What are we automating? On the need for vision and expertise when deploying AI systems

Alexander Rast[*‡], Vivek Singh[*†], Steve Plunkett[†], Andrew Crean[†], and Fabio Cuzzolin[*]

Abstract: *Implementing in-house AI in the modern business is a classic example of digital transformation, often appearing simple and attractive, particularly given the emergence and availability of powerful, easy-to-use frameworks like TensorFlow or PyTorch. Such AIs are commonly considered for replacing cumbersome manual or physical systems, where neural networks may appear to be almost a panacæa automation solution to solve scalability or diversification concerns. However, such systems have subtle and sometimes very surprising behaviours that require considerable domain expertise, in order to implement a functional system without expending more effort than the system ultimately gains. Fundamentally, they need to be deployed with a clear sense of what the AI system is going to achieve. Careful attention must be paid at the outset to draft a clear and concrete design specification that indicates the intended function, and equally, draws a line under capabilities that are out of scope. Likewise, an effort needs to be made either to identify in-house people with the required skill sets to develop the system, or alternatively to enter into close working partnerships with external providers who can identify the needs and clearly articulate an appropriate solution. Most challenging of all, especially at large scale, is the emerging 'data gap' - the need to have access to or generate enormous volumes of labelled data - which often comes only at costs outside the budget of all but the largest companies. A case study in design collaboration between an emerging company transitioning from a physical to a virtual technology, and a university research group with substantial expertise in AI systems is presented, both as an illustration of the complex design considerations and a model for how to build in-house expertise. The collaboration is ongoing and outcomes are still preliminary, but the company is now starting to gain an appreciation for the complexity of real-world AI deployments and has developed a strategic plan that enables future growth. The emerging overall message is that modern AI is more an exercise in data automation than process automation.*

Introduction

The last 2 decades in business computing have seen the extraordinary development of Artificial Intelligence (AI) from a set of specialised techniques for niche applications to a mainstream set of tools. Deploying Artificial Intelligence (AI) solutions in a business context has become extremely fashionable, but can easily be done without a critical appraisal of what the underlying use case is. *What is AI going to do* that cannot be achieved just as effectively by more mature (and arguably more transparent) technologies? A 'top-down' answer is typically not enough; it is, for instance, insufficient to say 'this AI will allow us to discover more profitable trading patterns' or 'that AI will will enable a dynamic Web site with content tuned to the needs of a particular customer'. Discussion of end goals like this may motivate considering AI-based solutions, but they do not constitute the kind of formal specification that makes it possible to evaluate whether, much less what implementation of, an AI system, will realise its high-level goals [Calegari2020]. AI solutions tend to be narrowly domain-specific, data dependent, and sensitive to implementation details. The crucial point to be understood here is that AI systems are not guaranteed to solve anything.

* Visual AI Lab, Oxford Brookes University; † Supponor, Ltd. ‡ Corresponding Author

Because of this sensitivity, a business aiming to deploy AI successfully generally has to have, or develop, substantial in-house expertise even before beginning the design process.

This is a crucial step in digital transformation. Most modern artificial intelligence methods use powerful but extremely nuanced mathematical techniques, and the in-house team will need to be at least conversant in these mathematical algorithms (and their associated pitfalls) to be able to make rational design choices. The first step in the design of a working AI system is (or perhaps, should be) the enumeration of the mathematical, data, and application assumptions involved; and this will need to be referred to often in order to make sure such assumptions are not being violated (or taken for granted). Most prominent in modern systems in the chain of assumptions are those about the nature and distribution of the data [Marcus2020].

An empirical approach to data collection or algorithm development generally yields inconsistent results. Design by trial and error tends to produce apparently working systems that then fail spectacularly when they encounter conditions outside the range of what had been thought of or tested *ab initio* [Cai2020]. Data sensitivity is the most well-known of these pathologies; even large companies have fallen afoul of this in situations that lead to serious PR embarrassments [Guardian2018]. Scaling is another major source of problems; it cannot be assumed that a system developed at prototype scale with a limited dataset will work at all, much less well, when scaled to a production environment [Brigato2020]. The converse is true as well; attempts to scale down large systems to a streamlined, efficient solution do not always succeed; and the results are often problem-dependent [Passalis2018]. Another emerging issue is cross-domain transfer and generalisability; it is almost hypnotically alluring to attempt to apply a working AI solution developed for one domain to another, which may even appear to be relatively similar - e.g. to transfer the knowledge from a music recommendation system to one designed to recommend books. But again, results have often been uneven [MLong2015], [Tan2017], [Ramirez2019]. The pattern that seems to emerge is that (well-designed) AI systems can transfer knowledge relatively well with *very* large systems and datasets, but smaller systems are less effective at such transfer [Hoefler2021]. Quite aside from the fact that the largest-scale systems are out of scope for all but the largest, most well-resourced firms, a dilemma emerges: expend the resources, manpower, and development time to create one large 'omnibus' AI, or develop a set of specialised AIs for each specialised task. Hence we return to the problem noted at the outset: *what is AI going to do*?

It is worth listing the major challenges involved in creating and deploying a successful AI system:

- Application scoping
- Problem specification
- Data acquisition, annotation, and curation
- Platform selection
- Model selection
- Model tuning
- System validation

This study will examine a practical implementation of a complex AI system for a real business case - a firm engaged in real-time media delivery for sporting events, transitioning from an older, physically-based system to an AI-based approach. It will analyse the above challenges with a particular focus on the data acquisition, annotation, and curation problem (which turns out to dominate the considerations).

The General Background of AI

Before describing the specifics, it is useful to describe the general background of Artificial Intelligence to put solutions in context. As a computational discipline, the roots of practical AI extend back to the late 1950s; the 'traditional' birth date being the Dartmouth Workshop of 1956 [McCarthy1955]. However, despite several start-stop waves of AI adoption involving various techniques, the modern era of AI did not truly start to take shape until the emergence and eventual dominance of neural network techniques, based largely on so-called 'deep' networks. What it means for a network to be 'deep' is somewhat vaguely defined, but it refers, in general, to a network with considerably more 'layers' (arrays of neurons) that either the 2 originally introduced in the Perceptron [Rosenblatt1958] (and later shown to be seriously limited by Minsky in 1969 [Minsky1969]) or the 3 shown to be sufficient to implement a universal function approximator [Hornik1991] . Neural networks are often represented diagrammatically as large-scale parallel systems (Fig. 1); this provides a convenient conceptual understanding, but in fact, most computers represent neural networks as linear algebra operations over very large matrices.

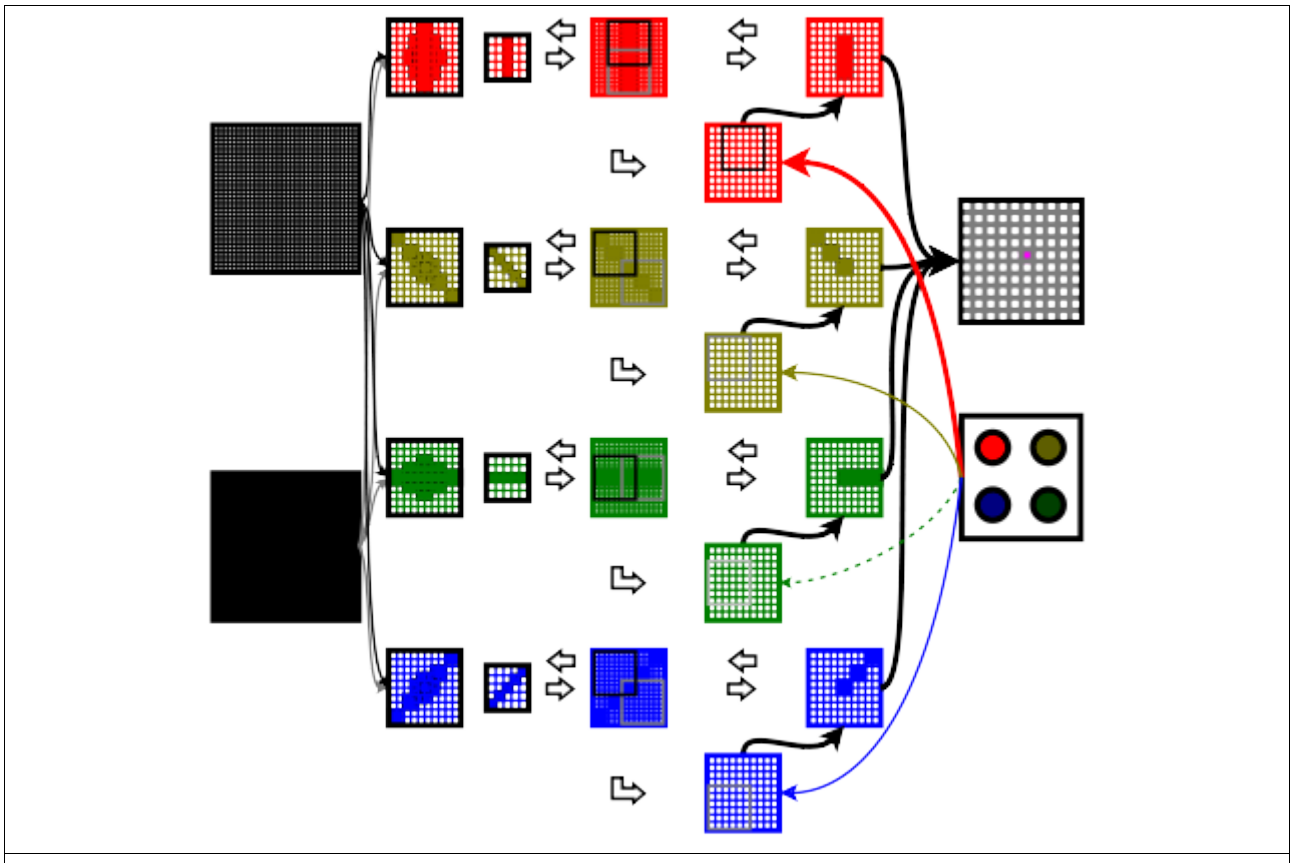


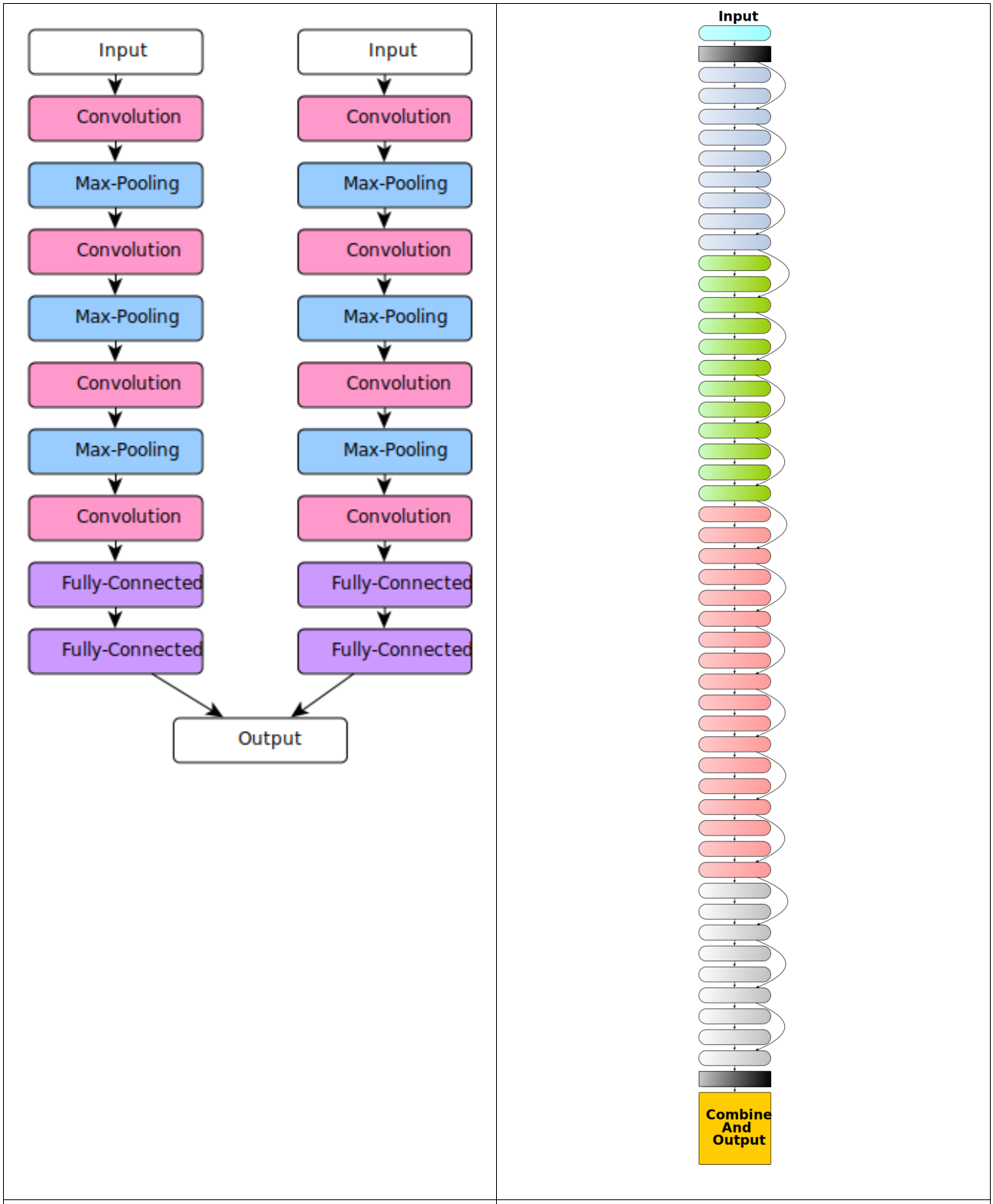
Figure 1: A 'typical' modern deep convolutional neural network (CNN). Layers alternate between match-to-template convolutional layers and grouping 'max-pooling' layers, ending in fully-connected layers to the 'heads'.

Although they have a 'different' parallelism than their diagrammatic counterparts, such operations are very naturally suited to parallel systems and have become routinely implemented on specialised hardware, typically, Graphics Processing Units (e.g. nVIDIA RTX series) (https://www.nvidia.com/en-us/design-visualization/ampere-architecture/),(https://www.nvidia.com/en-gb/design-visualization/rtx-a6000/) that efficiently compute matrix-vector products, the core of neural networks in modern implementations. It is

now understood that neural networks are an alternative formulation of Bayesian machine learning [Hoefler2021], which itself derives from the Bayesian reasoning systems that dominated AI in the early 2000s. Bayesian and other machine learning approaches still have significant applications and deployments in modern AI systems. In either case, similar challenges apply, but the deep neural network case can conveniently be used to illustrate many of the features of modern AI systems and indicate the main considerations.

Deep networks themselves have grown from relatively 'shallow' (in modern terms) networks involving 'only' 8-16 layers [Krizhevsky2012], to very large-scale systems with 100+ layers [He2016] [Justus2018] (Fig. 2). The most popular forms of modern neural network are Convolutional Neural Networks (CNNs) - which internally do a complex match-to-template with the data [LeCun1998]; and Transformers - which correlate data elements across sequences of data [Vaswani2017]. Many other models exist, each offering some tradeoff of different capabilities for different classes of problem, but CNNs, in particular, are widely used in machine vision applications [Jiao2019], [Abirami2021], which is completely rational considering their original inspiration in the visual cortex of humans and other mammals [Al-Aidroos2012]. It is not thought CNNs actually implement the processing the visual cortex does, but the structural similarities do suggest some affinity. Most CNNs for real applications can be split into 3 structures. The 'backbone' is the early layers of the network and is used for changing the input data representation into one the later network can use efficiently. Backbone networks are highly standard and go by names like AlexNet [Krizhevsky2012], VGG [Simonyan2015], or ResNet [He2016]. The 'neck' is the component that computes the statistical distributions and correlations that map input spaces to output spaces, e.g. 'features' to 'classes' or 'blocks' to 'regions'. Neck structures are typically tuned to suit the application, but remain general (and reusable) since their purpose is to produce universal representations. Typically a developer does not entirely hand-craft the neck but customises it from an off-the-shelf neck component. The 'head(s)' transform the general patterns represented in the neck to the specific output representations needed for the particular application, e.g. a set of labels such as 'ball', 'player', 'grass' etc. There may be several heads [Masaki2021] representing different label classes or properties - and these are entirely designed by the developer for the specific use. It should thus be seen that in essence, what CNNs are actually doing is nothing more than juggling the representation. The entire process of a CNN consists of finding an output representation that efficiently encodes the latent information desired from the input data.

What has led, arguably, to the widespread adoption of CNNs and other deep neural networks is the emergence of 'frameworks', integrated toolchains that link naturally with hardware platforms like nVIDIA and allow a high-level specification to be compiled directly down into code heavily optimised for the hardware. Frameworks not only reduce the design cycle by allowing networks to be assembled in 'building-block'-like fashion from a library of standard components, but also greatly improve the resultant performance by using powerful automated optimisations that replace what would otherwise be possibly months of hand-tuning. Popular frameworks include TensorFlow (https://www.tensorflow.org), PyTorch (https://pytorch.org), and Keras (https://keras.io) (names that may be familar to many readers). These tools allow rapid application development and may appear to make neural network development



a. A 'shallow' siamese CNN [Koch2015] | b. A 'deep' CNN (ResNet50) [He2016]

Figure 2: CNN development. The network on the left has only 11 layers; by contrast the network on the right has 50 layers (and larger versions of the same ResNet architecture have up to 152)

entirely straightforward, but feature significant quirks which must be dealt with and require considerable experience [Dai2022], (https://www.knowledgehut.com/blog/data-science/pytorch-vs-tensorflow). First, these frameworks are typically dependent on a very specific computer and operating system environment including particular versions of support libraries, hardware, and environment variables. Merely configuring a system to run any

framework for the first time can take up to a week (or more). Environment-manager software like Conda (https://conda.io) can, at least, manage the software dependencies reasonably, but hardware is another matter - and typically the user or developer must match their framework version and installation to the hardware they actually have on their system. Second, different framework versions can yield different performance for the same networks, and this means that an AI developed on one system does not necessarily migrate neatly to another; indeed, it is frequently the case that older versions or hardware might yield better performance than newer versions [Shahriari2022]. This, in turn, leads to a third quirk, that results may not be reproducible; one cannot, for example, rely on published benchmark studies to decide upon networks because the performance (or even functionality!) reported in such articles may depend upon a particular setup [Elshawi2021], and see e.g. (https://pytorch.org/docs/stable/notes/randomness.html). Overall, then, frameworks are an essential but temperamental part of modern AI development, rather like a Formula 1 race car: a high-performing vehicle, but one which requires a competent driver and continuous maintenance by an expert crew.

It has already been noted that CNNs are really doing nothing more than finding efficient data representations. For them to do so, they need data; a LOT of data. The problem is more subtle than it may initially seem; essentially, since large-scale neural networks contain many millions or billions of weights (commonly called 'parameters' in modern terminology), they may have the representational power to encode the *entire* dataset, if the dataset is small [Hoefler2021]. Therein lies the problem: if they do encode the entire dataset, the system is nothing more than a complex, opaque look-up table; the same results could have been had by simply storing the inputs as (input, class) pairs. This is useless; there is no 'intelligence' in this case because the network is not representing general properties of the system but simply mirroring the data, the well-known problem known as 'overfitting' [Shorten2019]. Overtraining happens whenever a suitably large network is presented with a suitably 'small' dataset for a long enough training time. For billions of weights, corresponding billions of data values need to be presented. But now this in turn creates a pair of related problems. First, training takes time, billions of data items may take weeks of time to process [Thompson2020], with enormous energy cost as well [Patterson2020], and this is likely entirely beyond the resources of most smaller firms, whilst being inefficient for larger firms (unless the result has unusually widespread application) [Brown2020]. Second and more critically, using *supervised* learning methods, a very substantial chunk of this data needs to be annotated data - marked up with 'correct' identifications, and annotation is generally a tedious, labour-intensive manual process [Hinterstoisser2019]. There are partial solutions from unsupervised or 'semi-supervised' learning methods [Ouali2020], [Li2022], but these are not usually complete, are typically less accurate, and involve even longer training times. Solving the data problem is now being discovered to be perhaps the crux issue in AI deployment, as indeed will be seen in the case study presented here.

Case Study: Background

The company to be considered: Supponor, Ltd. (http://www.supponor.com) is an emerging market leader in the field of targetted advertising provision. Specifically, Supponor operates at sporting events, to take LED billboards or other advertising placement points on the field or venue of play, and substitute the locally-visible content for content more suitably targetted to the regions or countries where the event is being broadcast live via television. The problem is very dynamic: Supponor's systems must be able to detect the regions of advertising content from the video stream, blank out these regions, and substitute different content, without accidentally blanking out critical video such as players, balls, etc. All of this must be done in real-time, at full frame rate, whilst considering problems of e.g. distortion in the image, altered aspect ratio due to camera angle, transient occlusions, weather, lighting contrast across

the scene, etc. Supponor initially entered into the market using proprietary physical technology directly installed on the field of play to be able to perform the real-time substitutions. However, the system was cumbersome, installation and setup time was significant, cost to the sporting organisation considerable, and the system represented a fixed capital investment with significant risks. As time went on it was clear there were further problems with diversification and expansion: a system installed for a given venue or sport did not easily transfer to different venues or sports; progressive changes within the sport either to play or to venue facilities could mean opportunities lost and/or costly changes to the installed system; the approach relied on long-term commitments from sporting organisations (generally, the leagues or associations in the relevant country for the relevant sport); the system was vulnerable to physical faults or disruptions; and perhaps most critically, the startup costs were more than all but the most well-resourced organisations (generally, the 'premier' leagues in big-market sports) could afford. In short, the fixed-installation nature of their existing technology prevented an agile business model. Supponor decided, therefore, to consider the possibility of full digital replacement, based purely on the video data as processed by an AI for video scene understanding. Their experience provides a good example of digital transformation in practice.

The company was put in contact with leading experts from Oxford Brookes University with a strong background in scene understanding; in particular, members of the Visual AI Lab (VAIL) (https://www.brookes.ac.uk/research/units/tde/groups/visual-artificial-intelligence-laboratory),
led by Prof. Fabio Cuzzolin. The VAIL team outlined a programme of work aimed at exploring the feasibility of using digital replacement approaches - technology that could not only substitute for the existing system but add additional capabilities, such as the ability to transfer directly to new sports or venues with little start-up time, or to overlay advertisements not just on the raw video, but potentially on that supplied by third-party broadcasters who have already overlain additional data layers (e.g. a running 'score ticker' at the bottom of the screen, etc.). The groups agreed to develop a Knowledge Transfer Partnership (KTP) with a dual purpose: on the one hand, to explore state-of-the-art AI video replacement solutions, and on the other to promote the development of the necessary in-house expertise in AI noted in the introduction to permit Supponor to continue forward with further developments. KTPs are a particular funding route supported by UK Research and Innovation (UKRI) (https://www.ukri.org/opportunity/knowledge-transfer-partnership/), the UK's national research funding body, to support close collaborations between industry and academia, particularly for de-risking exercises and/or development of in-house knowledge in leading-edge research at the margins of commercial viability. The project was eminently suited to this type of funding arrangement; work began in September, 2021.

Forming such partnerships, however, takes time, and in the period between the initial contact between Supponor and Brookes, and the start of the actual project, Supponor itself had already begun preliminary development of all-digital replacement technology. Much of this was in recognition of the clear limits of the original physical approach. But additional drivers included further developments in their target sports, particularly football, Supponor's initial primary area of focus (and in which they have by far the largest market penetration). There, the introduction of features like second rows of billboards and advertising 'carpets' placed directly on the pitch offered new opportunities that could not be exploited with the existing system. Furthermore, diversification into additional sports such as basketball (in the NBA) and hockey (the NHL), presented a complex rollout roadmap with considerable start-up time for organisations eager to go 'live' early and at a large scale. Only all-digital replacement could solve these problems, and so Supponor built its own internal technical development group and an initial all-digital system, borrowing heavily from the existing technology, with a plan to transition away from the physical installations as quickly as the AI-based technology could mature. The state-of-play at the beginning of the project was thus that Supponor had

what could be considered a prototype all-digital system (albeit in real deployments), but with significant limitations to its use or future potential.

The Supponor Experience

In spite of the extremely early nature of the new digital solution, Supponor's initial experience was reasonably positive. Deployments in football largely worked as a drop-in replacement for the physical technology with minimal start-up time; perhaps this was not surprising given Supponor's extensive experience and deployments in football. Viewer familiarity with the displayed effects also no doubt played a rôle; it is considerably easier to deploy replacement technologies with marginally different behaviour to an audience already familiar with the overall effect, than it is to introduce hitherto unseen technologies to completely inexperienced audiences not expecting significant changes to their existing experience.

An object lesson in how this can come into play was encountered in deployment to the NHL. Supponor itself was, in fact, well aware of the limitations and potential for teething problems and strongly recommended a cautious roll-out, but NHL management was eager to get the system live at full scale across the league early, and opted for a very aggressive roll-out plan. Given that there were no immediate technology concerns, just a *general* sense at Supponor that an ambitious deployment schedule would be inviting trouble, the groups proceeded with immediate roll-out. As things happened, the technical issues encountered revolved not around the AI components, which generally worked acceptably, but on integration issues such as video format, display resolution, hardware, etc. These generated a spike of technical support load for Supponor, but the larger problem was how these relatively minor issues affected the viewer experience. Although the roll-out was by and large successful, it produced something of an Internet backlash amongst die-hard NHL fans who felt that the resulting effects were too visually intrusive or noticeable [CBC2022]. In turn, this wave of outrage generated a group of fans using their own video tools to isolate and characterise particular artefacts in the video stream which would probably pass for unnoticed to the casual observer, but pointed out in this way, suddenly became very distracting. As a consequence, Supponor was bombarded by a wave of online criticism, arguably unfairly directed at them, because they had already been keenly aware of the limitations of the existing system. This demonstrates that it is not only important that a company have in-house expertise in AI, but also that they need to be able to communicate this effectively to their user base so that users do not end up with unrealistic expectations.

A level-headed analysis suggests that what was necessary in this situation was to temper expectations carefully for all parties. Undoubtedly, the NHL moved extremely aggressively and underestimated the strength of fan discontent, particularly with regard to heavily intrusive or distracting advertisements that quite literally drew attention to themselves. This is to be contrasted with the cautious plan adopted by the Bundesliga in Germany which requires extensive system validation *and* continuous analysis following each match to retain certification. Perhaps surprisingly, Supponor itself may have ended up being aided by the NHL experience, because fan response provided an unforeseen torrent of debugging information. In essence, the group of disgruntled fans provided free identification of artefacts at scale, without being so overwhelming in number as to colour the largely positive experience of the majority. Further disruptions are likely to decrease in scale as viewers become 'acclimatised' and the identified issues are ironed out. However, these transitions could have been made more smoothly if the rollout had been preceded with a period of pilot trial and perhaps focus groups amongst the fan base, lessons which apply equally to any would-be deployer of state-of-the-art AI solutions.

Teething pains aside, Supponor's immediate transition to AI technology has been surprisingly rapid and successful. At end of 2021 the company was just starting to transition

and their combined physical/virtual deployments amounted to 800, up from about 150 from the previous year. By contrast, at the end of 2022, the company foresees 3,000 deployments (matches/events covered), the greatest part of the increase coming from the NHL deployment of approximately 1,400 games (it should be noted that the scale of this rollout alone, compared to the comparatively conservative growth in football rollouts the previous year indicates the ambitiousness of the NHL schedule). However, the jump from 150 to 800 and then subsequently to 3,000 was entirely driven by the virtual technology - demonstrating how rapidly AI has overtaken the physical technology. Revenue growth likewise was strong, a gain of 140% between 2021-22 and 2022-23 financial years. The company has now started to work with Formula 1, who, following the more cautious approach recommended, have seen a successful pilot project form the basis of talks for a more long-term, wide-scale deployment. AI will be particularly important here because the international nature of F1 and the extremely unique, individual nature of each venue more-or-less necessitates off-site processing not tied to a physical installation - something the virtual solution enables which had previously been infeasible. The question then may be - what is left for the KTP to explore? Has not Supponor negotiated the learning curve in the digital transformation successfully and do they not as a result have the required in-house expertise already?

Crafting State-of-the-Art Solutions

At first glance, it may seem like Supponor has already solved most, if not all, of the major issues involved in transitioning to an AI virtualisation solution. However, the truth - hinted at in who Supponor is working with - is that this process is still relatively expensive and resource-intensive. The AI must, in essence, be hand-crafted for each new sport, and less potential has been observed for cross-domain transfer than might have been hoped, under the existing approach. Sports with relatively similar game play and venue setup like hockey and football, might, for example, offer hope that the system could be generalised to work with an arbitrary such 'players on a pitch' format, yet the systems themselves are individual for each. F1, meanwhile, presents an entirely new class of sport, with little expectation of direct transfer, and developing what is in essence a new system from the ground up requires considerable compute as well as human resources over many months. What is needed is an approach that can somehow generalise across sports, to the overall *class* of video infilling, and this involves moving from simple video segmentation (bounding different objects in a scene) to true video understanding (identifying the type of object bounded and being able to characterise - and predict - its behaviour). If then, the original aim of the KTP was to build expertise and develop a proof-of-concept, the goal has now changed: a proof-of-concept exists together with some in-house expertise, but what is wanted now is a more general system and a shift of design approach away from ad-hoc, hand-crafted AIs towards ones based on more universal, generalisable methods.

Originally the analysis focussed on so-called 'whole-scene' understanding. Supponor's existing virtual AI method trains only on cropped or masked patches of the original video stream - isolating the areas thought to be of interest and then implementing limited scene segmentation within those areas. This has strong similarities with the '2-stage' models [He2017] often used in perception systems for applications like autonomous driving: an initial model separates regions, and a second stage of processing segments within the region by applying class labels. However, recent research suggests such 2-stage models may be discarding global information across the scene that can inform segmentation, in other words, that can provide further conditioning on the posterior region probabilities. 'End-to-end' systems [JLong2015] produce segmented objects without any division into regions; this is particularly useful for the case of relatively slow-moving objects which, over a series of frames, smoothly change position in the visual field; many sports have this characteristic, and so an end-to-end system based on full semantic segmentation of the scene appears to make

sense (Fig. 3). It was also thought that this might allow better cross-domain transfer, as the network is learning general properties of segmentable shapes, rather than properties of objects of a specific expected size or shape within the visual field.

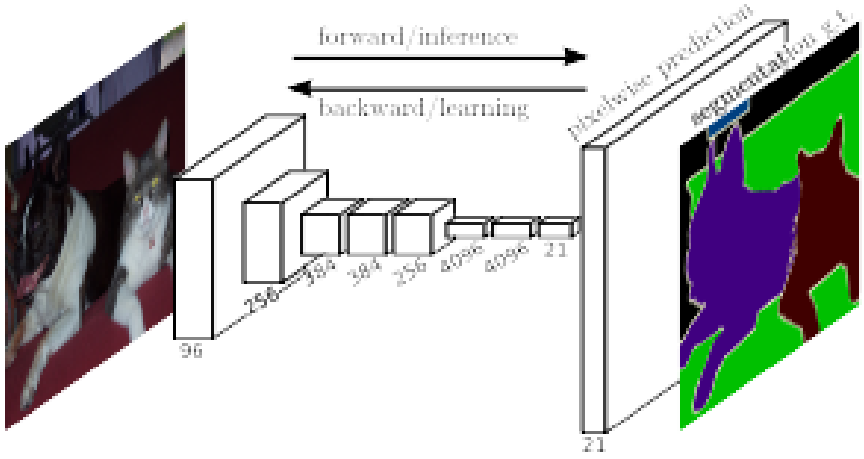


Figure 3: End-to-end convolutional network for semantic segmentation, from [JLong2015]

However, it became quickly apparent that traditional full scene segmentation faced a daunting barrier: the need for labelled data. For semantic segmentation to work, typical systems require training using large datasets annotated with the ground truth for both regions and labels. Thus, for example, a player on the pitch has to be given both an outline, and a label indicating this object is of class 'player'. It does not take long to realise that with complex, variable shapes like players, labelling video images will be a tedious and time-consuming process; many players in many postures and locations will have to be hand-annotated, the pixel boundaries may not be clear-cut, and if 2 objects intersect on the field it is not immediately obvious how they should be segmented, particularly if prediction is desired for the next frame. Existing training data was, for the most part, only partial frames, which makes sense in the case of Supponor's first-generation AI system, but would not be useful to develop a full-scene understanding model. For new generations of AI to be deployed, some method of generating labelled data rapidly, cheaply, and at scale would be essential, and it became very apparent that this challenge, indeed, would dominate the entire process of transitioning to a virtual AI system.

The Data Gap

The basic problem - the 'data gap' is starting to be understood throughout the AI industry as one of the most formidable ones to overcome. Generating large *labelled* datasets is both labour-intensive and time-consuming. There are firms that specialise in manual annotation of datasets, typically by outsourcing the annotation to a contract labour pool, e.g. Mindy (https://mindy-support.com/services-post/data-annotation-services), Qualitas Global (https://www.qualitasglobal.com), but for the scale of annotation that makes effective training possible, costs remain significant. Quotes for annotation of only 2000 frames of data for Supponor ranged from about $11,000 to about $27,000; and it was generally thought that this volume of labelled frames is not sufficient in itself (that is, without the use of other 'data augmentation' methods) to train a full-scene understanding system to a production standard.

This is a cost often overlooked in the deployment of AI systems. Research articles often quote impressive state-of-the-art performance figures for tasks like semantic segmentation [Chen2018], but these articles often conceal 2 hard truths. In the first instance, many articles are written focussing on established standard benchmark datasets. Such datasets have typically been assembled by large teams or consortia over many years and reflect a very large

prior investment [Lin2014]. Furthermore, standard benchmark datasets may be somewhat informative as to *abstract* performance capabilities of a system, but are typically not tuned to *specific* application requirements and are not suitable for training models for production applications. Some cross-domain transfer is possible, but results are often disappointing [Zhang2020], [Zhang2021]. In the second instance, state-of-the-art figures are often quoted in articles [Chen2020], [HWang2021] by large research teams working for the very largest firms in the field, e.g. Google, Amazon, or Microsoft, who are lavishly resourced in compute hardware, staffing, and access to data sources. Such a level of data access is not generally available to most firms. Indeed, many times access to data is blocked behind paywalls, as companies, increasingly aware of the value of data, understandably attempt to monetise their assets. Frequently, this is at rates that, while reflecting the labour involved in their creation, remain utterly inaccessible to small and medium-sized enterprises (SMEs). When one considers that even how much value the dataset would have to the customer company, may not be easily assessable without prior access to the data itself and some pilot trials, it becomes hard to justify what looks like a risky fixed investment in what could turn out to be a 'pig in a poke'.

For *specialised* domains such as Supponor's video-infill case, useful data may be unavailable even if the company has the resources to purchase it from an external source. Existing datasets, whether open-access or pay-for-play, tend to focus on general scene understanding for generic scenes or videos - the sort of application most useful e.g. to label photos on an Internet picture gallery or provide annotations for a film database. This may be useful to mass-market content providers, but is of less interest to domain specialist companies, especially those whose business model is primarily B2B focussing on the needs of client organisations rather than end consumers. A quick inspection made it clear that there was very, very little in the way of pre-existing datasets for the Supponor case. Hence there was a decision to be made: spend the time and money on extensive hand annotation, or look for alternative methods to bridge the data gap.

Solutions to the Data Gap

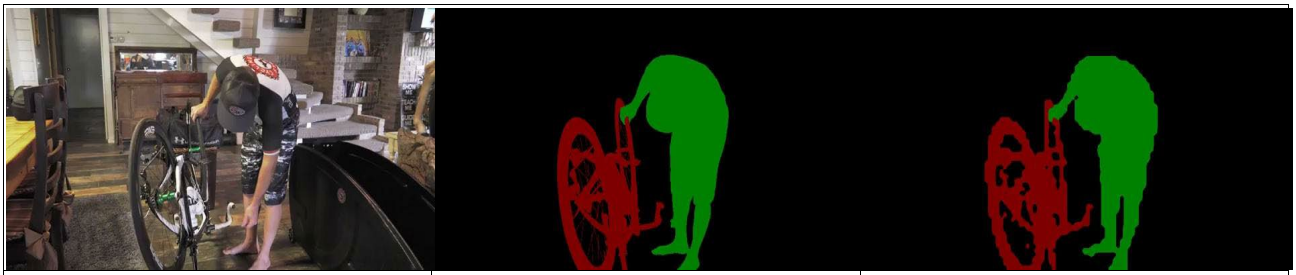
Efforts examined a variety of approaches. These include: using standard automated 'data augmentation' strategies [Shorten2019] to expand Supponor's existing labelled datasets; use of simulation to create complete 'virtual worlds' that come automatically available with ground truth information (because it is inherent to the simulation itself'); hybrid approaches combining both limited full-scene understanding with local scene segmentation similar to Supponor's existing system; unsupervised and 'semi-supervised' [Wang2019], [Chen2020] learning techniques that can bootstrap the learning to the point where a smaller labelled dataset is sufficient to fine-tune the system; enhanced annotation tools that permit faster and more automated manual annotation, and simply 'biting the bullet' and outsourcing a large dataset to an annotation firm.

Even the most cursory look at the costs involved in hand annotation - as seen above, quickly eliminated the full-annotation approach at the outset. Much more than $20K would not have been economically justified given uncertainty of outcomes - nor was it within the budget of the KTP. Data augmentation and hybrid strategies were also quickly eliminated, in part by early trials that performed poorly, but more realistically, because transferring a partial-scene dataset to full-scene dataset would be complex, error-prone, and dependent on prior assumptions. The simulation approach was looked at more seriously; the existence of game engines that, for example, include football, hockey, etc. hinted at the possibility of progress. As a future research direction for the general annotation problem, such simulation-based approaches look promising because, in principle, they can automatically generate arbitrary volumes of labelled data with almost infinite permutations [Hinterstoisser2019]. However, again, a realistic appraisal concluded that game engines and other simulation platforms are

complex systems with steep learning curves, the required expertise lay outside the skill sets of the team, and it was not at all immediately clear that off-the-shelf or even fully bespoke in-house simulation would be sufficiently similar to the real world to be useful for training. Simulation remains a very promising avenue for future exploration, but in the present is too dependent on personnel with matching skills.

The remaining options were to leverage special-purpose annotation tools and to explore unsupervised and partially-supervised methods. Supponor developed contacts with V7 Labs (https://www.v7labs.com), a firm offering a semi-automated annotation tool that promises dramatic reduction in annotation time. Although it was not yet definitely clear that it would reduce the costs to annotate significantly, it was reasoned that, given that the tool could potentially improve annotations for the *production* team working on Supponor's existing AI solution as well as for the *research* team in the KTP, a trial was justified. A quick series of tests followed and resulted in the following conclusions. First, it was found that indeed the annotation time dropped significantly - from hours per frame to around 45 minutes per frame. Second, the annotator could be used to edit frames automatically annotated using other AI methods, to produce accurate ground-truth data efficiently using a combination of automatic and manual methods. Finally, it was also concluded that whilst these results were effective and certainly justified the purchase of the V7 tool, they would still be inadequate for the extent of data required.

Almost inevitably, then, the group was looking at using unsupervised/semi-supervised methods. As it happened, such approaches were already built into the project, so these already looked like attractive options, but they were brought significantly forward in the project timeline by the pressing need to be able to use a minimum of hand-labelled data. Very recently, 'self-supervised' techniques which use information metrics to extract latent data have suddenly gained traction, and seen promising reported results in the literature [Lai2020], [NWang2021]. In part, this may be because the limitations of hand annotation and supervised learning are now becoming very evident to all research AI practitioners. Nevertheless, it is an emerging field, not one where many firms have existing in-house expertise, and indeed, it lies entirely outside the experience of the existing Supponor technical team. By contrast, the KTP partners at Brookes were already looking into these methods intensively and have some preliminary results, so it has proven a natural fit to extend these methods into the Supponor project, with the aim not only to improve the system, but to embed the knowledge and expertise brought to the in-house team and create a core capability extending Supponor's competitive advantage. Results (Fig. 4) thus far have been confined to benchmark datasets, but already suggest that a semi-supervised learning stage can automate annotations to about 65% accuracy - not enough, yet, for production-quality full-scene understanding, but enough to reduce the manual annotation requirements to a few thousand frames.



Original Video	Ground Truth Segmentation	Self-Supervised Prediction

Figure 4: Example of preliminary results from self-supervised learning on benchmark datasets. From presentation to KTP LMC4 meeting presentation, 29 November 2022

A staged path of development has been established with a network successively augmented by self-supervised pre-training, partially-supervised training, and fine-tuned training with full manual annotation using the V7 tools. (Fig. 5)

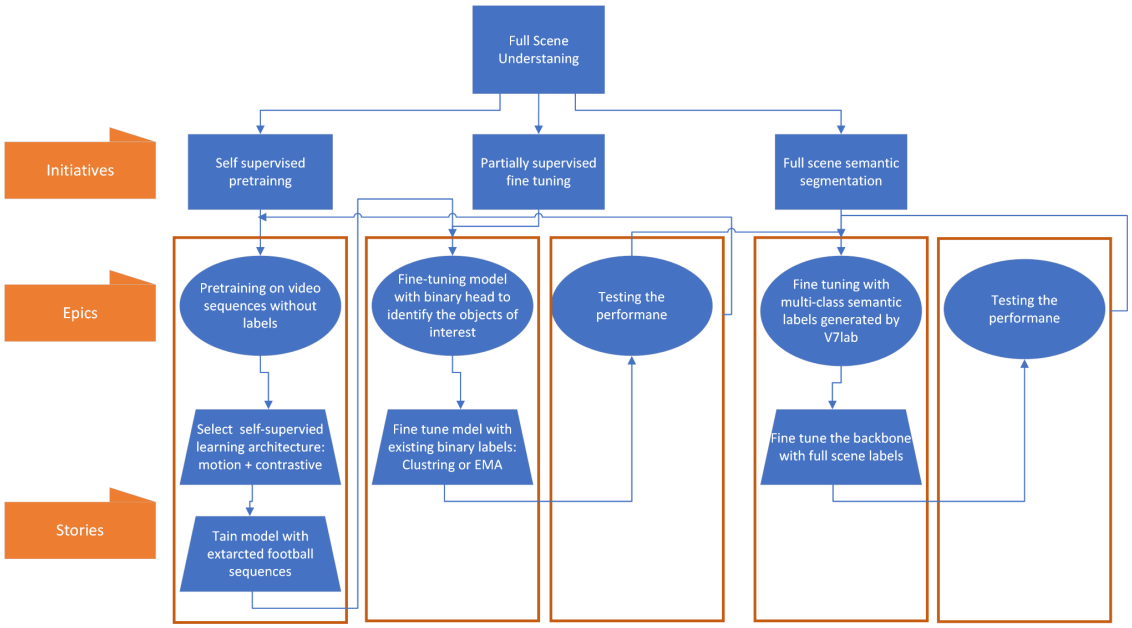


Figure 5: Development path. The system is being evolved from a self-supervised initial setup to a network ultimately implementing full-scene segmentation with a combination of annotated and unannotated data pretrained using self-supervised learning. From presentation to KTP LMC4, 29 November 2022.

It is still early to conclude exactly how effective this pipeline will be. It is, however, abundantly clear that this would not have been possible without the KTP partnership to transfer research expertise to the commercial environment.

Conclusions

The overall experience of both Supponor and Brookes in the KTP powerfully reinforces how critical it is for firms to understand fully the technical as well as business implications of automation using AI. In-house expertise is vital if initiatives are to be successful; without it, a company on the one hand is ill-prepared to handle the unexpected pitfalls of AI development, on the other unable to take advantage of recent developments in a field still rapidly advancing. Indeed, Supponor has quickly recognised the importance of this and as a result has begun forming an in-house *research* team (based in France) to supplement the *development* team (based in Finland). The 2 teams have complementary rôles; the development team is responsible for day-to-day deployments, bugfixes, and maintenance of the existing system, whilst the research team is responsible for taking recent advances from the academic research community and translating them into a strategic technology roadmap for the future. The Supponor-Brookes KTP, meanwhile, is a direct link into academic research at the edge of the state-of-the-art - an information conduit that gives Supponor access to skills and technologies beyond the commercial horizon. Where this leads to remains to be seen, but it seems clear that Supponor is now on a rapidly ascending trajectory with its wholehearted embrace of a fully-virtual AI content infill technology.

The entire process has also put the spotlight strongly on the data-generation problem. It appears that for AI development of the future, this will dominate research, over and above even the creation of new models. In the past AI was thought about mostly as automating the problem of data *processing*; the future appears headed towards data *creation*. Both the research and the software pipeline under investigation in the Supponor-Brookes KTP now looks more like a process to automate the generation of information using AI and machine learning methods, than it is to automate their interpretation. Perhaps fittingly, this is where Supponor started: the entire business is based on creative content infill; it now appears that

their AI systems of the future will be performing content infill on themselves. In the end, AI may work best when it embodies the intelligence built into the company deploying it.

References

[Abirami2021] R. N. Abirami, P. M. D. R. Vincent, K. Srinivasan, U. Tariq, and C.-Y. Chang, 'Deep CNN and Deep GAN in Computational Visual Perception-Driven Image Analysis', *Complexity*, vol. 2021, April 2020

[AlAidroos2012] N. Al-Aidroos, C. P. Said, and N. B. Turk-Browne, 'Top-down attention switches coupling between low-level and high-level areas of human visual cortex', *Proc. Nat. Acad. Sci. USA* vol. 109. no. 35, 4 Sep. 2012

[Brigato2020] L. Brigato and L. Iocchi, 'A Close Look at Deep Learning with Small Data', *Proc. 25th Int. Conf. Patt. Recog. (ICPR 2020)*, 2020

[Brown2020] T. Brown, et al. 'Language Models are Few-Shot Learners', *arXiv:2005.14165*, May 2020

[Calegari2020] R. Calegari, G. Ciatto, E. Denti, and A. Omicini, 'Logic-Based Technologies for Intelligent Systems: State of the Art and Perspectives', *Information*, vol. 11 no. 167, March 2020

[CBC2022] G. Nixon, 'The ads are virtual, but for some NHL fans, the irritation is real', *Canadian Broadcasting Company (CBC) News*, 15 October 2022

[Chen2018] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, 'Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation', *Proc. 2018 Euro. Conf. Comput. Vis. (ECCV 2018)*, 2018

[Chen2020] L.-C. Chen, R. G. Lopes, B. Cheng, M. D. Collins, E. D. Cubuk, B. Zoph, H. Adam, and J Shlens, 'Naive-Student: Leveraging Semi-Supervised Learning n Video Sequences for Urban Scene Segmentation', *Proc. 2020 Euro. Conf. Comput. Vis.* (ECCV 2020), 2020

[Dai2022] H. Dai, X. Peng, X. Shi, L. He, Q. Xiong, and H. Jin, 'Reveal training performance mystery between TensorFlow and PyTorch in the single GPU environment', *Science China: Information Sciences*, vol. 65, January 2022

[Elshawi2021] R. Elshawi, A. Wahab, A. Barnawi, and S. Sakr, 'DLBench: a comprehensive experimental evaluation of deep learning frameworks', *Cluster Computing*, vol. 24, 2021

[Guardian2018] 'Google's solution to accidental algorithmic racism: ban gorillas', T*he Guardian*, 12 January 2018

[He2016] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', *Proc. 2016 Int. Conf. Comput. Vis. Patt. Recog.* (CVPR 2016), 2016

[He2017] K. He, G. Gkioxari, P. Dollár, and R. Girshick, 'Mask R-CNN', *Proc. 2017 IEEE Int. Conf. Comput Vis. Patt. Recog.* (CVPR 2017), 2017

[Hinterstoisser2019] S. Hinterstoisser, O. Pauly, H. Heibel, M. Marek, and M. Bokeloh, 'An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Detection', *Proc. 2019 Int. Conf. Comput. Vis.* (ICCV 2019), 2019

[Hoefler2021] T Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden and A. Peste, 'Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks', *J. Mach. Learn. Res.*, vol. 23, September 2021

[Hornik1991] K. Hornik, 'Approximation capabilities of multilayer feedforward networks', *Neural Networks*, vol. 4 no. 2, 1991

[Koch2015] G. Koch, R. Zemel, and R. Salakhutdinov, 'Siamese Neural Networks for One-Shot Image Recognition', *Proc. 32nd Int. Conf. Mach. Learn.* (ICML 2015), 2015

[Krizhevsky2012] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks', *Adv. Neur. Inf. Process. Syst. 25* (NIPS 2012), 2012

[Jiao2019] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, 'A Survey of Deep Learning-based Object Detection', *IEEE Access*, vol. 7, September 2019

[Lai2020] Z. Lai, E. Lu, and W. Xie, 'MAST: A Memory-Augmented Self-Supervised Tracker', *Proc. 2020 IEEE Conf. Comput. Vis. Patt. Recog.* (CVPR 2020), 2020
[LeCun1998] Y. LeCun, Y. Bottou, Y. Bengio, and P. Haffner, 'Gradient-based learning applied to document recognition', *Proc. IEEE*, vol. 86 no. 11, 1998
[Li2022] L. Li, T. Zhou, W. Wang, L. Yang, J. Li, and Y. Yang, 'Locality-Aware Inter-and Intra-Video Reconstruction for Self-Supervised Correspondence Learning', *Proc. 2022 IEEE Conf. Comput. Vis. Patt. Recog.* (CVPR 2022), 2022
[Lin2014] T.-S. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, 'Microsoft COCO: Common Objects in Context', *Proc. 2014 Euro. Conf. Comput. Vis.* (ECCV 2014), 2014
[JLong2015] J. Long, E. Shelhamer, and T. Darrell, 'Fully Convolutional Networks for Semantic Segmentation', *Proc. 2015 Int. Conf. Comput. Vis. Patt Recog.* (CVPR 2015), 2015
{MLong2015] M. Long, Y. Cao, J. Wang and M. I. Jordan, 'Learning Transferable Features with Deep Adaptation Networks', *Proc. 32nd Int. Conf. Mach. Learn.* (ICML 2015), 2015
[McCarthy1955] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, 'Proposal for the Dartmouth summer research project on artificial intelligence, *Tech. Rep., Dartmouth College*, 1955
[Masaki2021] S. Masaki, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, 'Multi-Domain Semantic-Segmentation using Multi-Head Model', *Proc. 2021 IEEE Intell. Transp. Syst. Conf. (ITSC 2021)*, 2021
[Minsky1969] M. L. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*, MIT Press 1969
[Ouali2020] Y. Ouali, C. Hudelot, and M. Tami, 'Semi-Supervised Semantic Segmentation with Cross-Consistency Training', Proc. 2020 Conf. Comput. Vis. Patt. Recog. (CVPR 2020), 2020
[Passalis2018] N. Passalis and A. Tefas, 'Learning Deep Representations with Probabilistic Knowledge Transfer', *Proc. 2018 Euro. Conf. Comput. Vis.* (ECCV 2018), 2018
[Ramirez2019] P. Z. Ramirez, A. Tonioni, S. Salti, and L. Di Stefano, 'Learning Across Tasks and Domains', *Proc. 2019 Int. Conf. Comput. Vis.* (ICCV 2019), 2019
[Rosenblatt1958] F. Rosenblatt, 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psych. Review*, vol. 65 no. 6, 1958
[Shorten2019] C. Shorten and T. M. Koshgoftaar, 'A survey on Image Data Augmentation for Deep Learning', *J. Big. Data*, vol. 6, 2019
[Shahriari2022] M. Shahriari, R. Ramler, and L Fischer, 'How Do Deep-Learning Framework Versions Affect the Reproducibility of Neural Network Models?' *Mach. Learn. Knowl. Extr.* vol. 4, 2022
[Simonyan2015] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *Proc. 2015 Int. Conf. Learn. Represent.* (2015)
[Tan2017] B. Tan, Y. Zhang, S. J. Pan, and Q. Yang, 'Distant Domain Transfer Learning', *Proc 31st AAAI Conf. Artific. Intell.* (AAAI17), 2017
[Vaswani2017] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, 'Attention Is All You Need', *Adv. Neur. Inf. Proc. Syst. 31* (NIPS 2017), 2017
[Wang2019] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. Hoi, and H. Ling, 'Learning Unsupervised Video Object Segmentation through Visual Attention', *Proc. 2019 IEEE Conf. Comput. Vis. Patt. Recog.* (CVPR 2019), 2019
[HWang2021] H. Wang, X. Jiang, H. Ren, Y. Hu, and S. Bai, 'SwiftNet: Real-time Video Object Segmentation', *Proc. 2021 IEEE Conf. Comput. Vis. Patt. Recog.* (CVPR 2021), 2021
[NWang2021] N. Wang, W. Zhou and H. Li, 'Contrastive Transformation for Self-Supervised Correspondence Learning', *Proc. 35th AAAI Conf. Artific. Intell.* (AAAI-21), 2021
[Zhang2020] Y. Zhang and B. D. Davison, 'Impact of ImageNet Model Selection on Domain Adaptation', *Proc. 2020 IEEE Winter Applic. Comput. Vis. Wkshp.* (WACVW 2020), 2020

[Zhang2021] G. Zhang, H. Zhao, Y. Yu, and P. Poupart, 'Quantifying and Improving Transferability in Domain Generalization', *Adv. Neur. Info. Proc. Syst. 35* (NeurIPS 2021), 2021

About the Authors

Alexander Rast is a Senior Lecturer in Computer Science at Oxford Brookes University. He is a member of the Artificial Intelligence and Robotics research group and affiliated with the Visual AI Laboratory. He has close partnerships with several other research groups including the Autonomous Driving and Intelligent Transport group and the Institute for Ethical AI. His research interests focus on neural networks, with an emphasis on hardware implementations of spiking neural systems (neuromorphic chips). He is particularly concerned with embedded and robotic applications of neural networks (both spiking and nonspiking) and has worked on perception, visual attention, language grounding, and neuromorphic-robotic integration. Along with Fabio Cuzzolin he is one of the 2 academic supervisors for the Brookes-Supponor Knowledge Transfer Partnership.

Vivek Singh is the Knowledge Transfer Partnership (KTP) Associate for the Brookes-Supponor KTP. His background includes previous work with the Visual AI Laboratory on models for visual semantic segmentation in the surgical operating theatre under the SARAS project. His research focusses on scene understanding, semantic scene segmentation, object and action detection and recognition, and facial expression analysis.

Steve Plunkett is the Chief Product Officer for Supponor, Ltd. and has overall technical responsibility for Supponor's AIR digital-replacement technology. He is the KTP company supervisor for the project and has day-to-day responsibility for managing progress.

Andrew Crean is the Chief Financial Officer for Supponor, and is the Chairman of the KTP Local Management Committee (LMC), with overall oversight of the entire project. He sets strategic directions for the project, and ensures research objectives are matched with Supponor technical and market requirements.

Fabio Cuzzolin is a Professor of Computing at Oxford Brookes University and the director of the Visual AI Laboratory. He is the lead academic in the Brookes-Supponor Knowledge Transfer Partnership, and is currently PI for 8 funded projects with a total budget of £5,000,000. Professor Cuzzolin’s research focusses on artificial intelligence and its applications to computer vision and robotics. As founder in 2012 of OBU’s Visual Artificial Intelligence Laboratory he has been conducting work at the boundaries of computer vision. In particular the Lab is world-leading in deep learning for detecting and recognising human actions, as evidenced by some of the best detection accuracies to date and the first system (2017) able to detect multiple action instances in a streaming video in real time. The work has been highly cited in the field and has led publications in top computer vision journals such as IEEE PAMI and IJCV. Parts of the technology are being spun off as the foundations for a start-up company. The team’s focus has now shifted towards issues at the boundaries of visual AI, such as the modelling via deep learning of complex activities performed by multiple people and agents, the predicting of future events and intentions, and the creation of a machine theory of mind.