Transfer of Status Report
Generative Sequence Modelling for Model-based Reinforcement Learning

Adam Kosiorek[1]

Supervisor: Prof. Ingmar Posner

I. Introduction

Sequence prediction is a powerful framework for unsupervised learning. Its utility can be intuitively explained by the fact that predicting the future, if it is to be done well, requires a very good understanding of the present. If a model can learn the idea of an object and infer intuitive physics from data, it should be able to constrain its predictions to the ones where objects obey physically plausible trajectories: e. g., a car should not dissolve into thin air. Moreover, the availability and abundance of real-world sequential data make it possible to train large and highly-complex models for this task. While modern approaches based on neural networks have achieved considerable success, they generally do not take any domain-specific problem structure into account, nor do they provide transferable representations that could be easily used for downstream tasks.

Recent advances in variational inference and neural networks allow building scalable generative latent-variable models of high-dimensional data. On one hand, the latent variables explain observations, are typically low-dimensional, and can be used in downstream tasks. On the other, this approach results in an approximation to the true probability distribution of the data, which allows generation of multiple trajectories from a single starting point. One could argue that stochasticity is not necessary in fully-observable environments. In the real-world, however, partial-observability is most often the case and a stochastic system can act as a simulator conditioned on imperfect information. A stochastic and imperfect simulator of this kind can be used for model-based reinforcement learning (RL) as argued by Sutton, 1991 to improve sample-efficiency of model-free approaches. The majority of sequence predictors that act in the image space has the problem, however, that the further in time they venture, the more blurry and undefined the predictions become. With the increasing divergence

[1]Applied Artificial Intelligence Lab, Oxford Robotics Institute, University of Oxford.

between the true data distribution and the predictive distribution the quality of model-free policies decreases and planning becomes inaccurate.

All work performed under this thesis will aim at constructing a generative latent-variable model of sequential data, where the model structure introduces prior information about the task. Specifically, the structure of the model will encourage learning intuitive physics and decomposing scenes into their constituent components; by modelling moving objects separately, it will be possible to perform counter-factual stochastic simulation. One of the core features of the model is the ability to simulate in the latent space, therefore circumventing the issue of blurry predictions. Since latent variables describe the state of the world, we will aim at conditioning a model-free policy on these latent variables for action prediction. Finally, we will test the generative model of the environment in the Dyna framework (Sutton, 1991).

The rest of this paper is structured as follows. Section II covers prior work related to the areas in question. I summarise the task of sequence prediction, describe relevant variants of unsupervised learning, investigate how model structure helps to learn abstract concepts from data and examine prior work on Dyna. In section III, I describe our work on object tracking and how it ties with my interests and the planned future work. Section IV details how we are going to build a structured generative model of sequences and use it in model-based RL. Section V concludes this work.

II. Related Work

A. *Unsupervised Learning via Generative Modelling*

While data in general is abundant and cheap, data for supervised learning is often expensive and time-consuming to gather. The majority of ML algorithms require relatively large amounts of labelled training data. One of the explanation states that they start learning without any prior knowledge of the world. This is in stark contrast to humans, who not only have a vast amount of knowledge about the world, but also expand it continuously and without any supervision (Friston, 2009). One alternative is to perform generative modelling of the probability distribution $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z}$ of observations $\mathbf{x}$ in terms of some latent variables $\mathbf{z}$. The latent variables *explain* the observations and can make the joint distribution $p(\mathbf{x}, \mathbf{z})$ tractable even in the case of an intractable marginal distribution. The latent encoding can be used in downstream tasks e. g., for transfer or semi-supervised learning (Pan and Yang, 2010). Hinton, Osindero, and Teh, 2006 introduced Deep Belief Networks (DBN) which explain the observations in terms of Bernoulli latent variables. Alternatively, we can introduce an approximate posterior distribution and approximate the true data distribution by maximising the evidence lower bound (ELBO) on the log probability of the data. This approach results in variational autoencoders (VAE) (Kingma and Welling, 2014; Danilo J Rezende, Mohamed, and Wierstra, 2014). VAEs are much more flexible than DBNs as they allow latent variables from arbitrary probability distribution functions (pdf) and can be trained end-to-end with off-the-shelf gradient-based methods. Performance of VAEs depends on the choice of the approximate posterior distribution and a prior for the latent space. Since the latter is stochastic, the variance of the gradient estimator is increased compared to deterministic neural networks, which leads to slower convergence. These approaches are primarily suited to modelling datasets of independent and identically distributed (*i.i.d.*) points.

B. Sequence Modelling

Traditional approaches to sequence modelling often consider inference of latent variables that explain the data e. g., linear dynamical systems or hidden markov models (Bishop, 2006). They often require dynamics of the system to be known and often have too little capacity to model complex and high-dimensional real-world data. Neural networks, on the other hand, can learn both features and state dynamics from data and they can approximate functions of arbitrary complexity with arbitrary precision. Even early works on the topic demonstrated how useful neural networks are for prediction of chaotic time-series (Lapedes and Farber, 1988). Since then, neural networks have been successfully applied to sequence classification and prediction in different domains: written natural language, speech and audio, motion capture data or brain waves (Längkvist, Karlsson, and Loutfi, 2014). Sequence prediction is a promising method of unsupervised learning. The task is to predict the observation at time $t+1$ given a sequence of observations $\mathbf{x}_{1:t}$ up to time t. It is flexible in that it admits many different model types, including Gaussian processes, support vector machines or feed-forward neural networks, although models which can explicitly use temporal structure of data such as Gaussian process dynamic models (GPDM; Wang, Fleet, and Hertzmann, 2008) or recurrent neural networks (RNN) tend to perform better. Recently, sequential counterparts of VAEs have been proposed, which allow efficient generative modelling of sequences, with the additional advantage of better regularisation and superior uncertainty estimates (Fabius and Amersfoort, 2015; Bayer and Osendorfer, 2015; Karl et al., 2017; Fortunato, Blundell, and Vinyals, 2017). Unlike deterministic RNNs, sequential VAEs model time-series in terms of low-dimensional latent variables that can be used in downstream tasks. Moreover, they allow counter-factual simulation and generation of multiple trajectories from a single starting point due to their stochastic nature. The most relevant prior work is that on state estimation and next-frame prediction in videos. Ondruska and Posner, 2016 introduced Deep Tracking, which aims to estimate state in a partially-observable environment. It uses two-dimensional occupancy grids and predicts grid occupancy in the future. While related, it is unclear how this approach could be used to model any other type of data. Next-frame prediction has been done recently as predictive coding (Lotter, Kreiman, and Cox, 2016; Canziani and Culurciello, 2017). The idea dates back to the Kalman filter (Kalman, 1960) and states that the hidden state of the model should be updated only to remove any discrepancies between the predictions and the observations at the following time-steps. While very general, this approach imposes additional structure on the prediction problem: it (i) removes redundancies found in consecutive inputs, (ii) creates an inner feedback loop, which could adjust model dynamics at runtime to minimise any errors and (iii) could implement human-like attention mechanisms if realised probabilistically (Friston, 2009).

C. Model Structure as Prior Information

As the majority of neural models are over-parametrised (Denil et al., 2013), learning abstract notions from data can be extremely sample inefficient. Eslami et al., 2016 introduced Attend, Infer, Repeat (AIR), a VAE with a variable-length latent encoding for image reconstruction. This model imposes a geometric prior on the encoding length which encourages sparse solutions, therefore learning to decompose the scene into a number of independent parts — the objects. It is worth noting that, along the main model, the authors introduce difference-AIR, which exploits the specific structure of the problem and adheres to the predictive coding paradigm, thereby achieving better performance. In the extension of this work, Danilo Jimenez Rezende et al., 2016 learn to reconstruct three-dimensional (3D) structure of an object from even a single two-dimensional (2D) view by imposing 3D latent representation and structuring the decoder as a projection of the latent space into the 2D output space; they show that their model is able to infer the idea of an object from data. Häusser, Mordvintsev, and Cremers, 2017 learn the idea of an object and its class by learning to associate similar objects with each other in the embedding space, which is very much like a child learning about its identity by comparing itself with others (Decety and Chaminade, 2003). In case of reinforcement learning, a complex environment might itself be a cue which leads to learning abstract ideas. Heess et al., 2017 shows that articulated agents can learn real-world motion patterns by interacting with the environment. Specifically, they learn to crouch, jump, turn and run while maximising a very simple reward function based on forward progress. Using a specific model structure as a method of learning abstract ideas was also demonstrated by Battaglia et al., 2016. The authors propose an interaction network, a highly complex model that operates on a graph of objects and relations between. Their application is to simulate physical systems under full-observability, but additionally, the model structure enables learning invariants (e. g., energy conservation) and inferring latent variables describing the system as a whole (e. g., potential energy).

D. Generative Modelling for Reinforcement Learning

Model-free RL is data hungry and improving sample efficiency of model-free methods is a long-standing research problem. Sutton, 1991 introduced the Dyna architecture, which uses and jointly trains a model-free parametric policy and a generative model of the environment. The former allows efficient inference and optimal performance, the latter reduces number of samples required from the environment by providing model-based simulations. Despite the theoretical advantages, it has been very difficult in practice to implement Dyna for anything but the simplest RL problems due to instabilities introduced by the learning of the model (Gu et al., 2016). Nagabandi et al., 2017 managed to overcome this issue recently and used a mid-sized neural network as the model.

The Dyna framework has neuroscientific grounding. It has been hypothesised that the parametric models of neocortex admit efficient inference but require long time to train. According to Kumaran, Hassabis, and McClelland, 2016, this issue can be mitigated by hippocampus, which can quickly store experiences and either replay or simulate them during sleep. In this sense, simulations in Dyna are very similar to the experience-replay mechanism, which has been shown to stabilise and improve convergence of large-scale model-free RL models (Mnih et al., 2015).

Dyna is not the only approach based on generative modelling. On the contrary, RL based on control in latent spaces has been quite successful. Watter et al., 2015 introduced Embed to Control (E2C), a stochastic locally-optimal control framework, which uses VAEs for learning of the latent-space for control. It approximates the latent-space dynamics by a locally-linear transition, which has controls as one of the inputs. The VAE manages to recover the true latent variables describing the state of the environment, which results in good long-term prediction performance. This type of long-term imagination could be used for multi-step roll-outs of model-based simulations in Dyna.

III. Submitted Work

During my first year as a DPhil student we developed the Hierarchical Attentive Recurrent Tracking (HART) framework, which was submitted to NIPS 2017. This RNN-based model learns to track objects in videos by focusing on small image regions. It does so by using a differentiable attention mechanism, which can effectively crop a part of the image, thereby quickly removing irrelevant parts of the input. Upscaling HART to a challenging real-world dataset proved difficult, as end-to-end training on a randomly initialised neural network was very unstable and converged to poor results. To address this issue, we resorted to transfer learning and used AlexNet (A. Krizhevsky, I. Sutskever, and Hinton, 2012) as a feature extractor, which has stabilised the training and improved performance (*cf.* section 5.2. in the paper).

The task of object tracking is fully-supervised, but can be seen as a reinforcement learning problem (Zhang et al., 2017) with a continuous action space, where a policy chooses a bounding-box update at every time-step; the agent receives a reward either at every time-step or at the end of the episode and the reward structure can be chosen based on the distance between the ground-truth bounding-box and the model estimate. In this setup, HART can be seen as a model-free policy. Instead of using a pre-trained feature extractor, it would be possible to utilise a model of the environment to perform off-line training of the policy, similarly to the Dyna framework. If the model is structured and provides correct position estimates of the object, this approach could increase performance of the tracking framework via unlimited model-based data augmentation.

Alternatively, if a generative latent-variable model of image sequences is available, HART could use the latent representation as extracted features, without the need to rely on a feature extractor pre-trained on static image analysis. Even though static image analysis has different characteristics than sequential analysis (e. g., data redundancy at consecutive time-steps), image classification models are often used for processing of video (see e. g., Ning et al., 2016). This approach, while effective, has little justification in neuroscience. In contrary, there is a growing body of evidence indicating the importance of temporal connections in the human visual cortex (Kastner and Ungerleider, 2000), which suggests that the temporal integration of information is vital for building up high resolution representation of the world, and is also confirmed by the empirical results of predictive coding approaches, *cf.* section II-B.

Our work on HART resulted in a biologically-inspired algorithm, which advanced the state-of-the-art performance in attentive recurrent tracking. Contrary to modern trackers, it does not use heuristics to update the scale estimate of the tracked object or to choose the search region in the new frame (Bertinetto et al., 2016; Held, Thrun, and Savarese, 2016). It is efficient thanks to the attention mechanism and end-to-end trainable. Finally, it has taught us about learning in the presence of temporal dependencies and structured modelling.

IV. Research Proposal

During the next year, we will develop a series of structured generative models of videos. Firstly, we are going to leverage recent advances in variational inference and neural networks to build a generative model of moving objects as an extension to the AIR framework. Secondly, we are going to improve the AIR model to work on images with rich background and real-world data and then extend this modification to the generative model of moving objects. Finally, we are going to investigate using trained generative models of videos within the Dyna framework. We now detail the above steps.

A. A Generative Model of Moving Objects

While AIR reconstructs an image by detecting objects present therein and painting one object at a time in a blank canvas, the generative model of moving objects (GMMO?) extends AIR to track objects by generating them one-at-a-time in a sequence of blank canvases. To reconstruct an image, AIR decomposes it into a set of $\mathbf{z}^{\text{where}}$ and $\mathbf{z}^{\text{what}}$ latent variables, which describe location and appearance of an object, respectively. The sequential model will need to take time-dependencies into account. In particular, instead of directly using $\mathbf{z}^{\text{where}}$ and $\mathbf{z}^{\text{what}}$ inferred from an image $\mathbf{x}_t$ at time t to reconstruct the image at time $t + 1$, it will need to take into account the history of appearances and locations $\mathbf{z}_{1:t}$ at times 1 to t. This can be accomplished by using a dynamics model, e. g., an RNN.

Even though the modification to AIR looks simple, it is unclear whether this approach will work. Firstly. it is based on the assumption (like AIR) that the correlation between pixels within an object is much stronger than correlation between pixels inside and outside of the object. Secondly, this model is not allowed to peek at the image at time $t + 1$ to reconstruct it, which severely increases the difficulty of the task. To address this issues, we are going to start simple, with a toy dataset of moving two-dimensional shapes. We will extend it later to moving three-dimensional shapes in the presence of camera motion.

In the absence of data, the model allows simulation by updating the latent state $\mathbf{z}_t$ with samples drawn from a prior distribution $p(\mathbf{z})$. Choosing the right prior for a sequential task poses a research question by itself (Sölch et al., 2016) and might require significant effort to answer. The transition function of the dynamics model is another crucial component of the model. It defines dynamics in the latent space and it will determine whether the model adheres to the laws of physics. We expect this stage to take about two to four months.

B. Generalisation of the AIR framework

In order for the model to be useful in any real-world setting, it has to be able to handle video sequences with rich backgrounds and occluded objects. We expect that this will require a form of object/background segmentation or background subtraction and generative blending of objects and the background. Given that AIR uses a spatial transformer (Jaderberg et al., 2015) to draw objects in a canvas, it is straightforward to create an explanation mask, which marks which locations in the canvas has been drawn to. When objects are explained, it should be possible to use the explanation mask with a complementary background model to explain the remainder of the image. Separating reconstruction of the background and the objects might create discontinuities at the boundaries, however, and it is unclear how to prevent the background model from explaining the objects at the same time. It is our intuition that pixel correlations within objects are different than in the background or between objects and their neighbourhoods. If we parametrise background- and object-generating models with a minimum-length encoding scheme, it should force them to learn their problem-specific correlation structure, therefore forcing the parts of the scene to be explained by corresponding model components. Since the KL-divergence term in the VAE loss can be interpreted as an information-bottleneck (Achille and Soatto, 2016), VAE effectively minimises encoding-length of the latent representation.

We will start by working with a multi-MNIST dataset, similar to the one used by the original paper, but with a noisy background. The goal is to upscale the approach to real-world images and e. g., ImageNet dataset. We expect this phase to take between 4 and 6 months.

C. A Generative Model of Videos

Combining the generalised version of AIR with a generative model of moving objects will result in a structured generative model of image sequences. While simple in principle, we expect a number of issues to arise. Firstly, the moving object model does not take the background of the target image into account. We might have to modify the model to predict the background and use it to condition the locations of the objects in the target image; alternatively we can condition background generation on the object appearance and their location. Secondly, training a high-fidelity model on video sequences is computationally very expensive due to the huge amounts of data this approach requires. Additionally, it is unclear which output probability distribution to use; output probability distribution in VAEs is responsible for the shape of the loss landscape. Gaussian assumption about prediction errors is not well justified when dealing with images and temporal dependencies between model outputs aggravate this issue even further. This issue might require further research (Generative Adversarial Networks might hold an answer; Wenzhe et al., 2016) if satisfactory performance is to be attained. We expect this stage to take 4 to 6 months.

D. Model-based RL

A good generative model of videos can be used for improving sample efficiency within the Dyna framework. Given difficulties with training non-linear models of environment, however, it is unclear whether this approach will work. A generative model of videos with variable-length encoding factorised between parts of the scene can be very useful in latent-space control algorithms similar to E2C, especially when any form of relational reasoning are required. In this case, the object-based representation deliver by AIR-like modelling can be married with structured reasoning models such as relational nets of Santoro et al., 2017 or the dynamic neural computer of Graves et al., 2016. We expect the final stage to take the remainder of this DPhil.

V. Conclusions

This paper summarises the contributions I have made during my DPhil studies so far and details my future research plan. For the remainder of my studies we would like to explore representation learning for sequential data, with the goal of using developed techniques in model-based reinforcement learning.

References

A. Krizhevsky, I. Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *NIPS*, pp. 1097–1105.

Achille, Alessandro and Stefano Soatto (2016). "Information Dropout: Learning Optimal Representations Through Noisy Computation". In: *arXiv*, pp. 1–11. arXiv: 1611.01353.

Battaglia, Peter et al. (2016). "Interaction Networks for Learning about Objects, Relations and Physics". In: *Nips*, pp. 4502–4510. arXiv: 1612.00222.

Bayer, Justin and Christian Osendorfer (2015). "Learning Stochastic Recurrent Networks". In: *ICLR*. arXiv: 1411.7610.

Bertinetto, Luca et al. (2016). "Fully-Convolutional Siamese Networks for Object Tracking". In: *ArXiv*. Springer, pp. 850–865. arXiv: 1606.09549.

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Springer, p. 738.

Canziani, Alfredo and Eugenio Culurciello (2017). "CortexNet: a Generic Network Family for Robust Visual Temporal Representations". In: arXiv: 1706.02735.

Decety, Jean and Thierry Chaminade (2003). "When the self represents the other: A new cognitive neuroscience view on psychological identification". In: *Consciousness and Cognition*. Vol. 12. 4, pp. 577–596.

Denil, Misha et al. (2013). "Predicting Parameters in Deep Learning". In: *NIPS*, pp. 2148–2156. arXiv: 1306.0543.

Eslami, S. M. Ali et al. (2016). "Attend, Infer, Repeat: Fast Scene Understanding with Generative Models". In: *NIPS*. arXiv: 1603.08575.

Fabius, Otto and Joost R. van Amersfoort (2015). "Variational Recurrent Auto-Encoders". In: *Iclr 2013*, pp. 1–5. arXiv: 1412.6581.

Fortunato, Meire, Charles Blundell, and Oriol Vinyals (2017). "Bayesian Recurrent Neural Networks". In: arXiv: 1704.02798.

Friston, Karl (2009). "The free-energy principle: a rough guide to the brain?" In: *Trends in Cognitive Sciences* 13.7, pp. 293–301.

Graves, Alex et al. (2016). "Hybrid computing using a neural network with dynamic external memory". In: *Nature* 538.7626, pp. 471–476.

Gu, Shixiang et al. (2016). "Continuous Deep Q-Learning with Model-based Acceleration". In: arXiv: 1603.00748.

Häusser, Philip, Alexander Mordvintsev, and Daniel Cremers (2017). "Learning by Association - A versatile semi-supervised training method for neural networks". In: *CVPR*. arXiv: 1706.00909.

Heess, Nicolas et al. (2017). "Emergence of Locomotion Behaviours in Rich Environments". In: arXiv: 1707.02286.

Held, David, Sebastian Thrun, and Silvio Savarese (2016). "Learning to track at 100 FPS with deep regression networks". In: *European Conference on Computer Vision Workshop*. Vol. 9905 LNCS. Springer, pp. 749–765. arXiv: 1604.01802.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh (2006). "A Fast Learning Algorithm for Deep Belief Nets". In: *Neural Computation* 18.7, pp. 1527–1554.

Jaderberg, Max et al. (2015). "Spatial Transformer Networks". In: *Nips*, pp. 1–14. arXiv: arXiv:1506.02025v1.

Kalman, R E (1960). "New Approach to Linear Filtering and Prediction Problems". In: *Fluids Engineering* 82.82 (Series D), 35–45 (1960) (11 pages).

Karl, Maximilian et al. (2017). "Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data". In: *ICLR*. arXiv: 1605.06432.

Kastner, Sabine and Leslie G. Ungerleider (2000). "Mechanisms of visual attention in the human cortex". In: *Annual Reviews of Neuroscience* 23.1, pp. 315–341.

Kingma, Diederik P and Max Welling (2014). "Auto-Encoding Variational Bayes". In: *ICLR*. arXiv: 1312.6114.

Kumaran, Dharshan, Demis Hassabis, and James L. McClelland (2016). "What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated". In: *Trends in Cognitive Sciences* 20.7, pp. 512–534.

Längkvist, Martin, Lars Karlsson, and Amy Loutfi (2014). “A review of unsupervised feature learning and deep learning for time-series modeling”. In: *Pattern Recognition Letters* 42, pp. 11–24.

Lapedes, AS and RM Farber (1988). “How neural nets work”. In: *NIPS*.

Lotter, William, Gabriel Kreiman, and David Cox (2016). “Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning”. In: arXiv: `1605.08104`.

Mnih, Volodymyr et al. (2015). “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540, pp. 529–533.

Nagabandi, Anusha et al. (2017). “Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning”. In: arXiv: `1708.02596`.

Ning, Guanghan et al. (2016). “Spatially Supervised Recurrent Convolutional Neural Networks for Visual Object Tracking”. In: *arXiv preprint arXiv:1607.05781*. arXiv: `1607.05781`.

Ondruska, Peter and Ingmar Posner (2016). “Deep Tracking: Seeing Beyond Seeing Using Recurrent Neural Networks”. In: *AAAI*, pp. 3361–3367. arXiv: `1602.00991`.

Pan, Sinno Jialin and Qiang Yang (2010). *A survey on transfer learning*. arXiv: `PAI`.

Rezende, Danilo J, Shakir Mohamed, and Daan Wierstra (2014). “Stochastic backpropagation and approximate inference in deep generative models”. In: *ICML*. Vol. 32, pp. 1278–1286. arXiv: `arXiv:1401.4082v3`.

Rezende, Danilo Jimenez et al. (2016). “Unsupervised Learning of 3D Structure from Images”. In: *NIPS*.

Santoro, Adam et al. (2017). “A simple neural network module for relational reasoning”. In: *Arxiv*, pp. 1–16. arXiv: `1706.01427`.

Sölch, Maximilian et al. (2016). “Variational Inference for On-line Anomaly Detection in High-Dimensional Time Series”. In: *ICML*. arXiv: `1602.07109`.

Sutton, Richard S. (1991). “Dyna, an integrated architecture for learning, planning, and reacting”. In: *ACM SIGART Bulletin* 2.4, pp. 160–163. arXiv: `arXiv:1011.1669v3`.

Wang, Jack M., David J. Fleet, and Aaron Hertzmann (2008). “Gaussian process dynamical models for human motion”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2, pp. 283–298.

Watter, Manuel et al. (2015). “Embed to Control: A Locally Linear Latent Dynamics Model for Control from Raw Images”. In: *NIPS*. arXiv: `1506.07365`.

Wenzhe, Shi et al. (2016). “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network”. In: *CVPR*, pp. 1874–1883. arXiv: `1609.05158`.

Zhang, Da et al. (2017). “Deep Reinforcement Learning for Visual Object Tracking in Videos”. In: arXiv: 1701.08936.