Where Research meets Industry: New Challenges and Opportunities at AImageLab

Rita Cucchiara[1,*], Lorenzo Baraldi[1], Simone Calderara[1], Marcella Cornia[1], Matteo Fabbri[2], Costantino Grana[1], Angelo Porrello[1] and Roberto Vezzani[1]

[1]AImageLab, University of Modena and Reggio Emilia, Modena, Italy
[2]GoatAI S.r.l.

Abstract

The application of Artificial Intelligence (AI) is becoming increasingly ubiquitous in industrial fields, posing some of the major issues and challenges of the future decades. In this respect, the old dichotomy *Industry vs. Research* appears tired: instead, a strong synergy between these two worlds represents a key aspect toward the most breakthrough incoming innovations. This manuscript presents an example of such a connection: namely, some of the activities carried out at the AImageLab laboratory, one of the most active research lab in Italy and Europe in the fields of Computer Vision and Artificial Intelligence.

Keywords

Artificial Intelligence, Innovation, Industry 5.0, Video Surveillance, Generative AI, Continual Learning

1. Introduction

Artificial Intelligence (AI) has emerged as the next breakthrough technology and is going to shape the innovation landscape of various fields, in ways not yet fully understood. It has been a long and hard journey (still ongoing) that dates back to the early 1950s (with the definition of the former idealized artificial neurons).

In this respect, the landscape has radically changed in the last two decades: the major research fields falling under the umbrella term *AI – i.e.* Deep Learning and Computer Vision – have crossed the chasm, switching from frontier academic topics to mature sources of operative and foundational business tools. Indeed, there is a disruptive change of perspective in how industry players look at AI: not just as a cutting-edge technology, but as a public utility on which to build services and applications, as already holds for electricity and water.

In such a context, the boundaries between academic research and the industry field have been blurring: to comply with the extremely low time-to-market required, the adoption of scientific results has to come about not over years, but in a few months or even weeks. To make it viable, the fundamental ingredient is a close *synergy* between several actors: namely, those primarily involved in research (even the most theoretical ones), those who produce goods and services, and potentially those intermediate actors facilitating the adoption of these technologies, to meet the challenges of society. In particular, the experts of the industrial domain have to work closely with AI engineers, researchers, data suppliers, and supervisors (more generally, who can support their continuous life-cycle comprising development, evaluation, and deployment). Besides the strict synergy, the definition of a *programmatic agenda* also plays a fundamental role, as it could define the deliverables and the perimeter of what is expected by AI models.

Such an epochal change has a worldwide dimension and has reached Italy, which retains a good tradition of AI research. In this manuscript, we are showcasing one successful case of research/industry integration, represented by the AImageLab laboratory, an internationally recognized research lab of the Department of Engineering "Enzo Ferrari" located within the University of Modena and Reggio Emilia. The research covers topics of Computer Vision, Pattern Recognition & Machine Learning, Natural Language Processing, Medical Imaging, Robot-Human Interaction, and Multimedia applied to optical images and videos as well as data from different sensors.

Since its foundation – which dated more than 25 years ago – AImageLab has joined activities for industrial research with many manufacturers and health companies, recently as a partner of the Artificial Intelligence Research and Innovation Center (AIRI) of the Modena Technopole. The projects carried out and the topics cov-

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
rita.cucchiara@unimore.it (R. Cucchiara); lorenzo.baraldi@unimore.it (L. Baraldi); simone.calderara@unimore.it (S. Calderara); marcella.cornia@unimore.it (M. Cornia); matteo.fabbri@goatai.it (M. Fabbri); costantino.grana@unimore.it (C. Grana); angelo.porrello@unimore.it (A. Porrello); roberto.vezzani@unimore.it (R. Vezzani)
0000-0002-2239-283X (R. Cucchiara); 0000-0001-5125-4957 (L. Baraldi); 0000-0001-9056-1538 (S. Calderara); 0000-0001-9640-9385 (M. Cornia); 0000-0001-6522-1988 (M. Fabbri); 0000-0002-4792-2358 (C. Grana); 0000-0002-9022-8484 (A. Porrello); 0000-0002-1046-6870 (R. Vezzani)

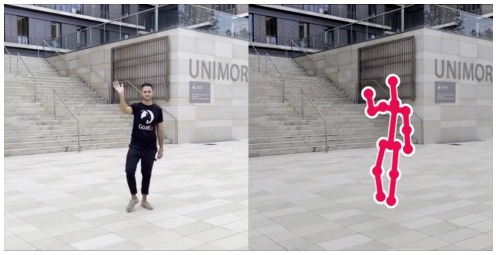

Figure 1: A frame captured by a common camera (left) compared to a frame captured by the GoatEye camera (right).

ered have achieved a strict connection with the economic fabric of the Emilia-Romagna region, tied principally to manufacturing production. However, in the last years, the AImageLab has been evolving to meet the growing demands of the service industry: more and more startups and established IT service companies are active in producing digital solutions for local industry and for export in Italy and the world.

Therefore, we kindly refer the reader to the next sections, each of which is presenting a cross-section of some of the industry-oriented research activities carried out within the AImageLab laboratory.

2. GoatEye: the Privacy Camera

Among the industry-oriented activities of AImageLab, the one carried out along with the startup GoatAI is with no doubt the more emblematic. GoatAI has been conceived within AImageLab and, as such, fully embraces a mindset where research and industry are strictly entwined. GoatAI purpose is to apply the most modern AI techniques for Human Behavior Understanding in health care, fitness, retail, safety and security.

As AI continues to advance, it is crucial that it is used in ways that respect individuals and their personal information. That's why GoatAI is on a mission to develop innovative solutions that not only provide benefits, but also prioritize and safeguard the privacy of those who use them. Nowadays and even more in the future, applying the wrong data privacy strategy can cost an organization billion in fees and damages.

Additionally, with the advent of the AI Act, several measures will be enforced to protect individual rights and privacy when processing sensitive data through AI models, intended to ensure that AI systems will be used in a responsible and ethical manner. The AI Act will have a significant impact on those developing AI systems, especially for Human Behavior Understanding.

That is why GoatAI wants to change the way people see cameras with GoatEye: the Privacy Camera. GoatEye wants to provide a new layer of abstraction by producing anonymized video material while maintaining all the non-

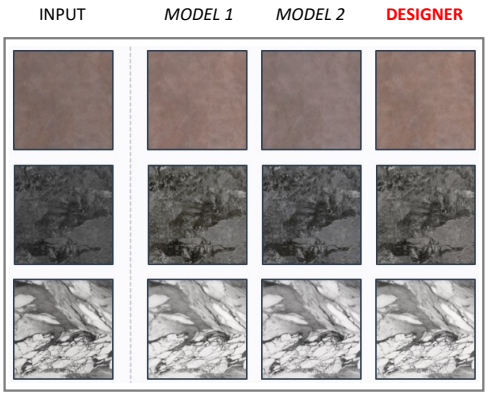

Figure 2: Qualitative results with 4-channel color space. The visual shows the RGB image, CONVERTED and ENHANCED image by designer, and the corresponding CONVERTED and ENHANCED versions generated by two models based on AI technologies.

sensitive information useful for data analytics purposes.

Featuring cutting-edge AI technology, GoatEye is able to deliver powerful video analysis capabilities while ensuring that user privacy is always protected. With GoatEye, customers enjoy all the benefits of AI without worrying about people privacy being compromised. It utilizes advanced algorithms to perceive humans in real-time, assessing position, posture, and even action performed with high accuracy. What sets GoatEye apart from the competition is its commitment to user privacy. It uses techniques to ensure that personal information is never collected or shared without users' consent. This means that customers are able to use GoatEye with confidence, knowing that the privacy of people is always respected.

3. AI for the Ceramic Industry

Image enhancement In the design industry, the automation of color management and image enhancement are challenging tasks, with several applications related to the creation and the print of design surfaces. As outlined in the next paragraphs, automated tools would speed up the work of graphic designers in the creation of novel ceramic tiles, with particular focus on the editing and the preparation of images fed to the printing pipeline.

Indeed, one of the most time-consuming activities of designers is to modify the image so that, once printed, its quality and level of details are consistent with what customers see on the monitor (or on paper). Such an operation – currently carried out manually through advanced image editing tools – is not only complex and time-consuming, but also hard to formalize, as it is based on the experience and sensitivity of designers.

The AImageLab team has recently proposed novel AI methods – based on Convolutional Neural Networks – to enhance the visual quality of industrial design surfaces,

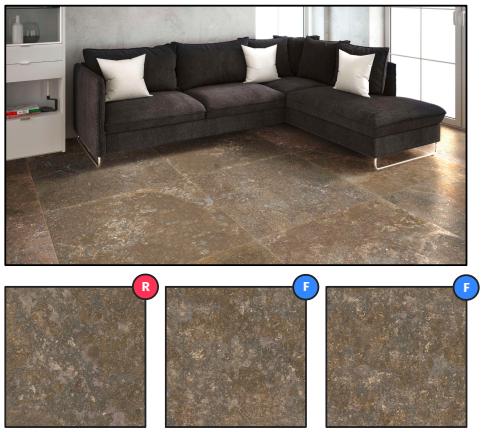

Figure 3: Given a single real tile (R), the system developed at AImageLab can generate many new high-resolution graphics (F) that can be used to decorate an entire floor surface.

according to specific properties dictated by color profiles (for a visual of the obtained results, see Fig. 2). These approaches were developed in collaboration with *Digital Design Srl*, a leading design company located in Fiorano Modenese (Modena) specialized in creating high quality design surfaces for different applications, in particular ceramic tile printing.

Generative design In order to provide further support to designers, a team of researchers at AImageLab developed a new approach to synthesize novel ceramic tiles. By generating new patterns and textures, the approach allows to generate the different tile faces that make up an entire ceramic surface. Such a case study is a practical example of how AI technologies can support the creation and evolution of new products, which represent the main goals of *generative design*. The latter is an important milestone in industrial production, especially for processes involving an important creative component.

To do so, the researchers have leveraged Generative Adversarial Networks (GANs): these models, by virtue of the extremely high-quality and high-fidelity images provided, have seen a significant growing interest for a variety of applications, ranging from super-resolution to image-to-image translation, from image inpainting to the field of image synthesis [1].

The idea of supporting designers and engineers by using a generative AI-based system opens new ways in the so-called Industry 5.0 paradigm, where humans and AI-based engines should collaborate for the industry of tomorrow. However, the proposed system has already been patented and is currently in use in the ceramic tile industry, whose designers, who supervise both the training and the generation process, are exploiting with considerable satisfaction the flexibility and creative capabilities of our generative model.

Figure 4: Given a front-view reference model and one or more try-on garments, the virtual try-on architecture developed at AImageLab can generate a new image where the input garments are virtually worn while preserving the body pose and identity of the reference model.

4. Generative AI for Fashion

With the advent of e-commerce, the variety and availability of online garments have become increasingly overwhelming for the final user. Consequently, user-oriented services and applications such as virtual try-on [2, 3, 4], customer-to-shop garment retrieval, and vision-and-language interactions [5, 6] are increasingly important for online shopping, helping fashion companies to tailor the e-commerce experience and maximize customer satisfaction.

In this context, the task of image-based virtual try-on aims at synthesizing an image of a reference person wearing a given try-on garment while preserving the person's intrinsic information such as body shape and pose. Existing datasets for the task are either proprietary and therefore not publicly available, or feature a very limited number of images which are usually low resolution [2]. To tackle these problems, AImageLab researchers recently presented *Dress Code* [4]: a new dataset of high-resolution images (1024×768) containing more than 50k image pairs of try-on garments and corresponding catalog images where each item is worn by a model. Differently from existing public datasets, which contain only upper-body clothes, *Dress Code* features upper-body, lower-body, and full-body clothes, as well as full-body images of human models.

In addition to the dataset, the researchers have also introduced a novel image-based virtual try-on architecture that can anchor the given garment to the right portion of the body. As a consequence, it is possible to perform a "complete" try-on over a given person by selecting different garments (Fig 4).

5. AI for Interior Design

An additional research and industrial innovation field on which AImageLab works is that of empowering aug-

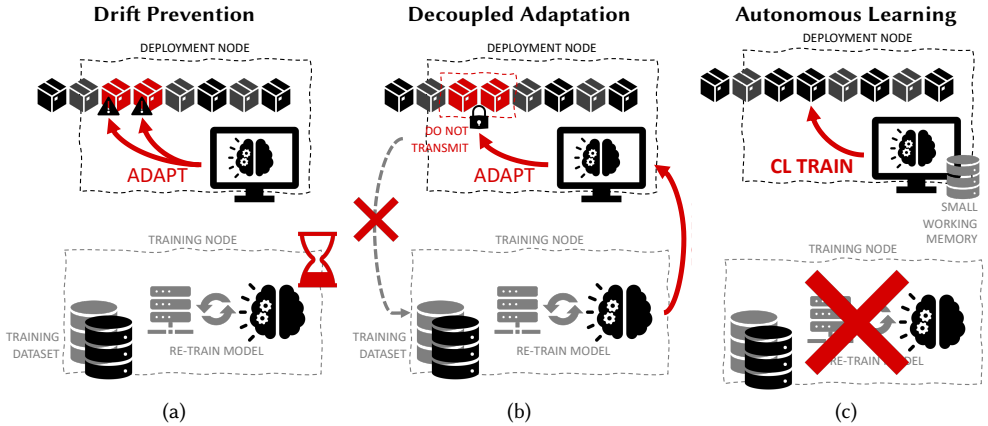

Figure 5: Innovative MLOps operation modes enabled by Continual Learning algorithms. (a) Adapting the in-deployment model on incoming out-of-training-distribution data while waiting for model re-training; (b) Allowing the in-deployment model fit specific data-points that cannot be transmitted to a separated training node (e.g., due to privacy concerns); (c) Enabling long-lasting model adaptation to changing data without retraining.

mented and virtual reality application for interior design. In this context, a fundamental task which is addressed by our research activities is that of automatically parsing and understanding pictures of indoor scenes. The goal of the task is that of providing detailed information about the objects in a scene, the layout of the space, and how objects interact with each other.

One of the core subtasks which need to be solved in this context is that of performing a semantic segmentation over the input image. The research on semantic segmentation models has focused on the introduction of either fully convolutional networks or Vision Transformers which leverage upsampling operations to increase the output resolution. Although this architectural choice is necessary to encode contextual information and deal with objects at large scales, it also leads to feature smoothing across object boundaries, and thus to a degraded quality in the final result.

With the aim of improving the quality of semantic segmentation in indoor scenarios, especially in boundary regions, we have investigated the design of boundary-aware losses for the optimization of semantic segmentation architectures, both in CNN-based and ViT-based architectures. We started from two recently proposed loss functions, namely the Boundary loss and the Active Boundary loss, and designed two improved versions that can significantly increase the overall quality of the segmentation at boundary level [7].

Our semantic segmentation solutions have been applied to the recognition of walls and floors in indoor images, and to the recognition of particular surfaces (e.g. counter-tops or steps). In addition to that, we also working towards the recognition of long-tail objects on the scene, and to the generation of natural language descriptions from indoor images [8].

6. Continual Learning

Human intelligence is distinguished from Artificial Neural Networks (ANNs) by the former's ability to acquire knowledge incrementally and retain long-term memory. Conversely, ANNs suffer sudden performance deterioration, termed *catastrophic forgetting*, in response to changes in training data distribution [9]. Continual Learning (CL) [10] is a rapidly growing area of machine learning which focuses on bridging this gap by devising technical solutions that allow AI systems to overcome forgetting. In recent years, the AImageLab team has actively contributed to CL research, focusing especially on the development of novel CL methods belonging to performant *rehearsal-based* category [11, 12].

While apparently abstract, this emerging discipline has significant implications for MLOps practices. Today, a model's life cycle typically begins with training on a dedicated **training node** (*e.g.*, a high-compute server), with the result being frozen and used for inference on a separate **deployment node** (a machine that operates *on the edge*, close to the source of operational data). Herein, we examine three scenarios (see also Fig. 5) that demonstrate the potential of CL algorithms to alter this paradigm.

Drift Prevention. In-deployment model re-training is required when significant differences occur between inference and training data distributions. As this process involves recording new data, transmitting it to the training node and learning a new model from scratch, the model's response to new data may be delayed or unreliable. By applying a CL algorithm on the deployment node, the model can adapt to incoming data gracefully controlling its performance degradation until an update is available (Fig. 5a). Recent advancements permit this procedure even with limited supervision on new data [13].

Decoupled Adaptation. Due to their physical separation, it is likely that some novel data-points recorded on the deployment might not be transmitted back to the training node (e.g., due to security constraints or technical limitations). In such a scenario (Fig. 5b), CL allows the model to fit these additional data-points without them leaving the node. When a re-trained model (unaware of secure data) becomes available, the CL learner cannot be trivially replaced, but must follow an adequate procedure to allow knowledge merging [14].

Autonomous Learning. As CL methods improve, it becomes increasingly feasible to let them operate over very long periods of time. Ideally, this will allow the in-deployment model to fully adapt to upcoming data dynamically by only leveraging a small working memory, removing the need of re-training.

While such a degree of resilience has not yet been achieved, AImageLab has developed a strong focus on removing bias from continual learners, allowing them to operate for longer timespans [15, 16].

7. Theft Recognition

The evolution of self-checkout methods in retail stores presents new challenges that computer vision and deep learning technologies are able to address. As there no cashier are involved in scanning the items, it is important to develop a system able to monitor the customer actions and detect anomalous behaviours such as shoplifting attempts or even good-faith mistakes. Modern embedded systems allow such solutions to work on the edge, providing useful feedback to on-site staff.

To achieve this goal, researchers of AImageLab developed an approach to track the movement of customers at the self-checkout station and used this data to classify their actions, distinguishing between malevolent and benevolent behaviours. This system can be divided into three main parts: (*a*) human behaviour detector, (*b*) action classifier, (*c*) retrieval network.

The role of the first part of the system is to extract from visual data useful information that are able to describe the type of actions of the person involved in the self-checkout. Once a good descriptor of the events has been computed, it is fed to a more complex architecture, able to link each scanning action to a set of pre-defined behaviors. Finally, the latter module attempts at bridging the gap between the camera and the traditional laser bar code scanner: briefly, it is carried out by learning a neural network specialized on a retrieval task, *i.e.*, identifying objects and ascertaining whether its visual appearance corresponds with the reading of its bar code.

The system has shown promising results in identifying different types of theft, demonstrating that research results can earn a relevant place in the retail sector.

8. 3D Pose Estimation

In the context of Industry 4.0, experts agree that the cooperation between humans and intelligent robots [17], rather than the complete removal of human operators, will be the best solution for the advancement in manufacturing. Therefore, safe interaction is a crucial element, especially regarding social and physical coordination between coworkers. In this scenario, the ability to predict the 3D pose of the agents in a collaborative environment is an enabling technology for real-world safety applications, *e.g.* collision avoidance and anomaly detection.

Thinking about a surveillance system installed in an industrial setting, vision-based pose estimation approaches can be exploited to retrieve the 3D pose of an articulated object with respect to the camera viewpoint. However, a model that jointly predicts human and robot poses is still an open problem in the research community. Thus, the current approach is to consider humans and robots as separate agents. Since the human pose task has been extensively investigated in the computer vision community [18], the focus of our work is on the robotic scenario.

Robot Pose Estimation (RPE). In robotics, the common approach to estimate the absolute pose of a robot with respect to the camera is the Hand-Eye Calibration. However, with recent advances in deep learning, many works have been proposed to estimate the camera-to-robot pose using CNNs. Recently, Lee *et al.* [19] demonstrate that learning-based approaches could replace classic marker-based calibration also for standard manipulators (*e.g.* Rethink Baxter and Franka Emika Panda). Using synthetic data for training, they feed RGB images into an encoder-decoder network that predicts the 2D location of the robot joints. Assuming the camera intrinsics and the configuration of the joint angles are known, the camera-to-robot transform is computed via PnP.

Going beyond learning-based methods, recent works propose approaches based on rendering. Labbe *et al.* [20] pave the way to this field of research presenting the first method for robot pose estimation based on the *render&compare* paradigm. This optimization algorithm iteratively refines an initial robot state defined as the joint angles configuration and the pose of an anchor part with respect to the camera.

SPDH Pose Representation. In contrast with the discussed works, our work aims to regress the camera-to-robot pose using a novel heatmap-based representation of the 3D pose. In this way, we can exploit any CNN-based architecture trained to predict heatmaps and adapt it to predict our pose representation. Moreover, since deep learning applied to robotics requires a lot of data and it is not feasible to record real data covering all the possible scenarios, we exploited depth images to close the domain gap between the synthetic domain of the sim-

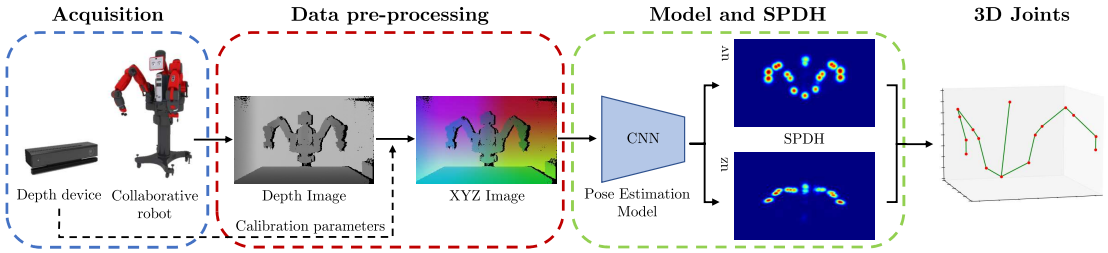

Figure 6: Overview of our 3D Robot Pose Estimation (RPE) system. A depth image is converted into an XYZ image which is given as input to a pose estimation deep model. The network predicts the proposed *Semi-Perspective Decoupled Heatmaps* (SPDH) from which the 3D robot pose is computed.

ulator and the real world. An overview of the proposed approach is depicted in Figure 6.

Regarding the pose representation, we propose the *Semi-Perspective Decoupled Heatmaps* (SPDH) [21] that rely on projections of the 3D space in two 2D spaces: *uv* and *uz* space. The *uv* space, *i.e.*the camera image plane, and the *uz* space, composed by quantized Z-values and the u dimension.A pose estimation algorithm is trained to generate two heatmaps for each joint, one for each space. The setting is a Sim2Real scenario, so the method is trained on synthetic data and tested directly on real data. Thus, we presented SimBa, a dataset containing RGB-D synthetic and real sequences with a Baxter robot performing *pick-n-place* movements. We proved that using depth maps as input reduces the domain gap obtaining promising results for the RPE task.

References

[1] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, M. E. Celebi, J. Yang, Image synthesis with adversarial networks: A comprehensive survey and case studies, Information Fusion 72 (2021) 126–146.

[2] X. Han, Z. Wu, Z. Wu, R. Yu, L. S. Davis, Viton: An image-based virtual try-on network, in: CVPR, 2018.

[3] M. Fincato, M. Cornia, F. Landi, F. Cesari, R. Cucchiara, Transform, Warp, and Dress: A New Transformation-guided Model for Virtual Try-on, ACM TOMM 18 (2022) 1–24.

[4] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, R. Cucchiara, Dress Code: High-Resolution Multi-Category Virtual Try-On, in: ECCV, 2022.

[5] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, From Show to Tell: A Survey on Deep Learning-based Image Captioning, IEEE Trans. PAMI 45 (2022) 539–559.

[6] N. Moratelli, M. Barraco, D. Morelli, M. Cornia, L. Baraldi, R. Cucchiara, Fashion-Oriented Image Captioning with External Knowledge Retrieval and Fully Attentive Gates, Sensors 23 (2023) 1286.

[7] P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara, Investigating bidimensional downsampling in vision transformer models, in: Image Analysis and Processing–ICIAP 2022, 2022.

[8] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: CVPR, 2020.

[9] M. McCloskey, N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of learning and motivation, volume 24, 1989, pp. 109–165.

[10] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, T. Tuytelaars, A Continual Learning Survey: Defying Forgetting in Classification Tasks, IEEE Trans. PAMI 44 (2022) 3366–3385.

[11] A. Robins, Catastrophic forgetting, rehearsal and pseudorehearsal, Connection Science 7 (1995) 123–146.

[12] P. Buzzega, M. Boschini, A. Porrello, S. Calderara, Rethinking experience replay: a bag of tricks for continual learning, in: ICPR, 2021.

[13] M. Boschini, P. Buzzega, L. Bonicelli, A. Porrello, S. Calderara, Continual semi-supervised learning through contrastive interpolation consistency, PRL 162 (2022) 9–14.

[14] M. Boschini, L. Bonicelli, A. Porrello, G. Bellitto, M. Pennisi, S. Palazzo, C. Spampinato, S. Calderara, Transfer without forgetting, in: ECCV, 2022.

[15] M. Boschini, L. Bonicelli, P. Buzzega, A. Porrello, S. Calderara, Class-incremental continual learning into the extended der-verse, IEEE Trans. PAMI (2022) 1–16.

[16] L. Bonicelli, M. Boschini, A. Porrello, C. Spampinato, S. Calderara, On the Effectiveness of Lipschitz-Driven Rehearsal in Continual Learning, in: NeurIPS, 2022.

[17] A. Weiss, R. Buchner, M. Tscheligi, H. Fischer, Exploring human-robot cooperation possibilities for semiconductor manufacturing, in: CTS, 2011.

[18] W. Liu, Q. Bao, Y. Sun, T. Mei, Recent advances of monocular 2d and 3d human pose estimation: a deep learning perspective, ACM Computing Surveys 55 (2022) 1–41.

[19] T. E. Lee, J. Tremblay, T. To, J. Cheng, T. Mosier, O. Kroemer, D. Fox, S. Birchfield, Camera-to-robot pose estimation from a single image, in: ICRA, 2020.

[20] Y. Labbé, J. Carpentier, M. Aubry, J. Sivic, Single-view robot pose and joint angle estimation via render & compare, in: CVPR, 2021.

[21] A. Simoni, S. Pini, G. Borghi, R. Vezzani, Semi-perspective decoupled heatmaps for 3d robot pose estimation from depth maps, IEEE RA-L 7 (2022) 11569–11576.