ERC Starting Grant 2020

Part B2: The Scientific Proposal

**Section A: State-of-the-art and objectives**

(14 pages excluding references + 2731 chars online budget (1 page) = 15/15 pages)

**Scientific context**

Visual artificial intelligence automatically interprets what happens in visual data like videos. Today’s research strives to tackle queries like: *“Is this person playing basketball?”* (action recognition) [G1, G4, 21, 22, 32]; *“Find the location of the brain stroke”* (video object segmentation) [48]; or *“Track the glacier fractures in satellite footage”* (visual object tracking) [G10, G15, 42-44]. All these queries are about predictions on visual observations that have already taken place. In other words, **today’s algorithms** focus on **explaining past** visual observations. Naturally, not all queries are about the past: *“Will this person draw something in or out of their pocket?”; “Where will the tumour be in the next 1,3, or 5 seconds given breathing patterns and moving organs?”*; or *“How will the glacier fracture given the current motion and melting patterns?”.* Forthese queries and all others,the **next generation of visual algorithms** must be able to **expect** what will happen next in the **future** given past visual observations. Visual artificial intelligence must also be able to prevent before the fact, rather than explain only after the fact. I propose an ambitious 5-year project to design **algorithms that learn to expect the possible futures** from visual sequences.

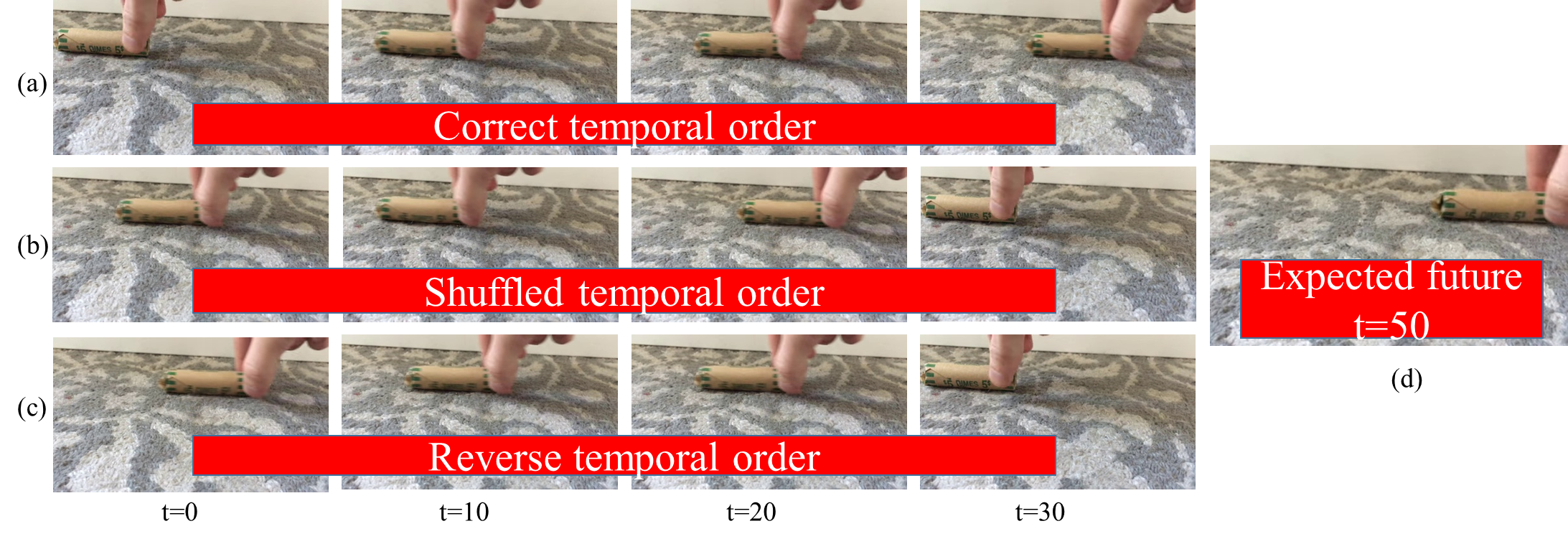
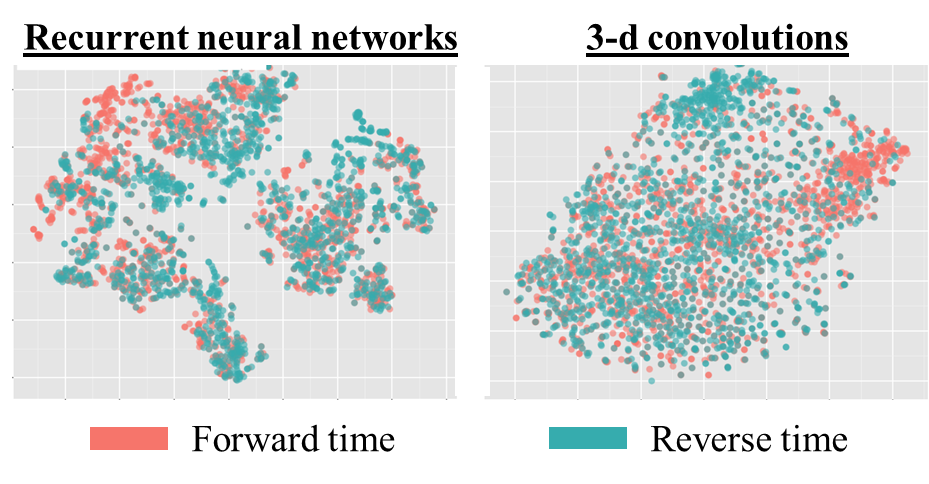
****Learning to expect the possible futures in visual sequences requires understanding of the sequence of events, *i.e.*, learning the **temporality** underpinning visual sequences. Consider the toy sequence in Figure 1 from the popular Something-Something [28] dataset, showing a paper cylinder moving. In Figure 1a the video frames are in the correct temporal order, in Figure 1b the frames are shuffled. From the shuffled video frames, it is impossible to predict the expected future location of the cylinder two seconds ahead of time. From the correct temporal order, however, it becomes immediately clear that in two seconds the cylinder is expected in the right side of the image (Figure 1d). While in this toy example the importance of temporality is perhaps superficial[[1]](#footnote-1), temporality will be critical in tomorrow’s tasks. It will be critical for determining whether a suspect is expected to draw something in or out of their pocket in surveillance videos. It will be critical for determining where the tumour is expected to move in the next few seconds in MRI sequences. And, it will be critical to determine the rate at which a glacier is expected to melt and fracture in the near future. To be able to expect the possible futures, it is imperative to have visual algorithms that model time and learn the temporality in visual sequences.

Figure 1. Understanding time is critical for algorithms to expect the future given past observations. For instance, when moving a paper cylinder, only with the correct temporal order it is possible to predict the expected future in (d), while (b)shuffled or (c) reverse create confusion.

**Scientific challenges**

Most of today’s algorithms that interpret visual sequences are **time-deterministic**. However, time and the expected future is anything but. In learning the future, multiple sources of uncertainty and stochasticity exist. A first source of stochasticity is that there exist many plausible futures given a visual sequence; eventually only one of them is observed. *Given a person going to the kitchen, will she make coffee, or tea or prepare dinner?* A second source of stochasticity is due to the uncertainty in the spatiotemporal appearances of the observed visual sequences. *In the given visual sequence, are these two persons fighting or they are having a conversation?* A third source of stochasticity relates to the so called “jumpy imagination” principle [49], namely when the various futures are expected to happen. When making coffee or tea, for instance, the immediate future steps (boiling water, stirring, and so on) will happen within seconds. When preparing dinner, however, the immediate future steps (open the fridge, wash vegetables, and so on) take place at different and non-uniformly spread timestamps. Today’s visual recognition algorithms do not incorporate stochasticity and uncertainty. They operate on restricted setups, where visual sequences are available in their entirety in advance. Each sequence example is always annotated with respect to a predefined set of classes. And, their predictions are in fixed and pre-determined time spans. Indeed, if visual artificial intelligence is to expect the future, **the first fundamental challenge** is to design **stochastic spatiotemporal** **visual algorithms that** **can handle uncertainty in** visual sequences. As a promising direction, I propose time-stochastic visualalgorithms with temporally invertible generative models**.**

In the recent years, the video understanding literature [G1, G4, G14, 21, 31-33, 35-39] has been exploring ways of breaking with the manual supervision paradigm in visual sequences. Instead, the capabilities of temporal self-supervision for training visual algorithms have been tested. Despite the larger training sizes available with temporal self-supervision, to date self-supervised algorithms have not been able to surpass manual annotation. Recently, [36] hypothesize and show that temporal self-supervision is more fruitful when having algorithms with an extended temporal footprint. This makes sense, as the goal of video understanding is different (quantify appearance and temporal nuances) than the goal of purely image understanding (focus only on appearance, ignore all nuances including temporal that are irrelevant to object semantics). Unfortunately, most of today’s visual algorithms do not model time convincingly. The fundamental problem is that they view time as an extrinsic variable, to be added either as an extra input or output dimension in their spatiotemporal neural networks. That is, today’s video algorithms are **time-extrinsic** by design. They rely on two components: *(i)* convolutions (2-d [20, 34], or 3-d [21, 33]), followed by *(ii)* a temporal aggregation mechanism [G4, G14, 21, 33]. Both components sideline time either on purpose or by consequence. 3-d convolutions can at best only encode short-range motions [21] that discriminate between coarse movements (*e.g.*, *“eating an apple”* vs. *“throwing an apple”*)but struggle with fine-grained movements (*e.g.*, *“apple falling down from the tree to the ground”* vs. *“apple moving from the ground up to the tree”*). Furthermore, the most popular temporal aggregation is taking a statistical average or max over time, both **ignoring temporality** by design. A close second for a temporal aggregation is recurrent nets [G13, G16, 22]. Many have shown [G13, 22] that recurrent networks ignore temporality in practice, as a consequence of chaotic optimization [50]. Be it 3-d convolutions or recurrent networks, today’s algorithms have trouble even telling apart whether a video is played in forward (like Figure 1a) or in the reverse (like in Figure 1c) as shown in [G13], see Figure 2. Indeed, to escape from manual supervision and move towards temporal self-supervision, **the second fundamental challenge** is to design visual recognition algorithms **that encode time-intrinsically** in their latent representations**.** As a promising direction,I propose time-intrinsic visual algorithmsby aligning their neural network structure with time**.**

Figure 2. Today’s visual recognition algorithms have trouble modelling time in visual sequences; note the overlap even in forward vs reverse time representations in videos.

A core reason why encoding time in visual sequences is hard is the conflict between what needs to be learnt and what can be learnt by algorithms. On one hand, **any visual recognition algorithm** –be it a recurrent neural network or 3-d convolutions– **is by nature finite**; finite number of parameters, finite number of convolutions, finite number of frames and finite number of labels. On the other hand, even in simple videos there exist several hundreds or thousands of frames with hundreds thousands of pixels per frame. To encode time, an algorithm must account for the practically innumerable dynamics between pixels within a frame and between frames. Today’s video algorithms learn their temporal representations directly on open and unconstrained latent spaces. This is problematic, as the algorithms learn to overfit to the innumerable spatiotemporal dynamics [G13, 33, 50] in the training data, rather than generalize to the general spatiotemporal dynamics patterns. Thus, **the third fundamental challenge** **is to design visual recognition algorithms that** despite their finite nature **can account for the** **innumerable spatiotemporal dynamics.** As a promising direction,I propose time-geometric visual algorithm**s**, by fitting known geometries to the spatiotemporal latent space.

**Aim and objectives**

**The overall aim** of this project are **visual recognition algorithms** that **given past** video observations **predict the expected future**, be it future object locations, future people’s actions, even future’s cause and effect patterns**.** I want to learn what people are likely to do next given their past actions and behavior. I want to learn where objects are expected given their past and co-located objects. Ideally, I want to even learn cause and effect mechanisms in visual interactions and reuse them in novel visual scenes. Notably, these are few examples only. **I propose to learn the expected futures given past observations by designing visual algorithms that are time-stochastic, time-intrinsic and time-geometric.**

Starting from the scientific challenges, the proposal has the following objectives.

|  |
| --- |
| O1. Study, develop and evaluate time-stochastic visual algorithms (first challenge)  O2. Study, develop and evaluate time-intrinsic visual algorithms (second challenge)  O3. Study, develop and evaluate time-geometric visual algorithms (third challenge)  O4. ShowTime: evaluation, dissemination & benchmarking |

The approach is timely and builds upon three key recent technological advances: *(i)* the rapid developments in video understanding with an emphasis on action recognition [G1, G4, G14, 21, 31-33, 35-39] and long-term tracking [G10, G15, 42-44], *(ii)* the immense progress in deep generative [51-54], geometric [30] and dynamical [29] learning, and *(iii)* the emergence of a plethora of rich video datasets [21, 28, 44, 45, 55-57] ranging from short [21] to very long [55], including specialized datasets that focus on temporality [28] or temporal relational reasoning [56].

**State-of-the-art**

From the pioneering spatiotemporal local descriptors [31] to spatiotemporal trajectories for actions [32], the interest in videos has a long history in visual recognition. With the arrival of deep learning the focus moved towards deep neural networks. Early attempts with recurrent neural networks [57] showed that pre-deep learning methods are surprisingly competitive and large datasets are key to better video understanding. Today, 3-d convolutions [21, 23] including spatiotemporal decompositions in my previous work [G14], and algorithms that consider spatiotemporal context [27] are the frontrunners in several video benchmarks. Generative models of visual sequences [54] with video pixel networks have also been explored. They model the entire visual sequence by a joint probability distribution, a design choice that is hard to satisfy for longer sequences and larger number of pixels. The above methods focus on making predictions in videos in past video observations.

Expecting the future by visual algorithms has received interest very recently. [59, 60] explore the role of motion in predicting the expected future. Both rely on domain knowledge: [59] uses optical flow and [60] pose and keypoint detectors. Considering motion from an egocentric point of view, [58] learn representations by extending the slow feature learning paradigm [77]. [63] propose pose detectors as an extra domain knowledge to act as supervision, so that to forecast future video. [61] also focus on predicting future motions by decomposing them into ego-motion and forward object motions. They also rely on the domain knowledge of optical flows, as well as on extra depth maps. [62] propose a multi-task approach that adds several sources of domain knowledge, including human behaviour and interactions between humans and objects, to detect future actions and locations. [64] proposed early action recognition where the beginning and end of the action is known in advance, whereas in previous work [G19] we explored online action prediction, a task that proved demonstrably harder. Predicting future trajectories [70, 71] has been also popular, assuming that perfect past trajectories are available a priori. Perfect future trajectories make sense when having reliable sensors like LIDAR. However, with RGB cameras the uncertainty into trajectories must also be taken into account [70]. Several works [65-67] proposed future frame prediction as a proxy task for unsupervised learning of representations. [65] propose a discretization of the visual semantic space such that to resort to language-like models of generation. [66] acknowledge the importance of uncertainty and propose to predict only future frames for which there is high confidence. And, [67] rely on the power of generative adversarial networks to predict future frames. Most algorithms that reconstruct future frames focus more on generating aesthetically pleasing visual reconstructions for the immediate next frames, rather than modelling the temporality in the visual sequence. Most of above works address the prediction of the expected future in the context of specific applications and propose application-specific algorithms using domain knowledge.

This proposal is interested in a general framework for learning to expect the future, with learning temporality as the central focus. Learning temporality has attracted sudden interest from a vision [35, 36, G13] and learning [49, 75, 76] perspective. The pioneering work of [35] studied the capacity of computer vision algorithms to recognize the arrow of time in visual sequences, namely whether the visual sequences are played in forward or reverse time. They proposed time-aware, handcrafted features, showing that considering temporality in the feature design is important. In their follow-up work [36], they attempted to improve further by casting the problem as representation learning. An important conclusion, especially in the context of the second workpackage of this proposal, is that algorithms with extended temporal footprint have better capacity in learning temporality. Further, they showed some early evidence that unsupervised and transfer learning is possible, by relying on pre-training with temporal self-supervision. In own work [G13], we generalized by considering different properties of temporality: temporal asymmetry, temporal continuity and temporal causality. We, further, proposed time-aligned DenseNets, which resemble unfolded recurrent neural networks with non-shared weights over time. The conclusion of this work was that adding time intrinsically, in the weights for time-aligned DenseNets, allowed for stronger temporal models that satisfy said properties of temporality better. These approaches are early attempts in understanding temporality in complex visual sequences. Understandably, they make several assumptions and have several limitations. For one, they cannot easily compete with strong supervision. They typically assume predetermined environments and data, which are easier to analyse. Importantly, it is not clear whether they generalize to harder visual sequences and setups, that last longer than a few seconds and comprise several complex actions. From a learning point [49, 75, 76] of view learning temporal abstractions has also provided promising evidence. [75] propose a variational version of recurrent neural networks, which however cannot automatically discover temporal structure, in the form of subsequences. Very recently, [49] propose a variational approximate inference approach for learning temporal abstractions with two objectives: being able to discover temporal structure, and also modelling its internal state variables stochastically. For their model, however, they assume that sequences have a strong hierarchical structure, which is not the case in realistic visual sequences where things are considerably fuzzier. Also, [76] recently proposed a temporal model that acts as a “mental simulator” of the world, building an abstract state of the world and its uncertainty. The strong assumption of this model, as well [49, 75], is that the world can be manipulated by reinforcement learning, which is only possible with artificial setups of synthetic environments. These algorithms present a promising start on modelling temporality. Future research is necessary, however, to design algorithms that are general, work on real data and do not assume artificial, interactive environments. This proposal focuses on a new synthesis of the vision and the learning perspective in learning temporality in visual sequences.

Today’s visual algorithms are time-deterministic, time-extrinsic and with unconstrained temporal latent spaces. This paradigm suffices when making predictions about past visual observations. However, it is unsuited when the goal is predictions about the expected future. The methodological originality of EVA comes from breaking with this paradigm. The proposition is to explore time-stochastic, time-intrinsic and time-geometric visual algorithms. By the end of the project, EVA will enable predicting the expected futures in long and complexvisual sequences with minimal manual supervision.

**Expected results, impact and knowledge utilization**

The overall aim of this project **ambitiously** seeks to contribute to shifting the field of visual recognition from explaining what has already been observed to predicting what is to be expected in the future. That is, answer questions *where* an object *will be* and *how will* it affect other objects, *what will* the person perform next or *what effect will* a particular visual intervention have. The high gain of this shift is evident, as it will allow for artificial intelligence that **prevents before the fact** and **plans ahead**, rather than explain after the fact.

My research in the past 5 years was on the frontier of another shift: taking the field of video understanding from handcrafted manual features to deep representation learning and large video datasets. I am confident that I can do a similar shift with the help of my new independent research team and close collaborators.

**Impact and knowledge utilization.** EVA addresses fundamental research challenges in visual recognition. However, its results are expected to serve as a basis for ground-breaking technological advances in practical applications. With the advent of deep learning, nowadays major technology and service vendors generate massive amounts of video sequence data. As video sensors are becoming extremely cheap to manufacture and purchase, soon most electronic devices will be equipped with a video sensor. It is estimated that by 2025, 45 billion cameras will record daily activities, providing vision to the Internet of Things and streaming content online. The demand (and hope) for improving technologies and services, or even for technological leaps, by leveraging this data is high. Learning to expect future outcomes given video sequences, as explored by EVA, will unlock high-impact research and applications.

A few examples:

I1. Today, video understanding and action recognition algorithms face the major problem of over-reliance on manual supervision. Also, the algorithms have trouble with videos of increased length and complexity. Advances from the proposed research, specifically time supervision for unsupervised training and the geometric memory mechanisms, will benefit greatly the state-of-the-art in video understanding.

I2. Use the researched visual recognition algorithms to monitor social media for security purposes by recognizing eminent malicious activities or threats in videos

I3. Use the researched visual recognition algorithms to monitor environmental changes, by analysing water levels and glacier melting rates from satellite footage and predicting their expected behaviour.

I4. Use the researched video algorithms to anticipate risky behaviour either from pedestrians, motorcyclists or other cars, thus improving the safety of autonomous driving platforms.

I5. Use the researched video algorithms for helping with cancer analysis by recognizing biomarkers in medical video imaging that specifies the tumour

I have already been taking concrete steps for translating the future research to utilization with collaborators for the last two cases, I3 and I4. The first case will be using the video recognition algorithms from the VIDI research for safety in autonomous driving, in conjunction with BMW. The second case will be using the video recognition algorithms from for biomedical analysis in cancer treatment, implemented within my own University spin off, Ellogon.AI. Next, I describe the concrete steps towards utilization of the VIDI research for these two cases.

**Cancer analysis with spin off Ellogon.AI.** In April 2019 I co-founded the University spin off, Ellogon.AI, with the mission to use visual recognition algorithms on histopathology, MRI and CT sequences for biomarker (imaging patterns found in biomedical data) recognition. The other co-founders of Ellogon.AI are: Dr. E. Kanoulas (associate professor in the University of Amsterdam, VIDI laureate), Dr. J. Teuwen (assistant professor in Radboud University and staff scientist in the Netherlands Cancer Institute), Dr. H. Horlings MD (head computational pathologist and group leader in the Netherlands Cancer Institute) and Dr. S. Dhawan (fund manager at Oncode Oncology Bridge Fund). Ellogon.AI is in talks with the University of Amsterdam and the Netherlands Cancer Institute on a deal that includes IP rights and licensing. The VIDI video recognition algorithms will be used for two technologies that Ellogon.AI currently explores. The first technology is using the VIDI video recognition algorithms to cross reference biomarker signatures in radiology and genome sequences. As video is cheaper than genome sequencing, the video recognition algorithms will be used to predict the particular type of cancer and the progress over time. The second technology is using the VIDI video recognition algorithms to track tumours in MRI sequences. As currently the tumours are localized offline only, the video recognition algorithms will enable live monitoring of the tumour. The location of the tumour will guide adaptive radiotherapy devices connected with the MRI recording device for delivering optimal radiation dosages. The goal is that early technology prototypes will be available within two years for valorisation.

**Safe autonomous driving with BMW.** Currently, I am working with BMW for forecasting the future spatiotemporal trajectories of road users given video, including pedestrians, cyclists and other cars. To predict the future trajectories, the past trajectories must be modelled so that to reason about the various future possibilities. Today’s algorithms assume perfect past trajectories with the use of expensive sensors like LIDAR. The VIDI video recognition algorithms will be used for modelling past trajectories from regular RGB videos, potentially replacing expensive LIDAR sensors. Specifically, the VIDI video recognition algorithms will take as input the videos from the vehicle camera and analyse the temporal behavior of pedestrians, cyclists and cars given past frames. Based on the past temporal behaviour, they will then forecast in RGB videos the likely next locations of all the objects in future time steps. The algorithms will be tested with the closed-loop BMW simulators. The most successful variants will further be tested with actual autonomous BMW vehicles. This has been agreed with the senior BMW researcher, N.S. Nagaraja (Naveen-Shankar.Nagaraja@bmw.de).

**Dissemination.** Besides translation of the researched technology to practical applications, other types of knowledge utilization include disseminating research to the public, nationally and internationally. Nationally, I will publicize research via the Deep Learning meetup that I currently organize and is open to the public from all backgrounds. Internationally, I will use the traditional academic channels, that is conferences, tutorials, workshops. Further, my intention is to disseminate research by continuing organizing the Video Understanding and Spatiotemporal Representations workshops also in the future editions of IEEE Conference in Computer Vision and Pattern Recognition. The temporal decathlon will be organized in the context of the said workshops. Algorithms, results, code and data will also be open sourced.

**Section B: Methodology**

EVA proposes four workpackages, one objective.

|  |
| --- |
| WP1. Time-stochastic visual algorithms (Objective O1: PhD1+PostDoc+PI)  WP2. Time-intrinsic visual algorithms (Objective O2: PhD2+PostDoc+PI)  WP3. Time-geometric visual algorithms (Objective O3: PhD3+PostDoc+PI)  WP4. ShowTime: evaluation, dissemination & benchmarking |

**WP1. Time-stochastic visual algorithms (O1: PhD1+PostDoc+PI)**

**Goal.** The goal for PhD1 in WP1 is visual algorithms that account for the high temporal uncertainty in visual sequences and their expected futures. The main idea is to design visual algorithms that do not assume that their inputs or outputs are fixed and deterministic; that is design time-stochastic visual algorithms. The challenge is to find a unified stochastic formalism that accounts for the different sources of stochasticity present in the data. In the description below, the tasks are organized chronologically.

**Description.** Stochastic visual algorithms do not assume that the inputs or the outputs are fixed to those observed in the training data. Rather, they assume there is a generating process, which produces all input and output patterns. Stochastic algorithms are not always necessary; not unless the setup is characterized by high uncertainty, for which deterministic models cannot learn general enough rules anymore. Predicting the expected future is a perfect example of high uncertainty regime. For one, there exist multiple plausible futures; in the end only one of the futures is observed. What is more, more often than not the appearances and dynamics in visual sequences entertain great amounts of noise that confuse deterministic algorithms [G13, G16, 35, 57]. Further, the expected future has no single temporal scale: there exist plausible expected future for both 5 seconds and 5 minutes ahead.

With PhD1 we investigate visual algorithms that incorporate stochasticity and randomness in their learning and design. In the first task (Task 1.1) we focus on the stochasticity in the appearances of the observed visual sequences. Specifically, we start from popular deterministic visual algorithms in video understanding, such as I3D [21] for short videos or our own Timeception [G14] for long videos. We reformulate these algorithms within the variational inference framework. Introducing variational inference is feasible with variational autoencoders [51]. The reason is that in the variational autoencoding framework three concepts must be defined. First, there should be an encoder that compresses the input to a code. Second, there should be a decoder that decompresses the code to a reconstruction. Third, there should be a stochastic layer in between the encoder and the decoder, which adds stochasticity to the code and represents like that the generating process. As shown in [51], the encoder and the decoder can effectively be any deterministic feature extractor and reconstructor. Thus, in the context of visual sequence we can use relevant visual algorithms like I3D [21] or Timeception [G14] as the encoders and decoders. Crucially, variational autoencoders minimize two losses: *(i)* a reconstruction loss of the generations returned by the decoder and compared to the true data, and *(ii)* a Kullback-Leibler regularization loss that measures the difference of the approximate posterior distribution from the prior distribution in the generating process. Normally, the regularization loss does not directly depend on the nature of the data. However, [72] showed that approximate posteriors in variational autoencoders may collapse to the prior distribution when the decoder is –or can be– too powerful. For images this can be a problem with powerful decoders like PixelCNN [53]. However, video sequences are much higher-dimensional and noisier, and there exist no powerful enough to date to cause problems. Thus, optimizing the regularization loss is not a problem. Regarding the reconstruction loss, when employing I3D in short videos there should be no problem either. However, with long visual sequences a pixel-reconstruction loss is less meaningful as the number of spatiotemporal appearances to reconstruct is likely too high. For long videos, we will follow with PhD1 a dual strategy, instead. First, we decompose the long visual sequences into shorter subsequences, whose pixels can be successfully reconstructed. That is to ensure that the encoders and decoders can receive relevant gradients and train themselves. Then, we additionally instruct the variational generator to also reconstruct the unit actions observed in the video. The unit actions are represented as one-hot vectors represented by a categorical distribution in the generation process. In a similar fashion, it is also possible to reconstruct the order of unit actions. In the end, we will have variational I3D and Timeception models that address stochasticity in the visual sequence appearances and actions. This is particularly important in longer videos, where higher uncertainty is expected.

While variational I3D and Timeception model uncertainty in appearances, they do not model uncertainty in frame transitions, especially the longer-range ones. To account for uncertain in the frame transitions, ttime-aligned DenseNets from previous work [G13] serves as inspiration. Time-aligned DenseNets connect frames and longer time steps explicitly via temporal residual connections. In the second task (Task 1.2) we start from this observation and define the temporal generating process on the basis of these residual connections. Recently, [73] showed that it is possible to transform residual layers such that they are invertible. Invertible residual layers have the advantage that they can be jointly optimized for classification, density estimation and stochastic generation without any adaptations. Their focus was on spatial residual blocks. Connecting their observation with the temporal residual connections from time-aligned DenseNets, in this task we will study with PhD1 temporally invertible stochastic residual networks. Specifically, we will explore the requirements for invertibility when residual connections are of temporal nature. What is more, different from [73], who focus on a single generation at the end of the network, in this task we have multiple time steps and multiple outputs. We, therefore, adapt the residual blocks such that each of them becomes a generator, conditioned on previous residual blocks.

As the ultimate goal is to have visual algorithms that predict the expected future, we should also account for the uncertainty in the temporal scale of the expected futures. In other words, we want to model “jumpy imagination” in future visual sequences, so that not only the content of the generation but also the time difference between futures. The advantage of having a stochastic visual algorithm for visual sequences is that we can introduce different prior distributions that fit our setup. For the third task (Task 1.3) a promising direction is Poisson generating processes as prior distributions to the proposed temporally-invertible residual networks from task 1.2. Poisson distributions are typically used [29] for modelling time series of discrete events where time differences are unknown and must be inferred. We will explore with PhD1 their use in high-dimensional and noisy visual sequences. Specifically, in task 1.3 the interest is not as much to the future appearances in the visual sequence, as to the future specific actions that are likely to take place. For instance, having observed a person moving to the kitchen, the algorithm should predict a possible future of washing the dishes in the next 20 seconds, or a possible future of preparing a meal in the next 2 minutes. In the last task (Task 1.4), PhD1 will focus on integration and efficiency.

**Datasets and evaluation.** As generative models are the focus, they can be trained on any video dataset. We will start with Something-Something [28], where temporality is critical for video understanding. For task 2.3, where the interest is also predicting the time difference un future actions, we need a video dataset with temporal annotations. In this case, we will start the Charades-Ego [55] dataset, which contains temporal annotations of sequences of actions.

**Pitfalls and solutions.** A potential risk is that combining temporally-invertible residual networks with Poisson prior distributions might be challenging. An alternative, more practical solution, will be to discretize the temporal scales we are interested when predicting the future. That way the space of time differences is a categorical variable, which could be easier to combine with.

**Final outcome.** The final outcome of this workpackage is time-stochastic visual algorithms that incorporate to their model design the stochasticity and randomness that is inherent in the appearances, frame transitions and action semantics in visual sequences. These time-stochastic visual algorithms will be able to generate multiple plausible expected futures given past visual observations. As such, they can be thereafter used for planning and decision taking.

**WP2. Time-intrinsic visual algorithms (O2: PhD2+PostDoc+PI)**

**Goal.** The ultimate goal is visual algorithms that can predict the expected future. As there are many plausible expected futures, manual strong supervision is undesirable. The goal for PhD2 in WP2 is, therefore, visual algorithms that can benefit from temporal self-supervision. The main idea and challenge is to incorporate time within the latent temporal representations of the visual algorithms, that is obtain time-intrinsic visual algorithms. The reason is that time-intrinsic visual algorithms will pertain neural networks that are more amenable to temporal self-supervision. In the description below, the tasks are organized chronologically.

**Description.** The literature [G2, 36, 37] has explored the capabilities of temporal self-supervision. While results show that temporal self-supervision does better than random initialization, the trained algorithms cannot outperform strong supervised pre-training by manual labels. Recently [36] showed that temporal self-supervision makes more sense when using visual algorithms with an extended temporal footprint. This is intuitive. The goal of video understanding algorithms is to quantify temporal nuances so that to automatically interpret the visual and temporal semantics in the visual sequence. In contrast, the goal of image understanding visual algorithms is to ignore temporal nuances, which are noise to the object semantics and the final classification. While reaching better accuracies, the algorithms in [36] can still not compete with manual supervision. A reasonable hypothesis is that their visual algorithms are still temporally not strong enough to take full advantage of the arrow of time.

Our recently published time-aligned DenseNets [G13] is a beneficial starting point. Revisiting the experiment with the temporal embeddings in Figure 2, we observe a much clearer separation of the forward and reverse time in videos with time-aligned DenseNets. This is further quantified in experiments, where time-aligned DenseNets are shown to predict more accurately the arrow of time, as well as temporal causality in Something-Something [28] videos. The conclusion is that time-aligned DenseNets are temporally stronger models.

Time-aligned DenseNets are temporally stronger by equating neural network layers to time steps. In the first task (Task 2.1) we generalize on time-aligned DenseNets. We can also view them as unrolled recurrent neural networks, whose weights and transition matrices are different per time step. That is, the recurrent neural network incorporates time intrinsically, as its structure (parameters, layers) are a function of time. We coin this time-aligned neural networks. Time-aligned neural networks are time-intrinsic, with each neural network layer receiving inputs from previous time step layers at , as well as from sequence inputs at time step . Time-aligned neural networks are markedly different from time-extrinsic visual algorithms [G4, G16, 21, 23, 59-63] that treat time as an external variable in the input or the output dimensions. However, time-aligned neural networks still rely on strong supervision with manual labels for training. This is in direct conflict with the focus of this workpackage. With PhD2 we will, therefore, research strategies how to integrate intrinsically temporal self-supervision constraints and replace manual supervision. The time-intrinsic structure changes will depend on the forms of temporal self-supervision. A first form of temporal self-supervision to consider is arrow of time. With the arrow of time supervision we will add frame-wise loss terms to consecutive neural network layer pairs (equivalent to time steps ). When the neural network layers make mistakes on the arrow of time, the respective layer time steps will immediately receive gradients that correct them. A second form of temporal self-supervision is odd-one-out predictions from previous works [G2]. When considering odd-one-out time supervision, a single loss term will connect to all neural network layer time steps. Mistakes in predicting the odd frame orders will then generate gradients that correct all layer time steps simultaneously. More time supervision strategies will be explored and integrated. Combinations of multiple time supervision loss objectives are also straightforward.

A drawback of time-aligned neural networks is that they have a fixed time horizon. Therefore, they cannot make predictions of the expected future at any temporal scale, rather they have a predetermined temporal scale that must be defined by the user. That is further problematic as time-aligned neural networks put a hard limit on the complexity of temporal patterns they consider. If the expected future has longer-range temporal dependencies with past observations, these will unavoidably be ignored. To address the limit on temporal patterns considered by time-aligned neural networks, the PhD2 will investigate in the second task (Task 2.2) a novel class of visual algorithms that allows for arbitrarily complex temporal patterns and video lengths. Specifically, PhD2 will explore the idea of parameterizing weights in video algorithms as functions of time in sequence, , instead of scalar values, , as with all visual algorithms to date. The difficulty is how to incorporate continuous functions for weights within the neural network structure. Last year’s best NeurIPS paper, NeuralODEs, [12] is inspirational. Time-aligned neural networks view the per layer neural network weights as a fixed length sequence, which is a discrete function of time. Similarly, NeuralODEs view the per layer neural network activation as an arbitrary length sequence, which is a continuous function of time. They then describe the per layer neural network activations by ordinary differential equations. Like in time-aligned neural networks, we are interested in having the layers and their weights as time functions. And like in NeuralODEs, we want these time functions to capture arbitrary dynamics and to not be confined to predetermined sequence lengths. We, therefore, explore neural networks whose weights are continuous functions of time, described by ordinary differential equations,

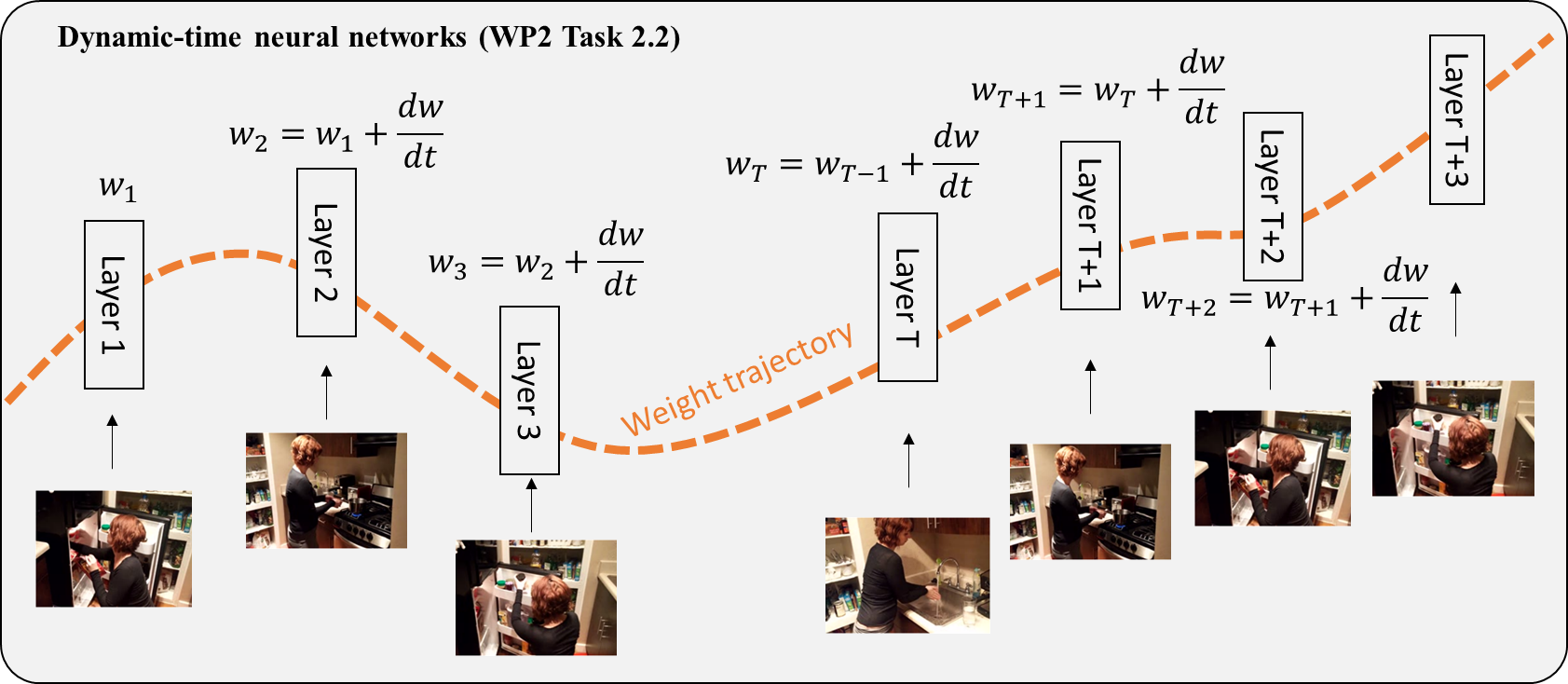
In the above equation is an auxiliary neural network. Our visual algorithm comprises two neural networks. The first main network processes the visual sequence and is parameterized by , where similar to time-aligned neural networks indicates both a time step and a neural network layer. The second networks it the auxiliary parameterized by , tasked to return the weights of the main neural network . As with [29], we implement the auxiliary neural network with an ODE solver that can be unrolled at will. The optimization then proceeds wither with Pontryagin’s adjoint method [29]. Temporal self-supervision from task 2.1 can easily be integrated in dynamic-time neural networks, as extra constraints in the ODE formulation. Importantly, by casting neural network weights as continuous functions of time, we meet both objectives for a time-intrinsic visual algorithm. For one, each neural network layer is a function of time and thus time is integrated in the structure of the algorithm. And, ODE solvers can be run at will, thus allowing for make predictions for expected futures at any temporal scale. An extra perk of using ODE solvers is that they can internally decide when to stop the weight trajectory based on an estimation of the remaining temporal error. An illustration of the dynamic-time neural networks is given in Figure 3.

Figure 3.Temporal self-supervision works best when time is incorporated within the neural network structure. What is more, when predicting the expected future one cannot assume a pre-determined temporal scale. In task 2.2 we explore dynamic-time neural networks, which express neural network weights as continuous functions of time, thus addressing both objectives.

As the ultimate goal is to have visual algorithms that predict the expected future, we then explore time-aligned and dynamic-time neural networks for spatiotemporal downstream tasks (Task 2.3). We focus on making future predictions for video object tracking and action recognition. For future video object tracking, we rely on siamese trackers from our previous work [G15] (simultaneously proposed by [42, 43]), which is the foundation of most long-term trackers nowadays [44]. For future action recognition, we rely on our previous work of Timeception [G14], which has showcased state-of-the-art performance in recognizing actions in long and complex videos. In the last task (Task 2.4), PhD2 will focus on integration and efficiency.

**Datasets and evaluation.** For this workpackage there are no special requirements regarding datasets. We will start with Something-Something [28], where temporality is critical for video understanding. The only exception is task 2.3, where spatiotemporal downstream tasks are considered. In this case, we will continue with long object tracking datasets including VOT [44] and our proposed OxUvA [G10] datasets. For future action recognition we will start with Charades-Ego, which contains continuous egocentric recordings of people performing sequential actions.

**Pitfalls and solutions.** Although literature has indicated that optimizing with ODE solvers [29] is straightforward, there is still a risk that optimization may be complex. Alternatively, we can recast and implement ODE solvers by backpropagation [29], as Euler’s method (backbone of ODE solvers) and Newton’s method (backbone of numerical gradient optimization) are both based on similar ideas. A second alternative is to fall back to fixed length, time-aligned neural networks and user hypernetworks [74] for learning the weights over time.

**Final outcome.** The final outcome of this workpackage is time-intrinsic visual algorithms that incorporate time in their latent structure. These time-intrinsic visual algorithms lead to temporally strong models, that eventually enabling accurate prediction of the expected futures given past visual observations.

**WP3. Time-geometric visual algorithms (O3: PhD3+PostDoc+PI)**

**Goal.** The goal for PhD3 in WP3 is visual algorithms that account for all the innumerable, past, present and future spatiotemporal dynamics despite their finite nature. This is important so that to avoid having visual algorithms that recognize only past spatiotemporal dynamics. The main idea of the workpackage is visual algorithms that constrain their temporal latent spaces by known geometries, that is obtain time-geometric visual algorithms. The challenge is to find the right formalism that forges a connection between time and geometry. In the description below, the tasks are organized chronologically.

**Description.** Traditional image and video recognition algorithms learn latent spaces directly from data. However, when the amount of patterns in the data is innumerable, as is the case with videos, or conversely when data is little, this can problematic. During training, the model learns representations that often underfit to the training spatiotemporal dynamics, as the finite number of parameters is far smaller than the number of patterns that must be learned. As a consequence, the learned models cannot generalize to test data and new spatiotemporal dynamics. To tackle this problem, forcing the latent space to have a specific geometry is beneficial [24]. The reason is that the neural network is then constrained to fit to the data a specific geometric landscape rather than any possible landscape. The algorithm learns the optimal fit for the specific geometry, given the spatiotemporal dynamics present in the training data. Then, the spatiotemporal dynamics in the test data will have to conform to the said geometry by construction, thus avoiding overfitting.

Breaking with the current paradigm of unconstrained temporal latent spaces, with PhD3 we investigate visual algorithms that constrain the latent temporal spaces to known geometries. A logical question is what type of geometry is a good fit for video sequences. To answer this question and introduce geometry in latent temporal spaces, the foundational work on geometric deep learning [30] sheds light. They argue that geometries can be either expressed in terms of graphs, or in terms of geometric manifolds. In order to account for the innumerable temporal patterns in video data, therefore, casting temporal latent spaces as graphs or manifolds is promising.

We first explore visual algorithms that learn undirected graph temporal representations (Task 3.1). With graphs two things need to be defined: the graph nodes, and the graph edges. The goal of the graph is to encode temporal relations given past video observations; then, eventually the graph should predict the expected future on the basis of predicting future temporal relations. Thus, a logical definition for the graph nodes is to represent salient temporal events in a specific video. For instance, salient temporal events can represent a person stirring in a cup, putting something into something, slicing something on a board or even extending their arm to open something. Then, the graph edges indicate temporal affinity between salient temporal events. For encoding a sequence of *putting sugar into the cup, then stirring*, the graph would comprise two nodes (node 1: putting something into something, node 2: stirring something in something) that are connected by an edge. A challenge is to define what constitutes salient temporal events. For defining salient temporal events ActionVLAD [39] and our Timeception [G14] are inspirational. ActionVLAD considers salient temporal events as centroids in linear Euclidean spaces. This is an assumption that does not necessarily hold in practice for complex data, like high-dimensional and noisy visual sequences. Instead, Timeception shows that complex temporal patterns can be captured by long temporal convolutions. Starting from these insights, with PhD3 we will view the salient temporal events as memories that can be shared between different visual sequences and between different time points (past, present and future). Then, we only need to learn the appropriate memory mechanism, whose outputs stores salient temporal events as a bank of temporal memory vectors. We do so using long temporal convolutions. The temporal memory vectors (akin to salient temporal events) are learned once during training and then reused ever after by all visual sequences. At inference time, any new frame is mapped to a single (hard assignment) or multiple (soft assignment) memories using either softmax or our recently proposed FouST estimator [G9]. A preliminary implementation of the said algorithm is illustrated in Figure 4a. Whenever a salient temporal event is detected in a video snippet, the respective temporal memory gets activated. Any two temporally adjacent memories will then be linked by an edge. For synthesizing larger graphs and graph hierarchies, graph neural networks are a good solution.

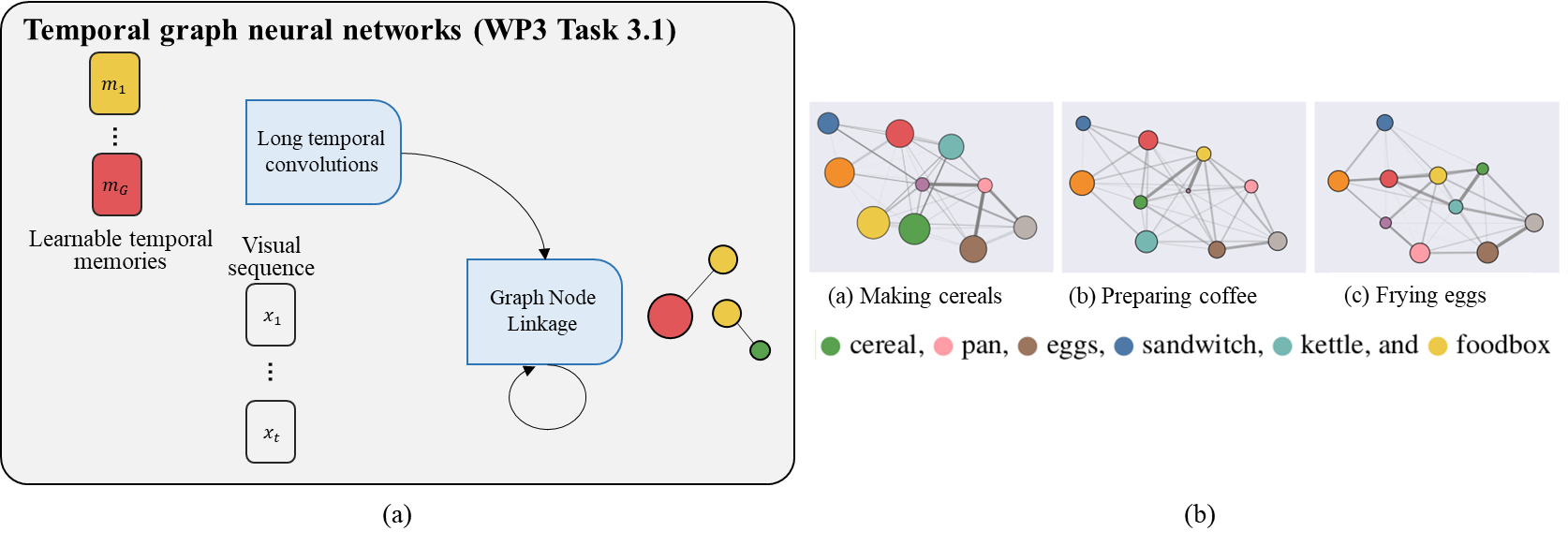
Our ultimate goal is to have visual algorithms that can predict the expected future. It is important to note that the salient temporal events, namely the graph nodes, are not tied to a particular past, present or future. However, the graph edges are tied to a past, present and future. Specifically, the past sequence forms a subgraph of all relevant salient temporal events, connected based on the past temporal relations. Then, predicting the expected future boils down to a graph linkage completion problem [40], namely finding the temporal edges that would satisfy a plausible future. For the graph linkage completion any of the available solutions [40] in the context of graph neural networks would suffice. With a graph representation of a visual sequence we improve also in explainability. Normally, since the temporal memory vectors are learned, they are not necessarily “nameable”. However, as the memory vectors must serve a great deal of different videos, the expectation is that in practice temporal memories will converge to semantically (but not absolutely) consistent temporal concepts. The final graph then will be able to explain a visual sequence as a conglomeration of semantically consistent temporal concepts, connected to each other in a particular way. An interesting extra perk is that since the graph stores only salient temporal events, the graph temporal representations scale sub-linearly with time. This will allow for processing even hour long sequences, which is quite important for practical applications. Results from a preliminary implementation that prove the principle are available and shown in Figure 4b.

Figure 4. (a) Learning undirected temporal graph representations in task 3.1. Graph nodes are salient temporal events, modelled as temporal memories, which are linked when they are temporally adjacent. (b) Preliminary experiments of temporal graph representations yield nodes that carry semantic meaning, even without user supervision. Given the temporal graphs, one can predict the expected future by graph linkage completion, using graph neural networks [40].

From a geometric point of view, undirected graphs bring a global understanding of temporal relations in a video. However, the lack of directionality in task 1.1 implies also lack of local temporal understanding. In Figure 4b, for instance, the graph can tell us that globally in the video, there in visual sequences of preparing breakfast, cereal is temporally related with a food box, with eggs, with a kettle. However, it cannot tell us whether cereal was stirred with eggs (unlikely), or whether the person first ate cereal, then eggs (likely). Such local temporal context is often critical. For instance, a person pushing violently another person all of a sudden implies that the algorithm shall expect a fight very soon. In contrast, a person pushing violently another person when a vehicle approaches implies that the algorithm shall expect a forthcoming accident. In the second task (Task 3.2) of WP3, PhD3 will research how to incorporate local context and directionality to temporal graph representations. In the context of graph geometry, time-variant spectral convolutions [30] (like heat kernels) are a good starting point. Another possibility would be to adapt causal convolutions –used in generative models like PixelCNN [53] – for temporal events. In that case, every pair of graph nodes will be connected with two edges, one per directionality.

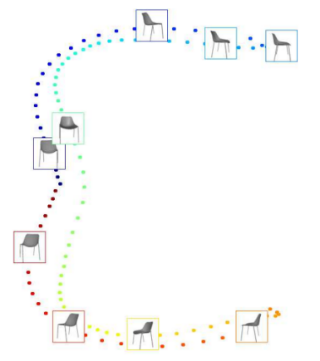
Parting from graphs as a definition of geometry, next we will explore geometric manifolds as representations of temporality learned from visual sequences. In previous work [G15] we showed that all appearance changes of a single object can be reduced to a similar neighbourhood in the latent space with great success. While this proves that it is possible to cluster wildly different appearances reliably and discriminatively together, this approach eventually does also discards time. Instead, what we want is to define a geometric manifold on the latent temporal space, such that any trajectory of coordinates on the manifold represents a visual sequence. The expected future is then recovered by continuing the trajectories on the geometric manifold. The work from [41] is inspirational. They show that when the shape and pose of an object is known, algorithms can learn latent spaces that resemble geometric manifolds (Figure 5). However, their approach requires knowing the shape of the object, which is impossible beyond CAD models and rendered objects. In the third task (Task 3.3), PhD1 will focus on the reverse problem. Using the temporality in video sequences of an object, learn geometric manifolds such that nearby manifold coordinates represent smooth appearance changes over time, yielding time equivariance. The challenge is how to define a flexible enough geometry, and, thereafter, how to optimize for it based on the training data. A good starting point is using hyperspheres as latent geometric manifolds, inspired by our recent work [G8]. In the last task (Task 3.4), PhD1 will focus on integration and efficiency.

Figure 5. Geometric transformations are similar to temporal appearance changes. In task 3.3 we represent temporal trajectories as trajectories on geometric manifolds. Image credit [41].

**Datasets and evaluation.** For easier evaluation, in this workpackage we need two types of datasets. For task 3.1 and 3.2 we need video datasets with long actions formulated as sequence of smaller, unit actions. This way it will be easier to evaluate to what extent the salient temporal events correspond to unit actions. Also, graphs can represent long actions and, therefore, long videos make sense. Fitting datasets are the Charades-Ego [55] and the recently proposed CATER [56], designed to evaluate temporal reasoning. In contrast, for task 3.3 we want video datasets depicting a single or few objects in the majority of the sequence. The reason is that it will be easier to evaluate the capacity of the geometric manifold to represent the temporal changes in the appearance and geometry of the object through time. Fitting datasets are object tracking ones, such as VOT [44] or OxUvA [G10] from previous work. That said, the algorithms in task 3.3 are expected to work with any visual sequence, although in that case the interpretation of the manifold will likely be harder.

**Pitfalls and solutions.** A potential risk is that hyperspheres as a geometric manifold might be hard to optimize. The reason is that paradoxically, when increasing the dimensions of a hypersphere (latent dimensions in our algorithm) the surface of hypersphere starts shrinking with a limit to zero [G8]. This is problematic, since a higher number of latent dimensions is required for complex data like visual sequences. An alternative will be to consider “onions” of multiple, concentric hyperspheres with different radii, each comprising a different subspace in the latent space.

**Final outcome.** The final outcome of this workpackage is time-geometric visual algorithms that enforce specific geometries to their temporal latent representations. These time-geometric visual algorithms encode time geometrically, allowing for predicting the expected future even with very long visual sequences and innumerable spatiotemporal dynamics.

**WP4. ShowTime: evaluation, dissemination & benchmarking (O4: PostDoc+PI)**

**Goal.** The goal in WP4 is dual: *(i)* organize a novel temporal decathlon benchmark for measuring the temporal behavior of visual algorithms, then *(ii)* evaluate the capacity of existing visual algorithms in key temporal properties*.* The PostDoc will take the lead and organize the decathlon and the evaluation as an open international competition. The PhDs will also help. I will supervise. The competition will be incorporated in a series of workshops on video understanding that I will annually organize with other leading researchers in the field. They will be held in conjunction with either the IEEE Conference on Computer Vision and Pattern Recognition, the International or the European Conference in Computer Vision. Ideally, the decathlon and the evaluation should work as platform for the PostDoc and the team to gain visibility and research embedding.

**Description.** The workpackage splits into two tasks.

Task 4.1. Temporal decathlon. Proper evaluation and benchmarking is key for learning successful and robust visual algorithms, as history has shown [45]. Evaluating the expected future given past observations in a visual sequence has advantages and disadvantages. An advantage is that given any video dataset, one can always split videos into “past subsequences” and “future subsequences” for free (single-mode future evaluation). The disadvantage is that single-mode future evaluation only represents one of the possible futures; there are several other plausible futures that an algorithm could predict. Solutions have been proposed, *e.g.*, best-of-many samples evaluation [70], all with limitations. Pragmatically speaking, this disadvantage incapacitates direct evaluation without resorting to synthetic visual environments, which in turn are hard to make realistic enough. Instead, with the postdoctoral fellow we explore an indirect evaluation of expecting visual futures, specifically by evaluating temporality in algorithms. As temporality is key to predicting the expected future, temporality can be used as an indirect evaluation platform.

Inspired by [44, 45], with the PostDoc we will devise and organize a temporal decathlon as a benchmark for measuring the temporal behaviour and properties of video algorithms. As nowadays dataset sizes become incompatible with the quick research cycles needed, the intention is a carefully selected set of axes of comparison that offer maximum insight for minimum data effort. This decathlon will evaluate the capacity of visual algorithms in satisfying key properties of temporality, like spatiotemporal continuity, spatiotemporal directionality, spatiotemporal localization and others. The plan is not to make one more dataset; rather the plan is to reuse existing excellent resources [21, 28, 35, G10, 55, 56]. At the moment, reasonable axes of comparisons that bring complementary insights are: (i) spatiotemporal aggregation: offline action classification, video compression, (ii) spatiotemporal continuity: future object trajectory prediction, detecting cell splitting and death [46], (iii) spatiotemporal forecasting: future trajectory prediction, future tracking forecasting, (iv) spatiotemporal localization: MRI-guided tumour tracking [47], video object segmentation. In this endeavour my previous experience [G10] on building a dataset and benchmark is useful. I plan is to collaborate with peers to share the effort and maximize impact. The initial edition of the decathlon will be carried out within the first twelve months.

Task 4.2. Evaluation. The goal of the temporal decathlon is to define axes of comparison between algorithms on visual sequences. On the basis of these axes different visual algorithms will be evaluated. The result will be an analysis of the when, where and under what conditions each visual algorithm captures temporality; and, by consequence, which of these algorithms can be used to predict the expected future. Algorithms, results, code and data will be open sourced for public access. That is important not only for designing novel video algorithms. It is also important for opening up the field of spatiotemporal visual algorithms to other domains, like medical or biomedical data, where it is important to predict the expect future, and researchers know what type of temporal properties are relevant. The evaluation is inspired by previous endeavours in visual object tracking [44, 45], where the introduction of systematic benchmarking and evaluation has led to tremendous progress [G15, 42, 43]. The evaluation and the results will be carried out by the second year of the project.

**Final outcome.** The final outcome of this workpackage is a decathlon benchmark and evaluation platform to compare on equal grounds visual algorithms with respect to temporality and indirectly, their ability to predict expected futures.

**Research context**

As part of this proposal, I have agreed with Prof. A. Zisserman from the Visual Geometry Group in the University of Oxford for a research visit and collaboration. Further, I will continue my ongoing collaborations. Locally, I will continue my collaboration with Prof. C.G.M. Snoek, world expert in multimedia and video, and with Prof. M. Welling, world expert in machine learning. Internationally, I have collaborations with leading scientists in top institutions: A. Vedaldi (University of Oxford), P. Torr (University of Oxford), S. Gould (Australian National University), B. Fernando (A\* Institute, Singapore), H. Bilen (University of Edinburgh). Last, I have agreed with Prof. A. Zisserman (University of Oxford), world expert in computer vision and actively working on video algorithms for a research visit. I plan rich exchanges for the students to the research labs I am collaborating with in Oxford, Edinburgh and Singapore.

**Feasibility**

Theproposed novel direction starts from several published works and experience I have collected over the years. Furthermore, the proposed research will be conducted within the context of world-class computer vision and machine learning specialists, including a research visit to one of the top research groups in computer vision. I am confident that the proposed research and direction is feasible.

**Research team**

I will devote 50% of my research time to EVA. As PI, I will be in charge of the overall scientific direction and management of the project. The team will further comprise three PhD students and a postdoctoral fellow. The PhD1-3 students will work on the workpackages WP1-3. They will be supervised by the postdoctoral fellow and me. Together, we will design the core methods that address the main challenges of the proposal. While the PhD students will work independently, collaborations between them and the rest of the group are encouraged. Further, I believe in a data and code sharing philosophy and I plan to have students benefit from each other in that respect. On WP4 containing the temporal decathlon, the PostDoc will take the lead. The PhD students will also help in the decathlon. I will supervise.

**Recruitment.** I plan to hire the first PhD in the beginning of the project, while the other two PhD students 8 months later. This is after personal experience with managing a large lab of students (12 students in QUVA Lab that I manage), where it seemed beneficial to spread the start times of students. The PostDoc will be hired after two years. That way the PhD students will have gathered experience. So, I will be able to give more freedom and the collaboration with the PostDoc will have maximum efficacy.

**Practical timetable.** The timetable for the said research, including the expected outcomes is as follows. The last year will serve as a buffer and also for further dissemination and knowledge utilization.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Workpackages | YEAR 1 | | | YEAR 2 | | | YEAR 3 | | | YEAR 4 | | YEAR 5 | |
| **WP1: TIME-STOCHASTIC VIDEO ALGORITHMS** |  |  |  | |  |  | |  |  | | *○* |  |  |
| TASK1.1: VISUAL ALGORITHMS ON SEQUENCES & VARIATIONAL INFERENCE |  | *▪* |  | |  |  | |  |  | |  |  |  |
| TASK1.2: TEMPORALLY-INVERTIBLE STOCHASTIC RESIDUAL NETWORKS |  |  | *▪* | | *▪* |  | |  |  | |  |  |  |
| TASK1.3: JUMPY FUTURES WITH POISSON GENERATIVE PROCESSES |  |  |  | |  | *▪* | | *▫* |  | |  |  |  |
| TASK1.4: INTEGRATION AND EFFICIENCY |  |  |  | |  |  | |  | *▫* | |  |  |  |
| **WP2: TIME-INTRINSIC VIDEO ALGORITHMS** |  |  |  | |  |  | |  |  | | *○* |  |  |
| TASK2.1: TIME-ALIGNED NEURAL NETWORKS |  | *▪* |  | |  |  | |  |  | |  |  |  |
| TASK2.2: DYNAMIC –TIME NEURAL NETWORKS |  |  | *▪* | | *▫* | *▪* | |  |  | |  |  |  |
| TASK2.3: EXTENDING MODELS FOR FUTURE PREDICTION |  |  |  | |  |  | |  | *▫* | |  |  |  |
| TASK2.4: INTEGRATION AND EFFICIENCY |  |  |  | |  |  | | *▪* |  | |  |  |  |
| **WP3: TIME-GEOMETRIC VIDEO ALGORITHMS** |  |  |  | |  |  | |  |  | | *○* |  |  |
| TASK4.1: VISUAL SEQUENCES AS UNDIRECTED TEMPORAL GRAPHS |  | *▪* |  | |  |  | |  |  | |  |  |  |
| TASK4.2: VISUAL SEQUENCES AS DIRECTED TEMPORAL GRAPHS |  |  | *▪* | | *▫* | *▪* | |  |  | |  |  |  |
| TASK4.3: TEMPORAL GEOMETRIC MANIFOLDS |  |  |  | |  |  | |  | *▫* | |  |  |  |
| TASK4.4: INTEGRATION AND EFFICIENCY |  |  |  | |  |  | | *▪* |  | |  |  |  |
| **WP4: SHOWTIME: EVALUATION & BENCHMARKING** |  |  |  | |  |  | |  |  | |  |  |  |
| TASK4.1: TEMPORAL DECATHLON |  | *▪*◊ |  | | ◊ |  | | ◊ |  | | ◊ |  | ◊ |
| TASK4.2:EVALUATION |  |  |  | | ▫ |  | |  |  | |  |  |  |
| ▪ 13x Conference papers ▫ 7x Journal papers ◊ 2x Demonstrators ◊ 3x Demonstrators (PI) ○ 3x PhD theses | | | | | | | | | | | | | |

# References by PI without supervisors

|  |
| --- |
| [G1] B. Fernando, E. Gavves, J.M. Oramas, A. Ghodrati and T. Tuytelaars, “Modeling video evolution for action recognition,” in *CVPR*, 2015. |
| [G2] B. Fernando, H. Bilen, E. Gavves and S. Gould, “Self-Supervised Video Representation Learning with Odd-One-Out Networks,” in *CVPR*, 2017. |
| [G3] X. Jia, E. Gavves, B. Fernando and E. Gavves, “Guiding the long-short term memory model for image caption generation,” in *CVPR*, 2017. |
| [G4] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi and S. Gould, “Dynamic image networks for action recognition,” in *CVPR*, 2017. |
| [G5] B. Fernando, E. Gavves, J.M. Oramas, A. Ghodrati and T. Tuytelaars, “Rank pooling for action recognition,” in *CVPR*, 2015. |
| [G6] C. Oh, E. Gavves and M. Welling, “[BOCK: Bayesian Optimization with Cylindrical Kernels](javascript:void(0)),” in ICML, 2018. |
| [G7] C. Oh, J. Tomczak, E. Gavves and M. Welling, “Combinatorial Bayesian Optimization using Graph Representations,” in *NeurIPS*, 2019. |

# References by PI as main supervisor

|  |
| --- |
| [G8] T. Davidson, J. Tomczak and E. Gavves, “Increasing Expressivity of a Hyperspherical VAE,” in NeurIPS, Workshop on Bayesian Deep Learning , 2019. |
| [G9] A. Pervez, T. Cohen and E. Gavves, “Low Bias Gradient Estimates for Very Deep Boolean Stochastic Networks,” ICLR (openreview preprint), 2019. |
| [G10] J. Valmadre, L. Bertinetto, J. Henriques, R. Tao, A. Vedaldi, A. Smeulders, P. Torr and E. Gavves, “Long-Term Tracking in the Wild: A Benchmark,” in 2018, ECCV |
| [G11] Y. Chen, T. Mensink and E. Gavves, “3D Neighborhood Convolution: Learning Depth-Aware Features for RGB-D and RGB Semantic Segmentation,” 3DV, 2019. |
| [G12] S. Shkodrani, M. Hoffman and E. Gavves, “Dynamic Adaptation on Non-Stationary Visual Domains,” ECCVW, 2018. |

# References by PI with supervisors

|  |
| --- |
| [G13] A. Ghodrati, E. Gavves and C. Snoek, “Video Time: Properties, Encodings, Evaluation,” in BMVC, 2018. |
| [G14] N. Hussein, E. Gavves and A. Smeulders, “Timeception for Complex Action Recognition,” in CVPR, 2019 |
| [G15] R. Tao, E. Gavves and A. Smeulders, “Siamese Instance Search for Tracking,” in CVPR, 2016 |
| [G16] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain and C. Snoek, “VideoLSTM convolves, attends and flows for action recognition,” CVIU, 2018 |
| [G17] E. Gavves, T. Mensink, T. Tommasi, C. Snoek and T. Tuytelaars, “Active transfer learning with zero-shot priors: Reusing past datasets for future tasks,” in ICCV, 2015. |
| [G18] E. Gavves, B. Fernando, C. Snoek, A. Smeulders and T. Tuytelaars, “[Fine-grained categorization by alignments](javascript:void(0)),” in ICCV, 2013. |
| [G19] R. de Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, T. Tuytelaars, “[Online action detection](javascript:void(0)),” in ECCV, 2016. |

# Other references

|  |
| --- |
| [20] Y. LeCun, Y. Bengio and G. Hinton, “Deep learning,” Nature, 2015 |
| [21] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” in CVPR, 2017. |
| [22] B. Zhou, A. Andonian, A. Oliva and A. Torralba, “Temporal Relational Reasoning in Videos,” in ECCV, 2018 |
| [23] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” in ICCV, 2015 |
| [24] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, 1997 |
| [25] J. Chung, Ç. Gülçehre, K. Cho and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” CoRR abs/1412.3555, 2014. |
| [26] G. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev and A. Gupta, “Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding,” in ECCV, 2016. |
| [27] X. Wang, R. Girshick, A. Gupta and K. He, “Non-Local Neural Networks,” in CVPR, 2018 |
| [28] R. Goyal and e. al., “The "Something Something" Video Database for Learning and Evaluating Visual Common Sense,” in ICCV, 2017 |
| [29] T. Q. Chen et al., “Neural Ordinary Differential Equations,” in NeurIPS, 2018 |
| [30] M. Bronstein, J. Bruna, Y. LeCun, A. Szlam and A. Vandergheynst, “Geometric Deep Learning: Going beyond Euclidean data,” IEEE Signal Processing, 2017 |
| [31] I. Laptev, “On Space-Time Interest Points,” IJCV, 2005. |
| [32] H. Wang and C. Schmid, “Action Recognition with Improved Trajectories,” in ICCV, 2013 |
| [33] A. we, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, “Large-Scale Video Classification with Convolutional Neural Networks,” in CVPR, 2014 |
| [34] C. Szegedy and e. al., “Going deeper with convolutions,” in CVPR, 2015. |
| [35] L. Pickup, Z. Pan, D. Wei, Y.-C. Shih, C. Zhang, A. Zisserman, B. Schölkopf and W. Freeman, “Seeing the Arrow of Time,” in CVPR, 2014. |
| [36] D. Wei, J. Lim, A. Zisserman and W. Freeman, “Learning and Using the Arrow of Time,” in CVPR, 2018. |
| [37] I. Misra, C. L. Zitnick and M. Hebert, “Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification,” in ECCV, 2016. |
| [38] G. Varol, I. Laptev and C. Schmid, “Long-Term Temporal Convolutions for Action Recognition,” PAMI, 2018. |
| [39] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic and B. Russell, “ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification,” in CVPR, 2017. |
| [40] T. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” in ICLR, 2017. |
| [41] M. Aubry and B. Russell, “Understanding Deep Features with Computer-Generated Imagery,” in ICCV, 2015. |
| [42] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik and P. Torr, “Fully-Convolutional Siamese Networks for Object Tracking.,” in ECCC Workshop, 2016. |
| [43] D. Held, S. Thrun and S. Savarese, “Learning to Track at 100 FPS with Deep Regression Networks,” in ECCV, 2016. |
| [44] M. Kristan and e. al., “The Sixth Visual Object Tracking VOT2018 Challenge Results,” in ECCV Workshop on Visual Object Tracking, 2018. |
| [45] A. Smeulders, D. Manh Chu, R. Cucchiara, S. Calderara, A. Dehghan and M. Shah, “Visual Tracking: An Experimental Survey,” PAMI, 2014. |
| [46] M. Hossain, A. K. Gostar, A. Bab-Hadiashar and R. Hoseinnezhad, “Visual Mitosis Detection and Cell Tracking Using Labeled Multi-Bernoulli Filter,” IEEE Fusion, 2018. |
| [47] L. Cerviño, D. J. and S. Jiang, “MRI-guided tumor tracking in lung cancer radiotherapy,” Physics in Medicine and Biology, 2011. |
| [48] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, “A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation,” in CVPR, 2016. |
| [49] T. Kim, S. Ahn and Y. Bengio, “Variational Temporal Abstraction,” arXiv preprint arXiv:1910.00775 [cs.LG], 2019. |
| [50] R. Pascanu, T. Mikolov and Y. Bengio, “On the difficulty of training Recurrent Neural Networks,” in ICML, 2013. |
| [51] D. Kingma, M. Welling, “Autoencoding Variational Bayes,” in ICLR, 2014. |
| [52] D. Rezende, S. Mohamed, “Variational Inference with Normalizing Flows,” in ICML, 2015. |
| [53] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, K. Kavukcuoglu, “Conditional Image Generation with PixelCNN Decoders,” in ICML, 2013. |
| [54] N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, K. Kavukcuoglu, “Video Pixel Networks,” in ICML, 2017. |
| [55] G. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, K. Alahari, “Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos,” in CVPR, 2018. |
| [56] R. Girdhar, D. Ramanan, “CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning,” arXiv preprint arXiv:1910.04744, 2019. |
| [57] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, “Large-scale Video Classification with Convolutional Neural Networks,” in CVPR, 2014. |
| [58] D. Jayaraman, “Learning image representations tied to ego-motion,” in ICCV, 2015. |
| [59] X. Liang, L. Lee, W. Dai and E. Xing, “Dual Motion GAN for Future-Flow Embedded Video Prediction,” in ICCV, 2017. |
| [60] Z. Liu, S. Wu, S. Jin, Q. Liu, S. Lu, R. Zimmermann and L. Cheng, “Towards Natural and Accurate Future Motion Prediction of Humans and Animals,” in CVPR, 2019. |
| [61] X. Qi, Z. Liu, Q. Chen and J. Jia, “3D Motion Decomposition for RGBD Future Dynamic Scene Synthesis”, in CVPR, 2019. |
| [62] J. Liang, L. Jiang, J. Niebles, A. Hauptmann and L. Fei-Fei, “Peeking Into the Future: Predicting Future Person Activities and Locations in Videos”, in CVPR, 2019. |
| [63] J. Walker, K. Marino, A. Gupta and M. Hebert, “The Pose Knows: Video Forecasting by Generating Pose Futures.”, in ICCV, 2017. |
| [64] M. Hoai and F. de la Torre, “Max-Margin Early Event Detectors”, in CVPR, 2012. |
| [65] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, “Video (language) modeling: a baseline for generative models of natural videos”, arxiv preprint arXiv:1412.6604, 2014. |
| [66] D. Jayaraman, F. Ebert, A. Efros, S. Levine, “Time-Agnostic Prediction: Predicting Predictable Video Frames”, in ICLR, 2019. |
| [67] Y. Kwon, M. Park, “Predicting Future Frames Using Retrospective Cycle GAN”, in CVPR, 2019. |
| [68] R. Krishnan, U. Shalit, David Sontag, “Deep Kalman Filters”, arxiv preprint arXiv:1511.05121, 2015. |
| [69] M. Gemici et al., “Generative Temporal Models with Memory”, arxiv preprint arXiv:1702.04649, 2017. |
| [70] N. Lee, W. Choi, P. Vernaza, C. Choy, P. Torr, M. Chandraker, “DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents”, in CVPR, 2017. |
| [71] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, “Social LSTM: Human Trajectory Prediction in Crowded Spaces”, in CVPR, 2017. |
| [72] I. Gulrajani, k. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, A. Courville, “PixelVAE: A Latent Variable Model for Natural Images”, in ICLR, 2017. |
| [73] Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, Jörn-Henrik Jacobsen, “Invertible Residual Networks”, arxiv preprint arXiv:1811.00995, 2019. |
| [74] D. Ha, A. Dai, Q. Le, “Hypernetworks”, in ICLR, 2017. |
| [75] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, Y. Bengio, “A Recurrent Latent Variable Model for Sequential Data”, in NeurIPS, 2015. |
| [76] K. Gregor, G. Papamakarios, F. Besse, L. Buesing, T. Weber, “Temporal Difference Variational Autoencoder”, in ICLR, 2019. |
| [77] L. Wiskott, T. Sejnowski, “Slow Feature Analysis: Unsupervised Learning of Invariances”, in Neural Computation, 2002. |

1. *One can hardcode handcrafted yet brittle rules to make predictions in simple videos like this with static background.* [↑](#footnote-ref-1)