rely on image similarity-based metrics such as, Mean-Squared Error (MSE), Structural Similarity Index Measure (SSIM) [195], and Peak Signal to Noise Ratio (PSNR). However, evaluating a prediction according to the mismatch between its visual appearance and the ground truth is not always reliable. In practice, these metrics penalize all predictions that deviate from the ground truth. In other words, they prefer blurry predictions nearly accommodating the exact ground truth than sharper and plausible but imperfect generations [159], [183], [196]. Pixel-wise metrics do not always reflect how accurately a model has captured the dynamic features and their temporal variability in a video. In addition, the precision of a metric is influenced by the loss function used to train the model. For instance, models minimizing the MSE loss function would blindly perform well on the PSNR metric as it is based on MSE. Suffering from similar problems, SSIM measures the similarity between two images, from -1 (very dissimilar) to $+1$ (the same image). As a difference, it measures similarities on image patches instead of performing pixel-wise comparison. These metrics are easily fooled by learning to match the background in predictions. To address this issue, some methods [18], [44], [57], [148] also evaluated predictions only on the dynamic parts of the sequence avoiding the background influence.

As the pixel space is multimodal and high-dimensional, it is challenging to evaluate how accurately a predicted sequence covers the full distribution of possible outcomes. Addressing this issue, some probabilistic approaches [159], [161], [183] assessed prediction coverage by sampling multiple random predictions; to then search for the best match

with the ground truth sequence using common metrics. This is the most widely used evaluation protocol in probabilistic models. Other methods [106], [107], [183] also reported their results using perceptual metrics such as: Learned Perceptual Image Patch Similarity (LPIPS) [196] which is a linear weighted ℓ_2 distance between deep features of images, Fréchet Video Distance (FVD) [197] measuring prediction realism at a distribution level using a 3D CNN to capture the temporal coherence across a video sequence, and DeePSiM [45]. Moreover, Lee *et al.* [159] used the VGG Cosine Similarity metric that performs cosine similarity to the features extracted by VGG network [99] from the predictions.

Among other metrics we have the Inception Score (IS) [198] introduced to deal with GANs mode collapse problem by measuring the diversity of generated samples; measuring sharpness based on difference of gradients [44]; Parzen window [199], yet deficient for high-dimensional images; and the Laplacian of Gaussians (LoG) [200], [201] used in [108]. In the semantic segmentation space, authors used the popular Intersection over Union (IoU) metric. IS was also widely used to report results on different methods [55], [126], [132], [133]. Differently, on the basis of the EPVA model [53] a quantitative evaluation was performed, based on the confidence of an external method trained to identify whether the generated video contains a recognizable person. To support quantitative evaluation, a qualitative assessment based on a visual inspection could be carried out via Amazon Mechanical Turk (AMT) workers.

5.2 Results

In this section we report the quantitative results of the most relevant methods reviewed in the previous sections. To achieve a wide comparison, we limited the quantitative results to the most common metrics and datasets. We have distributed the results in different tables, given the large variation in the evaluation protocols of the video prediction models.

Many authors evaluated their methods on the Moving MNIST synthetic environment. Although it represents a restricted and quasi-deterministic scenario, long-term predictions are still challenging. The black and homogeneous background induce methods to accurately extrapolate black frames and vanish the predicted digits in the long-term horizon. Under this configuration, the CrevNet model demonstrated a leap over the previous state of the art. As the second best, the E3d-LSTM network reported stable errors in both short-term and longer-term predictions showing the advantages of their memory attention mechanism. It also reported the second best results on the KTH dataset, after [106] which achieved the best overall performance and demonstrated quality predictions on natural videos.

Performing short-term predictions on the KTH dataset, the Recurrent Ladder Network (RLN) outperformed MCnet and fRNN by a slight margin. The RLN architecture draws similarities with fRNN, except that while the former uses bridge connections, the latter relies on state sharing that improves memory consumption. On the Moving MNIST and UCF-101 datasets, fRNN outperformed RLN. Other interesting methods to highlight are PredRNN and PredRNN++, both providing close results to E3d-LSTM. State-of-the-art results using different metrics were reported on Caltech Pedestrian by Kwon *et al.* [108], CrevNet [112], and Jin *et al.* [106]. The former, by taking advantage of its retrospective prediction scheme, was also the overall winner on the UCF-101 dataset. The latter outperformed state of the art on all metrics except on LPIPS, as predictions of probabilistic approaches are clearer and realist but less consistent with the ground truth. However [106] is absolute the winner on the BAIR Push dataset.

On the one hand, some approaches have been evaluated on other datasets: SDC-Net [114] outperformed [44], [132] on YouTube8M, TrIVD-GAN-FP outperformed [127],

TABLE 3
Results on M-MNIST (Moving MNIST). Predicting the next y frames from x context frames ($x \rightarrow y$). † results reported by Oliu *et al.* [110], ‡ results reported by Wang *et al.* [111], $*$ results reported by Wang *et al.* [202], ◁ results reported by Wang *et al.* [203]. **MSE** represents per-pixel average MSE (10^{-3}). **MSE**◇ represents per-frame error.

	M-MNIST ($10 \rightarrow 10$)					**M-MNIST** ($10 \rightarrow 30$)	
method	**MSE**↓	**MSE**◇↓	**SSIM**↑	**PSNR**↑	**CE**↓	**MSE**◇↓	**SSIM**↑
BeyondMSE [44]	27.48†	122.6$*$	0.713$*$	15.969†	-	-	-
Srivastava *et al.* [80]	17.37†	118.3$*$	0.690$*$	18.183†	341.2	180.1◁	0.583◁
Shi *et al.* [20]	-	96.5‡	0.713‡	-	367.2$*$	156.2◁	0.597◁
DFN [124]	-	89.0‡	0.726‡	-	285.2	149.5◁	0.601◁
CDNA [5]	-	84.2‡	0.728‡	-	346.6$*$	142.3◁	0.609◁
VLN [204]	-	-	-	-	187.7		
Patraucean *et al.* [66]	43.9	-	-	-	179.8	-	-
MCnet [132]†	42.54	-	-	13.857	-	-	-
RLN [205]†	42.54	-	-	13.857	-	-	-
PredNet [70]†	41.61	-	-	13.968	-	-	-
fRNN [110]	**9.47**	68.4‡	0.819‡	**21.386**	-	-	-
PredRNN [202]	-	56.8	0.867	-	97.0	-	-
VPN [206]	-	64.1‡	0.870‡	-	**87.6**	129.6◁	0.620◁
Znet [105]	-	50.5	0.877	-	-	-	-
PredRNN++ [203]	-	46.5	0.898	-	-	**91.1**	**0.733**
E3d-LSTM [111]	-	41.3	0.910	-	-	-	-
CrevNet [112]	-	**22.3**	**0.949**	-	-	-	-

TABLE 4
Results on KTH dataset. Predicting the next y frames from x context frames ($x \rightarrow y$). † results reported by Oliu *et al.* [110], ‡ results reported by Wang *et al.* [111], $*$ results reported by Zhang *et al.* [105], ◁ results reported by Jin *et al.* [106]. Per-pixel average MSE (10^{-3}). Best results are represented in bold.

	KTH ($10 \rightarrow 10$)		**KTH** ($10 \rightarrow 20$)		**KTH** ($10 \rightarrow 40$)	
method	**MSE**↓	**PSNR**↑	**SSIM**↑	**PSNR**↑	**SSIM**↑	**PSNR**↑
Srivastava *et al.* [80]†	9.95	21.22	-	-	-	-
PredNet [70]†	3.09	28.42	-	-	-	-
BeyondMSE [44]†	1.80	29.34	-	-	-	-
fRNN [110]	1.75	29.299	0.771◁	26.12◁	0.678◁	23.77◁
MCnet [132]	1.65†	30.95†	0.804‡	25.95‡	0.73◁	23.89◁
RLN [205]†	**1.39**	**31.27**	-	-	-	-
Shi *et al.* [20]‡	-	-	0.712	23.58	0.639	22.85
SAVP [159]◁	-	-	0.746	25.38	0.701	23.97
VPN [206]$*$	-	-	0.746	23.76	-	-
DFN [124]‡	-	-	0.794	27.26	0.652	23.01
fRNN [110]‡	-	-	0.771	26.12	0.678	23.77
Znet [105]	-	-	0.817	27.58	-	-
SV2P invariant [38]◁	-	-	0.826	27.56	0.778	25.92
SV2P variant [38]◁	-	-	0.838	27.79	0.789	26.12
PredRNN [202]	-	-	0.839	27.55	0.703‡	24.16‡
VarNet [207]◁	-	-	0.843	28.48	0.739	25.37
SAVP-VAE [159]◁	-	-	0.852	27.77	0.811	26.18
PredRNN++ [203]	-	-	0.865	28.47	0.741‡	25.21‡
MSNET [208]	-	-	0.876	27.08	-	-
E3d-LSTM [111]	-	-	0.879	29.31	0.810	27.24
Jin *et al.* [106]	-	-	**0.893**	**29.85**	**0.851**	**27.56**

TABLE 5
Results on Caltech Pedestrian. Predicting the next y frames from x context frames ($x \rightarrow y$). † reported by Kwon *et al.* [108], ‡ reported by Reda *et al.* [114], $*$ reported by Gao *et al.* [134], ◁ reported by Jin *et al.* [106]. Per-pixel average MSE (10^{-3}). Best results in bold.

	Caltech Pedestrian ($10 \rightarrow 1$)			
method	**MSE↓**	**SSIM↑**	**PSNR↑**	**LPIPS↓**
BeyondMSE [44]‡	3.42	0.847	-	-
MCnet [132]‡	2.50	0.879	-	-
DVF [22]$*$	-	0.897	26.2	5.57◁
Dual-GAN [56]	2.41	0.899	-	-
CtrlGen [94]$*$	-	0.900	26.5	6.38◁
PredNet [70]†	2.42	0.905	27.6	7.47◁
ContextVP [113]	1.94	0.921	28.7	6.03◁
GAN-VGG [107]	-	0.916	-	3.61
G-VGG [107]	-	0.917	-	**3.52**
SDC-Net [114]	1.62	0.918	-	-
Kwon et al. [108]	**1.61**	0.919	29.2	-
DPG [134]	–	0.923	28.2	5.04◁
G-MAE [107]	-	0.923	-	4.30
GAN-MAE [107]	-	0.923	-	4.09
CrevNet [112]	-	0.925	**29.3**	-
Jin *et al.* [106]	-	**0.927**	29.1	5.89

TABLE 6
Results on UCF-101 dataset. Predicting the next x frames from y context frames ($x \rightarrow y$). † results reported by Oliu *et al.* [110]. Per-pixel average MSE (10^{-3}). Best results are represented in bold.

	UCF-101 ($10 \rightarrow 10$)		UCF-101 ($4 \rightarrow 1$)		
method	**MSE↓**	**PSNR↑**	**MSE↓**	**SSIM↑**	**PSNR↑**
Srivastava *et al.* [80]†	148.66	10.02	-	-	-
PredNet [70]†	15.50	19.87	-	-	-
BeyondMSE [44]†	9.26	22.78	-	-	-
MCnet [132]	9.40†	23.46†	-	0.91	31.0
RLN [205]†	9.18	23.56	-	-	-
fRNN [110]	**9.08**	**23.87**	-	-	-
BeyondMSE [44]	-	-	-	0.92	32
Dual-GAN [56]	-	-	-	**0.94**	30.5
DVF [22]	-	-	-	**0.94**	33.4
ContextVP [113]	-	-	-	0.92	34.9
Kwon et al. [108]	-	-	**1.37**	**0.94**	**35.0**

[209] on Kinetics-600 test set [168], E3d-LSTM compared their method with [110], [202], [203], [206] on the TaxiBJ dataset [163], and CrevNet [112] on Traffic4cast [167]. On the other hand, some explored out-of-domain tasks [20], [111], [112], [124], [125] (see ood column in Table 2).

5.2.1 Results on Probabilistic Approaches

Probabilistic video prediction methods have been mainly evaluated on the Stochastic Moving MNIST, Bair Push and Cityscapes datasets. Different from the original Moving MNIST dataset, the stochastic version includes uncertain digit trajectories, i.e. the digits bounce off the border with a random new direction. On this dataset, both versions of Castrejon *et al.* models (1L, without a hierarchy of latents, and 3L with a 3-level hierarchy of latents) outperform SVG by a large margin. On the Bair Push dataset, SAVP reported sharper and more realistic-looking predictions than SVG which suffer of blurriness. However, both models were outperformed by [183] as well on the Cityscapes dataset. The model based on a 3-level hierarchy of latents [183]

TABLE 7
Results on SM-MNIST (Stochastic Moving MNIST), BAIR Push and Cityscapes datasets. † results reported by Castrejon *et al.* [183]. ‡ results reported by Jin *et al.* [106].

	SM-MNIST ($5 \rightarrow 10$)		BAIR Push ($2 \rightarrow 28$)			Cityscapes ($2 \rightarrow 28$)	
method	**FVD↓**	**SSIM↑**	**FVD↓**	**SSIM↑**	**PSNR↑**	**FVD↓**	**SSIM↑**
SVG [161]	90.81†	0.688†	256.62†	0.816†	17.72‡	1300.26†	0.574†
SAVP [159]	-	-	143.43†	0.795†	18.42‡	-	-
SAVP-VAE [159]	-	-	-	0.815‡	19.09‡	-	-
SV2P inv. [38]‡	-	-	-	0.817	20.36	-	-
vRNN 1L [183]	63.81	**0.763**	149.22	0.829	-	682.08	0.609
vRNN 3L [183]	**57.17**	0.760	**143.40**	0.822	-	**567.51**	**0.628**
Jin *et al.* [106]	-	-	-	**0.844**	**21.02**	-	-

TABLE 8
Results on Cityscapes dataset. Predicting the next y semantic segmented frames from 4 context frames ($4 \rightarrow y$). ‡ IoU results on eight moving objects classes. † results reported by Chiu *et al.* [149]

	Cityscapes			
	($4 \rightarrow 1$)	($4 \rightarrow 3$)	($4 \rightarrow 9$)	($4 \rightarrow 10$)
method	**IoU↑**	**IoU↑**	**IoU↑**	**IoU↑**
S2S [57]‡	-	55.3	40.8	-
S2S-maskRCNN [16]‡	-	55.4	42.4	-
S2S [57]	62.6†	59.4	47.8	-
Nabavi *et al.* [145]	71.37	60.06	-	-
F2F [16]	-	61.2	41.2	-
Vora *et al.* [150]	-	61.47	45.4	-
S2S-Res101-FCN [144]	-	62.6	-	50.8
Terwilliger *et al.* [18]‡	-	65.1	46.3	-
Chiu *et al.* [149]	72.43	65.53	50.52	
Jin *et al.* [144]	-	66.1	-	**53.9**
Terwilliger *et al.* [18]	73.2	67.1	51.5	52.5
Bayes-WD-SL [17]	**75.3**	66.7	52.5	-
F2MF [148]‡	–	**67.7**	**54.6**	–
F2MF [148]	–	**69.6**	**57.9**	–

outperform previous works on all three datasets, showing the advantages of the extra expressiveness of this model.

5.2.2 Results on the High-level Prediction Space

Most of the methods have chosen the semantic segmentation space to make predictions. Although they relied on different datasets for training, performance results were mostly reported on the Cityscapes dataset using the IoU metric. Authors explored short-term (next-frame prediction), mid-term (+3 time steps in the future) and long-term (up to +10 time step in the future) predictions. On the semantic segmentation prediction space, Bayes-WD-SL [17], F2MF [148], and Jin *et al.* [52] reported the best results. Among these methods, it is noteworthy that Bayes-WD-SL was the only one to explore prediction diversity on the basis of a Bayesian formulation.

In the instance segmentation space, the F2F pioneering method [16] was outperformed by Sun *et al.* [156] on short and mid-term predictions using the AP50 and AP evaluation metrics. On the other hand, in the keypoint coordinate space, the seminal model of Minderer *et al.* [89] qualitatively outperformed SVG [161], SAVP [159] and EPVA [53], yet pixel-wise metrics reported similar results. In the human pose space, and by regressing future frames from human pose predictions, Tang *et al.* [160] outperformed SAVP [159],

MCnet [132] and [54] on the basis of the PSNR and SSIM metrics on the Penn Action and J-HMDB [210] datasets.

6 Discussion

The video prediction literature ranges from a direct synthesis of future pixel intensities, to complex probabilistic models addressing prediction uncertainty. The range between these approaches consists of methods that try to factorize or narrow the prediction space. Simplifying the prediction task has been a natural evolution of video prediction models, influenced by several open research challenges discussed below. Due to the curse of dimensionality and the inherent pixel variability, developing a robust prediction based on raw pixel intensities is overly-complicated. This often leads to the regression-to-the-mean problem, visually represented as blurriness. Making parametric models larger would improve the quality of predictions, yet this is currently incompatible with high-resolution predictions due to memory constraints. Transformation-based approaches propagate pixels from previous frames based on estimated flow maps. In this case, prediction quality is directly influenced by the accuracy of the estimated flow. Similarly, the prediction in a high-level space is mostly conditioned by the quality of some extra supervisory signals such as semantic maps and human poses, to name a few. Erroneous supervision signals would harm prediction quality.

Analyzing the impact of the inductive bias on the performance of a video prediction model, Villegas *et al.* [190] demonstrated the maximization of the SVG model [161] performance with minimal inductive bias (e.g. segmentation or instance maps, optical flow, adversarial losses, etc.) by increasing progressively the scale of computation. A common assumption when addressing the prediction task in a high-level feature space, is the direct improvement of long-term predictions as a result of simplifying the prediction space. Even if the complexity of the prediction space is reduced, it is still multimodal when dealing with natural videos. For instance, when it comes to long-term predictions in the semantic segmentation space, most of the models reported predictions only up to ten time steps into the future. This directly suggests that the choice of the prediction space is still an unsolved problem. Finding a trade-off between the complexity of the prediction space and the output quality is challenging. An overly-simplified representation could limit the prediction on complex data such as natural videos. Although abstract predictions suffice for many of the decision-making systems based on visual reasoning, prediction in pixel space is still being addressed.

From the analysis performed in this review and in line with the conclusions extracted from [190] we state that: (1) including recurrent connections and stochasticity in a video prediction model generally lead to improved performance; (2) increasing model capacity while maintaining a low inductive bias also improves prediction performance; (3) multi-step predictions conditioned by previously generated outputs are prone to accumulate errors, diverging from the ground truth when addressing long-term horizons; (4) methods predicted further in the future without relying on high-level feature spaces; (5) combining pixel-wise losses with adversarial training somewhat mitigates the regression-to-the-mean issue.

6.1 Research Challenges

Despite the wealth of currently existing video prediction approaches and the significant progress made in this field, there is still room to improve state-of-the-art algorithms. To foster progress, open research challenges must be clearly identified and disentangled. So far in this review, we have already discussed about: (1) the importance of spatio-temporal correlations as a self-supervisory signal for predictive models; (2) how to deal with future uncertainty and model the underlying multimodal distributions of natural videos; (3) the over-complicated task of learning meaningful representations and deal with the curse of dimensionality; (4) pixel-wise loss functions and blurry results when dealing with equally probable outcomes, i.e. probabilistic environments. These issues define the open research challenges in video prediction.

Currently existing methods are limited to short-term horizons. While frames in the immediate future are extrapolated with high accuracy, in the long term horizon the prediction problem becomes multimodal by nature. Initial solutions consisted on conditioning the prediction on previously predicted frames. However, these autoregressive models tend to accumulate prediction errors that progressively diverge the generated prediction from the expected outcome. On the other hand, due to memory issues, there is a lack of resolution in predictions. Authors tried to address this issue by composing the full-resolution image from small predicted patches. However, as the results are not convincing because of the annoying tilling effect, most of the available models are still limited to low-resolution predictions. In addition to the lack of resolution and long-term predictions, models are still prone to the regress-to-the-mean problem that consists on averaging the output frame to accommodate multiple equally probable outcomes. This is directly related to the pixel-wise loss functions, that focus the learning process on the visual appearance. The choice of the loss function is an open research problem with a direct influence on the prediction quality. Finally, the lack of reliable and fair evaluation models makes the qualitative evaluation of video prediction challenging and represents another potential open problem.

6.2 Future Directions

Based on the in-depth analysis conducted in this review, we present some future promising research directions.

Consider alternative loss functions: Pixel-wise loss functions are widely used in the video prediction task, causing blurry predictions when dealing with uncontrolled environments or long-term horizon. In this regard, great efforts have been made in the literature to identify adequate loss functions for the prediction task. However, despite the existing wide spectrum of loss functions, most models still blindly rely on deterministic loss functions.

Alternatives to RNNs: Currently, RNNs are still widely used in this field to model temporal dependencies, and achieved state-of-the-art results on different benchmarks

[110], [111], [202], [203]. Nevertheless, some methods also relied on 3D convolutions to further enhance video prediction [111], [149] representing a promising avenue.

Use synthetically generated videos: Simplifying the prediction is a current trend in the video prediction literature. A vast amount of video prediction models explored higher-level features spaces to reformulate the prediction task into a more tractable problem. However, this mostly conditions the prediction to the accuracy of an external source of supervision such as optical flow, human pose, pre-activations (percepts) extracted from supervised networks, and more. This issue could be alleviated by taking advantage of existing fully-annotated and photorealistic synthetic datasets or by using data generation tools. Video prediction in photorealistic synthetic scenarios has not been explored in the literature.

Evaluation metrics: Since the most widely used evaluation protocols for video prediction rely on image similarity-based metrics, the need for fairer evaluation metrics is imminent. A fair metric should not penalize predictions that deviate from the ground truth at the pixel level, if their content represents a plausible future prediction in a higher level, i.e., the dynamics of the scene correspond to the reality of the labels. In this regard, some methods evaluate the similarity between distributions or at a higher-level. However, there is still room for improvement in the evaluation protocols for video prediction and generation [211].

7 Conclusion

In this review, after reformulating the predictive learning paradigm in the context of video prediction, we have closely reviewed the fundamentals on which it is based: exploiting the time dimension of videos, dealing with stochasticity, and the importance of the loss functions in the learning process. Moreover, an analysis of the backbone deep learning-based architectures for this task was performed in order to provide the reader the necessary background knowledge. The core of this study encompasses the analysis and classification of more than 50 methods and the datasets they have used. Methods were analyzed from three perspectives: method description, contribution over the previous works and performance results. They have also been classified according to a proposed taxonomy based on their main contribution. In addition, we have presented a comparative summary of the datasets and methods in tabular form so as the reader, at a glance, could identify low-level details. In the end, we have discussed the performance results on the most popular datasets and metrics to finally provide useful insight in shape of future research directions and open problems. In conclusion, video prediction is a promising avenue for the self-supervised learning of rich spatio-temporal correlations, providing prediction capabilities to existing intelligent decision-making systems. While great strides have been made, there is still room for improvement in video prediction using deep learning techniques.

Acknowledgments

This work has been funded by the Spanish Government PID2019-104818RB-I00 grant for the MoDeaAS project, supported with Feder funds. This work has also been supported by two Spanish national grants for PhD studies, FPU17/00166, and ACIF/2018/197 respectively. We also acknowledge Zuria Bauer and Victor Villena-Martinez for their valuable discussion and support on this work.

References

[1] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv:1806.11230*, 2018.
[2] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, Sep. 2018.
[3] V. Villena-Martinez, S. Oprea, M. Saval-Calvo, J. A. López, A. F. Guilló, and R. B. Fisher, "When deep learning meets data alignment: A review on deep registration networks (DRNs)," *arXiv:2003.03167*, 2020.
[4] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, 2015.
[5] C. Finn, I. J. Goodfellow, and S. Levine, "Unsupervised Learning for Physical Interaction through Video Prediction," in *NeurIPS*, 2016.
[6] F. Ebert, C. Finn, S. Dasari, A. Xie, A. X. Lee, and S. Levine, "Visual foresight: Model-based deep reinforcement learning for vision-based robotic control," *arxiv:1812.00568*, 2018.
[7] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *TPAMI*, vol. 38, no. 1, pp. 14–29, 2016.
[8] A. Xie, F. Ebert, S. Levine, and C. Finn, "Improvisation through physical understanding: Using novel objects as tools with visual foresight," in *Robotics: Science and Systems*, 2019.
[9] D. Hafner, T. P. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 97, 2019, pp. 2555–2565.
[10] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, A. Mohiuddin, R. Sepassi, G. Tucker, and H. Michalewski, "Model based reinforcement learning for atari," in *ICLR*, 2020.
[11] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-Term On-Board Prediction of People in Traffic Scenes Under Uncertainty," in *CVPR*, 2018.
[12] C. Choi, "Shared cross-modal trajectory prediction for autonomous driving," *arxiv:2004.00202*, 2020.
[13] A. Hu, F. Cotter, N. Mohan, C. Gurau, and A. Kendall, "Probabilistic future prediction for video scene understanding," in *ECCV*, 2020, pp. 767–785.
[14] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Predicting the future: A jointly learnt model for action anticipation," in *ICCV*, 2019, pp. 5561–5570.
[15] C. R. Opazo, B. Fernando, and H. Li, "Action anticipation by predicting future dynamic images," in *ECCV Workshops (3)*, 2018, pp. 89–105.
[16] P. Luc, C. Couprie, Y. LeCun, and J. Verbeek, "Predicting Future Instance Segmentation by Forecasting Convolutional Features," in *ECCV*, 2018, pp. 593–608.
[17] A. Bhattacharyya, M. Fritz, and B. Schiele, "Bayesian prediction of future street scenes using synthetic likelihoods," in *ICLR*, 2019.
[18] A. Terwilliger, G. Brazil, and X. Liu, "Recurrent flow-guided semantic forecasting," in *WACV*, 2019.
[19] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection - A new baseline," in *CVPR*, 2018.
[20] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *NeurIPS*, 2015.
[21] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. WOO, "Deep learning for precipitation nowcasting: A benchmark and a new model," in *NeurIPS*, 2017.
[22] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *ICCV*, 2017.
[23] W. R. Softky, "Unsupervised pixel-prediction," in *NeurIPS*, 1995.
[24] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, no. 1, 1999.

[25] G. Deco and B. Schürmann, "Predictive coding in the visual cortex by a recurrent network with gabor receptive fields," *Neural Processing Letters*, vol. 14, no. 2, 2001.

[26] A. Hollingworth, "Constructing visual representations of natural scenes: the roles of short- and long-term visual memory." *Journal of experimental psychology. Human perception and performance*, vol. 30 3, 2004.

[27] A. Cleeremans and J. L. McClelland, "Learning the structure of event sequences." *Journal of Experimental Psychology: General*, vol. 120, no. 3, 1991.

[28] A. Cleeremans and J. Elman, *Mechanisms of implicit learning: Connectionist models of sequence processing*. MIT press, 1993.

[29] R. Baker, M. Dexter, T. E. Hardwicke, A. Goldstone, and Z. Kourtzi, "Learning to predict: Exposure to temporal sequences facilitates prediction of future events," *Vision Research*, vol. 99, 2014.

[30] H. E. M. den Ouden, P. Kok, and F. P. de Lange, "How prediction errors shape perception, attention, and motivation," in *Front. Psychology*, 2012.

[31] Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *TPAMI*, vol. 35, no. 8, 2013.

[32] X. Wang and A. Gupta, "Unsupervised Learning of Visual Representations Using Videos," in *ICCV*, 2015.

[33] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in *ICCV*, 2015.

[34] D.-A. Huang, V. Ramanathan, D. Mahajan, L. Torresani, M. Paluri, L. Fei-Fei, and J. Carlos Niebles, "What makes a video a video: Analyzing temporal information in video understanding models and datasets," in *CVPR*, June 2018.

[35] L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Schölkopf, and W. T. Freeman, "Seeing the arrow of time," in *CVPR*, 2014.

[36] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification," in *ECCV*, 2016.

[37] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman, "Learning and using the arrow of time," in *CVPR*, 2018.

[38] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic variational video prediction," in *ICLR*, 2018.

[39] S. Aigner and M. Körner, "The importance of loss functions for increasing the generalization abilities of a deep learning-based next frame prediction model for traffic scenes," *Machine Learning and Knowledge Extraction*, vol. 2, no. 2, pp. 78–98, 2020.

[40] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *Trans. on Computational Imaging*, vol. 3, no. 1, 2017.

[41] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *arXiv:1702.05659*, 2017.

[42] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *CVPR*, 2017.

[43] J.-J. Hwang, T.-W. Ke, J. Shi, and S. X. Yu, "Adversarial structure matching for structured prediction tasks," in *CVPR*, 2019.

[44] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *ICLR (Poster)*, 2016.

[45] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *NIPS*, 2016.

[46] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[47] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.

[48] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *ICCV*, 2017.

[49] J. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 9909, 2016.

[50] W. Lotter, G. Kreiman, and D. D. Cox, "Unsupervised learning of visual structure using predictive generative networks," *arXiv:1511.06380*, 2015.

[51] X. Chen, W. Wang, J. Wang, and W. Li, "Learning object-centric transformation for video prediction," in *ACM-MM*, ser. MM '17. New York, NY, USA: ACM, 2017.

[52] X. Jin, X. Li, H. Xiao, X. Shen, Z. Lin, J. Yang, Y. Chen, J. Dong, L. Liu, Z. Jie, J. Feng, and S. Yan, "Video Scene Parsing with Predictive Feature Learning," in *ICCV*, 2017.

[53] N. Wichers, R. Villegas, D. Erhan, and H. Lee, "Hierarchical long-term video prediction without supervision," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 80, 2018.

[54] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, "Learning to generate long-term future via hierarchical prediction," in *ICML*, 2017.

[55] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *ICCV*, 2017.

[56] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion GAN for future-flow embedded video prediction," in *ICCV*, 2017.

[57] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun, "Predicting Deeper into the Future of Semantic Segmentation," in *ICCV*, 2017.

[58] Z. Hu and J. Wang, "A novel adversarial inference framework for video prediction with action control," in *ICCV Workshops*, 2019.

[59] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *ICPR*, 2004.

[60] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *TPAMI*, vol. 29, no. 12, 2007.

[61] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *ICCV*, 2011.

[62] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv:1212.0402*, 2012.

[63] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *ICCV*, 2013.

[64] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *TPAMI*, vol. 36, no. 7, 2014.

[65] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The THUMOS challenge on action recognition for videos "in the wild"," *CVIU*, vol. 155, 2017.

[66] V. Patraucean, A. Handa, and R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory," *(ICLR) Workshop*, 2015.

[67] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *CVPR*, 2009.

[68] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, vol. 32, no. 11, 2013.

[69] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[70] W. Lotter, G. Kreiman, and D. Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning," in *ICLR (Poster)*, 2017.

[71] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," *arXiv: 1803.06184*, 2018.

[72] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *arXiv:1903.11027*, 2019.

[73] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020.

[74] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[75] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv:1609.08675*, 2016.

[76] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li, "YFCC100M: the new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, 2016.

[77] I. Sutskever, G. E. Hinton, and G. W. Taylor, "The recurrent temporal restricted boltzmann machine," in *NIPS*, 2008.

[78] C. F. Cadieu and B. A. Olshausen, "Learning intermediate-level representations of form and motion from natural movies," *Neural Computation*, vol. 24, no. 4, 2012.

[79] R. Memisevic and G. Exarchakis, "Learning invariant features by harnessing the aperture problem," in *ICML*, vol. 28, 2013.

[80] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised Learning of Video Representations using LSTMs," in *ICML*, 2015.

[81] F. Ebert, C. Finn, A. X. Lee, and S. Levine, "Self-supervised visual planning with temporal skip connections," in *CoRL*, ser. Proceedings of Machine Learning Research, vol. 78, 2017.

[82] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," *arXiv:1910.11215*, 2019.

[83] R. Vezzani and R. Cucchiara, "Video surveillance online repository (visor): an integrated framework," *Multimedia Tools Appl.*, vol. 50, no. 2, 2010.

[84] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, "PROST: parallel robust online simple tracking," in *CVPR*, 2010.

[85] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *J. Artif. Intell. Res.*, vol. 47, 2013.

[86] G. Seguin, P. Bojanowski, R. Lajugie, and I. Laptev, "Instance-level video segmentation from object tracks," in *CVPR*, 2016.

[87] A. Garcia-Garcia, P. Martinez-Gonzalez, S. Oprea, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Jover-Alvarez, "The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions," in *IROS*, 2018, pp. 6790–6797.

[88] Z. Bauer, F. Gomez-Donoso, E. Cruz, S. Orts-Escolano, and M. Cazorla, "UASOL, a large-scale high-resolution outdoor stereo dataset," *Scientific Data*, vol. 6, no. 1, 2019.

[89] M. Minderer, C. Sun, R. Villegas, F. Cole, K. P. Murphy, and H. Lee, "Unsupervised learning of object structure and dynamics from videos," in *NeurIPS*, 2019.

[90] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating Visual Representations from Unlabeled Video," in *CVPR*, 2016.

[91] D. Jayaraman and K. Grauman, "Look-ahead before you leap: End-to-end active recognition by forecasting the effect of motion," in *ECCV*, vol. 9909, 2016.

[92] J. Walker, C. Doersch, A. Gupta, and M. Hebert, "An Uncertain Future: Forecasting from Static Images Using Variational Autoencoders," in *ECCV*, 2016.

[93] B. Chen, W. Wang, and J. Wang, "Video imagination from a single image with transformation generation," in *ACM Multimedia*, 2017.

[94] Z. Hao, X. Huang, and S. J. Belongie, "Controllable video generation with sparse trajectories," in *CVPR*, 2018.

[95] Y. Ye, M. Singh, A. Gupta, and S. Tulsiani, "Compositional video prediction," in *ICCV*, October 2019.

[96] D. Hafner, T. P. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *ICLR*, 2020.

[97] 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:1412.6604*, 2014.

[98] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH*, 2010.

[99] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[100] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV*, T. Pajdla and J. Matas, Eds., vol. 3024, 2004.

[101] Y. Zhou and T. L. Berg, "Learning Temporal Transformations from Time-Lapse Videos," in *ECCV*, 2016.

[102] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *ICLR*, 2018.

[103] S. Aigner and M. Körner, "Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing autoencoder gans," *arXiv:1810.01325*, 2018.

[104] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *NIPS*, 2017.

[105] J. Zhang, Y. Wang, M. Long, W. Jianmin, and P. S. Yu, "Z-order recurrent neural networks for video prediction," in *ICME*, 2019.

[106] B. Jin, Y. Hu, Q. Tang, J. Niu, Z. Shi, Y. Han, and X. Li, "Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction," *arXiv:2002.09905*, 2020.

[107] O. Shouno, "Photo-realistic video prediction on natural videos of largely changing frames," *arXiv:2003.08635*, 2020.

[108] Y.-H. Kwon and M.-G. Park, "Predicting future frames using retrospective cycle gan," in *CVPR*, 2019.

[109] R. Hou, H. Chang, B. Ma, and X. Chen, "Video prediction with bidirectional constraint network," in *FG*, 2019.

[110] M. Oliu, J. Selva, and S. Escalera, "Folded recurrent neural networks for future video prediction," in *ECCV*, 2018.

[111] 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 *ICLR*, 2019.

[112] W. Yu, Y. Lu, S. Easterbrook, and S. Fidler, "Efficient and information-preserving future frame prediction and beyond," in *ICLR*, 2020.

[113] W. Byeon, Q. Wang, R. K. Srivastava, and P. Koumoutsakos, "Contextvp: Fully context-aware video prediction," in *CVPR (Workshops)*, 2018.

[114] 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 *ECCV*, 2018.

[115] R. Memisevic and G. E. Hinton, "Learning to represent spatial transformations with factored higher-order boltzmann machines," *Neural Computation*, vol. 22, no. 6, 2010.

[116] R. Memisevic, "Gradient-based learning of higher-order image features," in *ICCV*, 2011.

[117] ——, "Learning to relate images," *TPAMI*, vol. 35, no. 8, 2013.

[118] V. Michalski, R. Memisevic, and K. Konda, "Modeling deep temporal dependencies with recurrent grammar cells," in *NeurIPS*, 2014.

[119] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial Transformer Networks," in *NeurIPS*, 2015.

[120] C. Lu, M. Hirsch, and B. Schölkopf, "Flexible Spatio-Temporal Networks for Video Prediction," in *CVPR*, 2017.

[121] J. R. van Amersfoort, A. Kannan, M. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-based models of video sequences," *arXiv:1701.08435*, 2017.

[122] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. P. Singh, "Action-Conditional Video Prediction using Deep Networks in Atari Games," in *NeurIPS*, 2015.

[123] B. Klein, L. Wolf, and Y. Afek, "A dynamic convolutional layer for short rangeweather prediction," in *CVPR*, 2015.

[124] B. D. Brabandere, X. Jia, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *NeurIPS*, 2016.

[125] C. Vondrick and A. Torralba, "Generating the Future with Adversarial Transformers," in *CVPR*, 2017.

[126] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating Videos with Scene Dynamics," in *NeurIPS*, 2016.

[127] A. Clark, J. Donahue, and K. Simonyan, "Adversarial video generation on complex datasets," 2019.

[128] P. Luc, A. Clark, S. Dieleman, D. de Las Casas, Y. Doron, A. Cassirer, and K. Simonyan, "Transformation-based adversarial video prediction on large-scale data," *arXiv:2003.04035*, 2020.

[129] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *TPAMI*, vol. 37, no. 9, 2015.

[130] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014.

[131] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *CVPR*, June 2018.

[132] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *ICLR*, 2017.

[133] E. L. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," in *NeurIPS*, 2017.

[134] H. Gao, H. Xu, Q. Cai, R. Wang, F. Yu, and T. Darrell, "Disentangling propagation and generation for video prediction," in *ICCV*, 2019.

[135] Y. Wu, R. Gao, J. Park, and Q. Chen, "Future video synthesis with object motion prediction," 2020.

[136] J. Hsieh, B. Liu, D. Huang, F. Li, and J. C. Niebles, "Learning to decompose and disentangle representations for video prediction," in *NeurIPS*, 2018.

[137] K. Greff, S. van Steenkiste, and J. Schmidhuber, "Neural expectation maximization," in *NIPS*, 2017, pp. 6691–6701.

[138] S. van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber, "Relational neural expectation maximization: Unsupervised discovery of objects and their interactions," in *ICLR (Poster)*, 2018.

[139] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem, "Are disentangled representations helpful for abstract visual reasoning?" in *NeurIPS*, 2019, pp. 14 222–14 235.

[140] S. Chiappa, S. Racanière, D. Wierstra, and S. Mohamed, "Recurrent environment simulators," in *ICLR*, 2017.

[141] K. Fragkiadaki, P. Agrawal, S. Levine, and J. Malik, "Learning visual predictive models of physics for playing billiards," in *ICLR (Poster)*, 2016.

[142] A. Dosovitskiy and V. Koltun, "Learning to Act by Predicting the Future," in *ICLR*, 2017.

[143] P. Luc, "Self-supervised learning of predictive segmentation models from video," Theses, Université Grenoble Alpes, Jun. 2019. [Online]. Available: https://tel.archives-ouvertes.fr/tel-02196890

[144] X. Jin, H. Xiao, X. Shen, J. Yang, Z. Lin, Y. Chen, Z. Jie, J. Feng, and S. Yan, "Predicting Scene Parsing and Motion Dynamics in the Future," in *NeurIPS*, 2017.

[145] S. shahabeddin Nabavi, M. Rochan, and Y. Wang, "Future Semantic Segmentation with Convolutional LSTM," in *BMVC*, 2018.

[146] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *NIPS*, 2014.

[147] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015.

[148] J. Saric, M. Orsic, T. Antunovic, S. Vrazic, and S. Segvic, "Warp to the future: Joint forecasting of features and feature motion," in *CVPR*, 2020, pp. 10 645–10 654.

[149] H.-k. Chiu, E. Adeli, and J. C. Niebles, "Segmenting the future," *arXiv:1904.10666*, 2019.

[150] S. Vora, R. Mahjourian, S. Pirk, and A. Angelova, "Future segmentation using 3d structure," *arXiv:1811.11358*, 2018.

[151] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-preserving interpolation of correspondences for optical flow," in *CVPR*, 2015.

[152] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.

[153] F. Yu, V. Koltun, and T. A. Funkhouser, "Dilated residual networks," in *CVPR*, 2017.

[154] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, "Variational approaches for auto-encoding generative adversarial networks," *arXiv:1706.04987*, 2017.

[155] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *ICCV*, 2017.

[156] J. Sun, J. Xie, J. Hu, Z. Lin, J. Lai, W. Zeng, and W. Zheng, "Predicting future instance segmentation with contextual pyramid convLSTMs," in *ACM Multimedia*. ACM, 2019.

[157] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee, "Deep visual analogy-making," in *NIPS*, 2015.

[158] N. Fushishita, A. Tejero-de-Pablos, Y. Mukuta, and T. Harada, "Long-term video generation of multiple futures using human poses," *arXiv:1904.07538*, 2019.

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

[160] J. Tang, H. Hu, Q. Zhou, H. Shan, C. Tian, and T. Q. S. Quek, "Pose guided global and local gan for appearance preserving human video prediction," in *ICIP*, Sep. 2019.

[161] E. Denton and R. Fergus, "Stochastic video generation with a learned prior," in *ICML*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80, 2018.

[162] E. Santana and G. Hotz, "Learning a driving simulator," *arXiv:1608.01230*, 2016.

[163] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *AAAI*, 2017.

[164] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *ICCV*, 2017.

[165] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *ICCV*, 2017.

[166] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. S. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *ICIP*, 2017.

[167] "Traffic4cast: Traffic map movie forecasting," https://www.iarai.ac.at/traffic4cast/, accessed: 2020-04-14.

[168] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv:1808.01340*, 2018.

[169] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *CVPR*, 2004.

[170] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. A. Funkhouser, "Semantic scene completion from a single depth image," in *CVPR*, 2017.

[171] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *CVPR*, 2015.

[172] J. Janai, F. Güney, A. Ranjan, M. J. Black, and A. Geiger, "Unsupervised learning of multi-frame optical flow with occlusions," in *ECCV*, vol. 11220, 2018.

[173] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV*, vol. 5302, 2008.

[174] E. Zhan, S. Zheng, Y. Yue, L. Sha, and P. Lucey, "Generating multi-agent trajectories using programmatic weak supervision," in *ICLR*, 2019.

[175] R. Goroshin, M. Mathieu, and Y. LeCun, "Learning to linearize under uncertainty," in *NeurIPS*, 2015.

[176] R. Goroshin, J. Bruna, J. Tompson, D. Eigen, and Y. LeCun, "Unsupervised learning of spatiotemporally coherent metrics," in *ICCV*, 2015.

[177] K. Fragkiadaki, J. Huang, A. Alemi, S. Vijayanarasimhan, S. Ricco, and R. Sukthankar, "Motion prediction under multimodality with conditional stochastic networks," *arXiv:1705.02082*, 2017.

[178] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *ECCV*, vol. 6315, 2010.

[179] M. Henaff, J. J. Zhao, and Y. LeCun, "Prediction under uncertainty with error-encoding networks," *arXiv:1711.04994*, 2017.

[180] P. Agrawal, A. Nair, P. Abbeel, J. Malik, and S. Levine, "Learning to poke by poking: Experiential learning of intuitive physics," in *NeurIPS*, 2016, p. 5092–5100.

[181] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *ICML*, vol. 48, 2016.

[182] J. Zhang and K. Cho, "Query-efficient imitation learning for end-to-end simulated driving," in *AAAI*, 2017.

[183] L. Castrejon, N. Ballas, and A. Courville, "Improved conditional vrnns for video prediction," in *ICCV*, 2019.

[184] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving video database with scalable annotation tooling," *arXiv:1805.04687*, 2018.

[185] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017.

[186] T. Xue, J. Wu, K. L. Bouman, and B. Freeman, "Visual Dynamics: Probabilistic Future Frame Synthesis via Cross Convolutional Networks," in *NeurIPS*, 2016.

[187] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[188] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *ECCV*, 2016.

[189] H. Wu, M. Rubinstein, E. Shih, J. V. Guttag, F. Durand, and W. T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ToG*, vol. 31, no. 4, 2012.

[190] R. Villegas, A. Pathak, H. Kannan, D. Erhan, Q. V. Le, and H. Lee, "High fidelity video prediction with large stochastic recurrent neural networks," in *NeurIPS*, 2019, pp. 81–91.

[191] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *NIPS*, 2016.

[192] R. Pottorff, J. Nielsen, and D. Wingate, "Video extrapolation with an invertible linear embedding," *arXiv:1903.00133*, 2019.

[193] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *NeurIPS*, 2018.

[194] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma, "Videoflow: A conditional flow-based model for stochastic video generation," in *ICLR*, 2020.

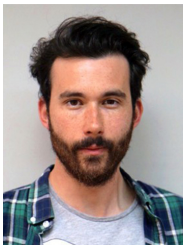
[195] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Trans. on Image Processing*, vol. 13, no. 4, 2004.

[196] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[197] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv:1812.01717*, 2018.
[198] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NIPS*, 2016.
[199] O. Breuleux, Y. Bengio, and P. Vincent, "Quickly generating representative samples from an rbm-derived process," *Neural Computation*, vol. 23, no. 8, 2011.
[200] E. Hildreth, "Theory of edge detection," *Proc. of Royal Society of London*, vol. 207, no. 187-217, 1980.
[201] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," in *NeurIPS*, 2015.
[202] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," in *NeurIPS*, 2017.
[203] Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu, "Predrnn++: Towards A resolution of the deep-in-time dilemma in spatiotemporal predictive learning," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 80, 2018.
[204] F. Cricri, X. Ni, M. Honkala, E. Aksu, and M. Gabbouj, "Video ladder networks," *arXiv:1612.01756*, 2016.
[205] I. Prémont-Schwarz, A. Ilin, T. Hao, A. Rasmus, R. Boney, and H. Valpola, "Recurrent ladder networks," in *NIPS*, 2017.
[206] N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu, "Video pixel networks," in *ICML*, 2017, pp. 1771–1779.
[207] B. Jin, Y. Hu, Y. Zeng, Q. Tang, S. Liu, and J. Ye, "Varnet: Exploring variations for unsupervised video prediction," in *IROS*, 2018.
[208] J. Lee, J. Lee, S. Lee, and S. Yoon, "Mutual suppression network for video prediction using disentangled features," *arXiv:1804.04810*, 2018.
[209] D. Weissenborn, O. Täckström, and J. Uszkoreit, "Scaling autoregressive video models," in *ICLR*, 2020.
[210] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *ICCV*, 2013.
[211] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *ICLR*, 2016.

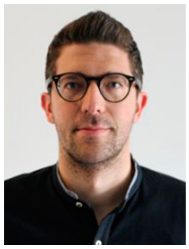
Sergiu Oprea is a PhD student at the Department of Computer Technology (DTIC), University of Alicante. He received his MSc (Automation and Robotics) and BSc (Computer Science) from the same institution in 2017 and 2015 respectively. His main research interests include video prediction with deep learning, virtual reality, 3D computer vision, and parallel computing on GPUs.

Pablo Martinez Gonzalez is a PhD student at the Department of Computer Technology (DTIC), University of Alicante. He received his MSc (Computer Graphics, Games and Virtual Reality) and BSc (Computer Science) at the Rey Juan Carlos University and University of Alicante, in 2017 and 2015, respectively. His main research interests include deep learning, virtual reality and parallel computing on GPUs.

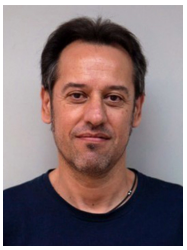


Alberto Garcia Garcia is a Postdoctoral Researcher at the Institute of Space Sciences (ICE-CSIC, Barcelona) working on the MAGNESIA ERC Consolidator project. He received his PhD (Machine Learning and Computer Vision) from the University of Alicante in 2019. Previously he was an intern at NVIDIA Research/Engineering, Facebook Reality Labs, and Oculus Core Tech. His main research interests include deep learning, virtual reality, 3D computer vision, and parallel computing on GPUs.

John Alejandro Castro Vargas is a PhD student at the Department of Computer Technology (DTIC), University of Alicante. He received his MSc (Automation and Robotics) and BSc (Computer Science) from the same institution in 2017 and 2016 respectively. His main research interests include human behavior recognition with deep learning, virtual reality and parallel computing on GPUs.



Sergio Orts-Escolano received a PhD in Computer Science from the University of Alicante in 2014. His research interests include computer vision, assistive robotics, 3D sensors, GPU computing, virtual/augmented reality and deep learning. He has authored +50 publications in top journals and conferences like CVPR, SIGGRAPH, 3DV, BMVC, IROS, etcetera. He has experience as a professor in academia and industry, working as a research scientist for companies such as Google and Microsoft Research.



Jose Garcia-Rodriguez received his Ph.D. degree, with specialization in Computer Vision and Neural Networks, from the University of Alicante (Spain). He is currently Full Professor at the Department of Computer Technology of the University of Alicante. His research areas of interest include: computer vision, machine learning, pattern recognition, robotics, man-machine interfaces, ambient intelligence, and parallel and multicore architectures.

Antonis Argyros is a professor of computer science at the Computer Science Department, University of Crete and a researcher at the Institute of Computer Science, FORTH, in Heraklion, Crete, Greece. His research interests fall in the areas of computer vision and pattern recognition, with emphasis on the analysis of humans in images and videos, human pose analysis, recognition of human activities and gestures, 3D computer vision, as well as image motion and tracking. He is also interested in applications of computer vision in the fields of robotics and smart environments.