UNLOCKING VISUAL INSIGHTS WITH VISION TRANSFORMERS

Advantages, applications, and economic benefits of Vision Transformers in computer vision tasks

REPLY [MTA, STAR: REY] specialises in the design and implementation of solutions based on new communication channels and digital media. Through its network of specialist companies, Reply supports some of Europe's leading industrial groups in Telco & Media, Industry & Services, Banks & Insurance, and Public Administration to define and develop business models, suited to the new paradigms of Artificial Intelligence & Machine Learning, Big Data, Cloud Computing, Digital Media and the Internet of Things. Reply services include: Consulting, System Integration and Digital Services.

www.reply.com

EXECUTIVE SUMMARY

Artificial intelligence has become an integral part of the society we live in. Not only does it affect the lives of each and every one of us, but it is also transforming processes across all industries.

Technologies such as ChatGPT, GPT-4 or Dall-E are admired not only by AI (Artificial Intelligence) enthusiasts, but also by everyone else who couldn't have imagined this level of competence a few years ago. It is both fascinating and thought-provoking to see the pace at which AI is progressing. Just a few weeks after the release of the GPT4 model in March 2023, an open letter was published calling for a halt to the training of models more powerful than GPT4. This letter was signed by such powerful people in the community as Yoshua Bengio, Turing Prize winner, Elon Musk, CEO of SpaceX, Tesla & Twitter, Steve Wozniak, co-founder of Apple and Emad Mostaque of Stability AI (futureoflife.org, 2023).

So, what has changed? What was the key to the success of these technologies? We believe that the primary element behind these models is the combination of the transformer architecture with a learning approach that leverage large scale data (millions of samples scale) without any explicit supervision. This is also the main reason why large pre-trained models nowadays show a superior generalization power.

Following their release in 2017, transformers (Vaswani, 2017) have solved challenges in AI that have been limiting for cutting-edge neural models such as Long Short-Term Memory (LSTM) (Hochreiter, 1997) or Convolutional Neural Networks (CNNs) (LeCun, 1998). As for the pre-trained models, questions of why and how exactly large models generalize so well are still captivating AI researchers around the world.

WHAT MAKES VISION TRANSFORMERS SO SPECIAL?

Vision Transformers (ViTs) are a type of deep learning architecture specifically developed for computer vision tasks such as image classification, object detection, and semantic segmentation. Introduced by Dosovitskiy et al. (2020), it follows the original Transformer architecture by Vaswani (2017), which was designed initially for machine translation tasks.

There are two key advantages of Vision Transformers for computer vision tasks over traditional convolutional networks:

- State-of-the-art performance on benchmarks with better ability to capture the global information of the data, independent of the input size. Long-term dependencies in images are essential for tasks such as object detection and segmentation but have been a limiting factor for CNN (Convolutional Neural Network) models. Most importantly, according to Dosovitskiy (2020), when trained on larger datasets with 14M-300M images ViTs outperform CNNs.
- ViTs achieve the same performance as CNN with fewer parameters while being more parallelizable and requiring significantly less time to train; making them a more suitable option for large-scale image processing tasks.

VIT (VISION TRANSFORMERS) IN ACTION

Figure 1: Spot, industrial robot, with DINO, Object Detection Solution

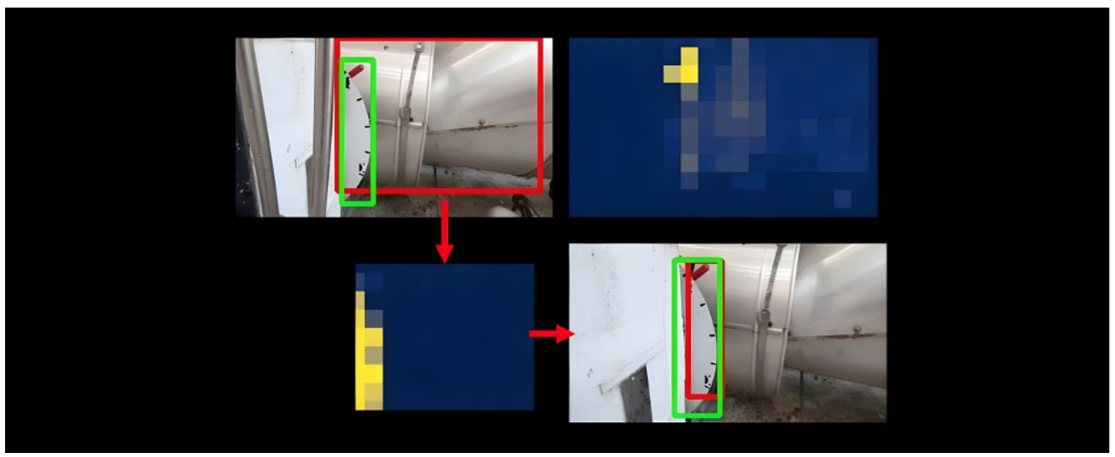


Figure 2: Object localization in industrial sites with DINO

In Reply, we've decided to validate Vision Transformers in real applications. The model that caught our attention was DINO. DINO is a state of-the-art AI model for computer vision tasks that stands for " self-DIstillation with NO labels", and was introduced by Meta AI in 2021 (Caron, 2021). This model, or rather this learning method, can generate numerical representations of images and what they depict. These representations, called embeddings, can ideally be used to unlock any kind of computer vision problem, as they are general purpose.

Within a year, we implemented the DINO model in three different use cases. All use cases involve DINO, for automatic feature extraction, and the usage of simpler and smaller ML (Machine Learning) models on top of these features (e.g., k-nearest neighbors classifier) to perform specific downstream tasks, such as object classification and localization. The use cases included the integration of DINO on board of autonomous robots and edge devices to perform automatically readings of numbers and texts like serial numbers and QR codes on product surfaces in production lines, number plates and recordings of measuring systems. In this white paper, we present one of the use cases that demonstrates the fundamental architecture of our solution, on board of the SPOT robot.

SPOT, Boston Dynamics' most friendly and agile autonomous dog-like robot, is widely used to safely perform monitoring and inspection of industrial sites, taking data-driven actions accordingly. We integrated our DINO based solution for object detection to enable automatic reading of industrial processes measurements.

The goal was to automate dataset labelling step for object detection tasks and perform image classification without actively training a Machine Learning model, but instead using a lazy k-NN classifier. The result led us to believe that the features extracted by DINO could encode the semantic layout of an image without a labelled dataset. Then, we verified how DINO performs unsupervised object localization of measuring tools in the environment of an industrial site. On the left side, the predicted bounding box (red) versus the true one (green), and on the right side, the image illustrates the attention map processed by TokenCut (Wang, 2022).

After a few weeks of intensive development, a team of our AI engineers achieved the following goals:

- used DINO to extract features from video frames without a labelled dataset.
- used TokenCut algorithm on top of DINO features to predict bounding boxes that locate a salient object in the foreground of an image.
- used k-NN classifier on DINO features to predict the class of the salient object without intensive model training.
- evaluated object localization and object classification accuracy, which resulted in 100% in 8 out of 11 classes, with the lowest being 87.5% in only one category.
- created a Docker image ready to run on edge devices.
- successfully integrated the object detection solution into SPOT robot.

HOW DO BUSINESSES BENEFIT FROM USING VIT?

The DINO use case for industrial inspection that we have described above is just one of the existing and potential use cases in which ViTs can optimize computer vision tasks without compromising on performance. ViTs offer a technology that is a part of some models such as DINO (Caron, 2021), VC-1 (Majumdar, 2023), ViViT (Arnab, 2021), Cait (Touvron H. C., 2021), Data-efficient image Transformers (Touvron H. C., 2021), CLIP (Radford A., 2021), and BeiT (Bao, 2021). Some are purely scientific in nature, while others have already found their application usage in industry. Here we will focus on the significant economic potential and cost savings associated with the use of Vision Transformers.

The main application areas where ViT can be integrated are:

- Quality control (e.g., anomalies, defects, action movements)
- Safety and risk management (e.g., visual inspection, disease detection, collision avoidance, surveillance)
- Automation (e.g., monitoring, counting)
- Forecasts (e.g., image reconstruction, climate prognosis)

QUALITY CONTROL

Product quality is a key factor in the success of a company. High quality production is directly related to the company's image and customer satisfaction. To produce a product with consistent quality, rules and requirements are set for the product. Also, it is essential to detect defective products early and to prevent their delivery to keep the associated complaint costs down. If quality control detection is well mapped via an algorithm with ViTs, a company will not only save the cost of complaints, but also the potential cost of labour to maintain quality standards. Below we will illustrate some of the application use cases, where the use of ViTs makes the process of quality compliance more efficient.

In their paper, Mishra et al. (2021), they show how production defects such as cracks in bottles, damage to nuts, drugs and wire harnesses can be detected. Even subtle anomalies, such as the fraying of toothbrush heads, can be localized with the help of ViTs. With a relatively small amount of training data, it keeps pace with current state-of-the-art technologies and even surpasses some of them.

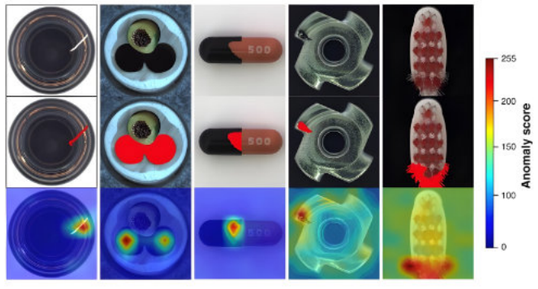


Figure 3: Anomaly detection by Mishra (2021). First row shows the actual anomaly image of the bottle, cable, capsule, metal nut and brush, row two shows the actual ground truth and row three displays the anomaly score and anomaly localization generated by ViTs.

Beside manufacturing companies, the service sector is also benefiting from the ability to recognize objects. For instance, a team of researchers at the University of Wuppertal taught an AI to inspect loading areas (Hütten N., 2022). During the loading of freight, the cargo or the loading and unloading process sometimes damages the loading areas of trains. This damage must be documented and inspected to ensure a safe and secure shipment. The model is capable of detecting damage to various materials such as wood and metal.

SAFETY AND RISK MANAGEMENT

Every business relies on the knowledge, processes and systems that enable it to deliver stable economic success. The employees hold most of the knowledge, while the processes and systems are executed by them. Protecting people’s health is a top priority because it is employees who run the systems and processes in their company. Machine vision can help minimize risks, especially when business processes are risky and life-threatening. The parcel delivery sector in particular has experienced a massive boom in recent years.

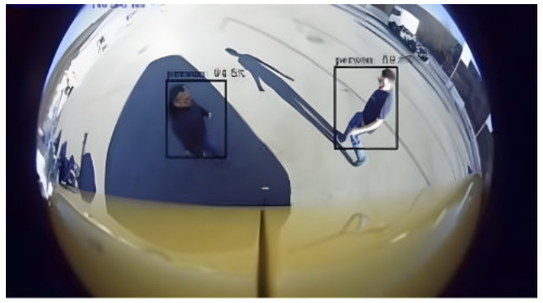


Figure 4: Theft detection with viso.ai by Intel® AI technologies

The large number of orders must be delivered on time, and the delivery drivers must cope with the flood of parcels under time pressure. As a result, the number of accidents involving delivery vehicles has increased. To address these challenges, an Intel AI team processed the images from the vehicle's cameras and warned the driver in advance of a collision (Klingler N. & Boesch G. , 2019). At the same time, cameras on the vehicle can be used to detect objects in the immediate area. This enables AI to detect and signal acts of vandalism, theft, or robbery of parcel carriers. Figure 4 illustrates a real-time monitoring of the environment during delivery and can warn staff about potential dangers.

Another major application where ViTs can be implemented is the object detection of helmets on industrial plants or construction

sites. In hazardous areas, it is important to wear helmets for health and legal reasons. Reported occupational accidents in Germany

in 2021 increased by 6 percent to 806,201. Of these, 103,525 accidents occurred in the construction industry (Boesch, 2022). As exemplified in Figure 5, AI is widely used to support employees with visual material and warns those who are not wearing a helmet.

This makes it possible to provide targeted advice to workers or to automatically block the operation of equipment. Here, ViTs can detect a wide range of protective clothing, as well as blurry or distracting parts of the image, without explicit training on a labelled dataset, which is financially and computationally expensive.

Figure 5: PPE recognition for helmet and vest detection (Boesch, 2022)

AUTOMATION

The definition of a process that machines can perform independently without human intervention is commonly referred to as automation. Analysts expect automated processes to reduce indirect costs in industrial operations by 15 to 20 percent in 12 to 18 months (Edlich, 2019). As a step towards automation in its driver assistance systems, Tesla is deploying Vision Transformer Models. As a result, curbs and lanes are better detected and autonomous functions such as turning at intersections are improved (Chen, 2021). Another interesting field of automation by computer vision is satellite imagery which is crucial for crop monitoring in the agricultural sector.
A team at Imperial College London is using advanced computer vision algorithms to analyse satellite images and identify the type of crop being grown and the size of the area being cultivated (Tarasiou, 2023). Regulatory processes are automated using this technology. For example, when farmers must report to the authorities on the management of their fields and agricultural processes, an algorithm can tell farmers about the condition of their crops.

FORECASTS

Climate affects us all, and despite its critical importance, it remains difficult to predict the weather in 7 days. In response, Microsoft's ClimaX approach uses cutting-edge AI solutions to improve weather and climate forecasting (Nguyen, 2023). At the heart of this approach is an architecture that uses transformers. The generated climate predictions answer the question: "What will be the annual mean temperature at a given CO2 level? To calculate forecasts with ViT, image data is used as input to the model.

THE MAIN TECHNOLOGIES BEHIND VISION TRANSFORMERS

The scope of this whitepaper is to highlight the potential of Vision Transformer models in computer vision applications. However, to understand why ViTs work so well, we will briefly introduce some of the basic knowledge required. The key idea behind Vision Transformers is to treat image data as a sequence of patches, aka regions of pixels, and use attention mechanisms to capture the relationships between regions to make a prediction.

Let us dive into two main technologies behind ViTs:

- The **self-attention mechanism** is the core element of ViT (Vision Transformers) networks. It is a type of attention mechanism, a function inspired by the way humans' reason, that tends to selectively focus on one part of information when and where it is needed but ignore other perceivable information at the same time. As is the case for us humans, when we notice that there is often something in a scene that we want to observe in a certain part, we learn to focus on that part when similar scenes recur and focus more on the useful part. Therefore, this is the way harnessed by the attention mechanism to quickly select high-value information from massive information, using limited processing resources. The attention mechanism greatly improves the efficiency and accuracy of perceptual information processing by allowing models to handle long-term dependencies regardless of input size and to incorporate global information about the data simultaneously at each processing stage (Vaswani, 2017). Based on this idea, ViTs break down the image into a sequence of small patches (8x8, or 16x16 pixels), each of which is then processed by a transformer-based neural network (Dosovitskiy, 2020). The transformer network is composed of multiple layers of self-attention and feedforward neural networks.

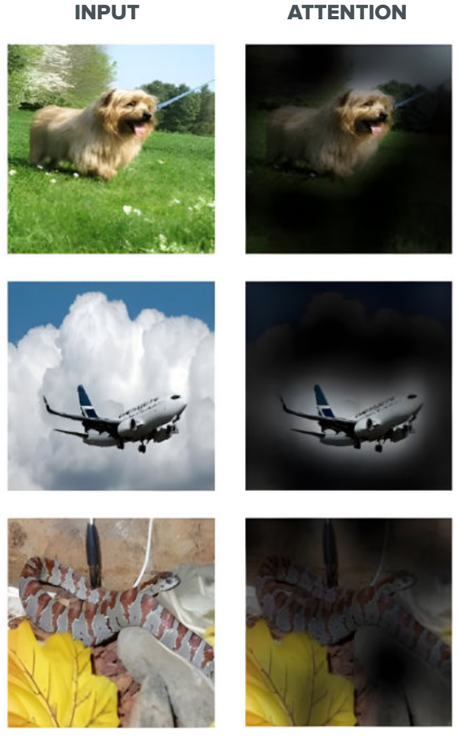


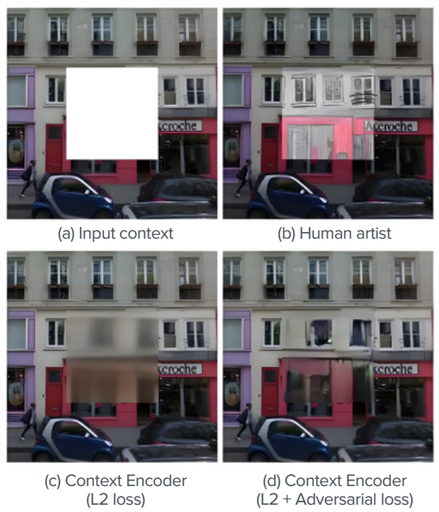
Figure 6: Self-Attention Mechanism recognizing the main subject of the image (Dosovitskiy, 2020)

Illustrated by Figure 6, the self-attention mechanism allows the model to capture global dependencies between the patches, enabling it to recognize complex global patterns and structures in the image. In contrast, traditional CNNs work hierarchically, processing the input image through a series of convolutional layers followed by pooling layers (LeCun, 1998). This approach works well for many image processing tasks where local information is essential but struggles with big images that require a high number of convolutional layers and a correspondingly large number of parameters.

Self-supervised pre-training on large scale dataset is another key ingredient for the success of Transformers.

- **Self-supervised learning (SSL)** doesn't need human labelled data to learn meaningful data representations. For this learning method the supervision signals (labels) are created from the raw input data itself (Yann LeCun, 2021). For example, in NLP (Natural Language Processing), a model would learn textual representations by masking a word in a sentence and using the masked word as a label. In computer vision, multiple transformations are applied to the input image (e.g., cropping, masking, rotating, blurring, etc), and the model learns the image representations by reconstructing the transformed image to its raw version, or recognizing whether two augmented versions represent the same object or not. Before a model can predict the generated label, it needs to understand the contextual information surrounding it. Image inpainting is one of the simplest examples of image manipulation used in self-supervised learning.

The image below illustrates the results of the original paper from Deepak Pathak on image inpainting with a CNN model trained on ImageNet and Paris Street View Dataset (Pathak, 2016). The strongest advantage of SSL approach is that it enables AI systems to learn from greater orders of magnitude of data (tens or hundreds of millions), which is important to recognize and understand patterns of more subtle, less common representations of the world. By pre-training large transformers with massive parameter size on self-supervised large-scale datasets, we allow the models to learn general representations of the data that can be used to tackle unseen tasks and data distribution. Moreover, these pre-trained models can then be fine-tuned or customized for specific tasks and data distributions by using smaller, supervised datasets (few hundreds of samples) (Dosovitskiy, 2020).



(a) Input context (b) Human artist

(c) Context Encoder (L2 loss) (d) Context Encoder (L2 + Adversarial loss)

Figure 7: Image inpainting with a CNN model (Pathak, 2016)

ETHICS AND REGULATIONS

With the rise in popularity of AI algorithms, the ethical and legal aspects of the use of AI are increasingly important. There are several critical issues that need to be addressed in the immediate future to steer the development and use of AI towards positive impact and with full compliance with fundamental human rights. Here, we will only touch on some of the issues that we consider important in regard to the computer vision field. Most of these concerns are backed by an intensive study (Rodrigues, 2020). As systems based on ViT use data for object detection, including face, location or motion, here are some of the key ethical and legal concerns to consider:

- Privacy, data protection and lack of informed consent of data subjects
- Transparency and accountability
- Bias in the data

PRIVACY, DATA PROTECTION AND LACK OF INFORMED CONSENT OF DATA SUBJECTS

Privacy concerns are ubiquitous for any technology using user data. Cases particularly relevant for object detection tasks include the usage of user data for model training. Wachter (2019) highlights concerns about algorithm accountability, emphasizing that "individuals are given little control or oversight over how their personal data is used to draw inferences about them". AI researchers and developers are key parties who must ensure conformity with moral and legal obligations. Vayena (2018) comments that developers should "pay close attention to ethical and regulatory constraints at each stage of data processing. Data provenance and consent for use and reuse are considered particularly important".

TRANSPARENCY AND ACCOUNTABILITY

AI object recognition models are increasingly involved in decision-making in high-risk areas such as autonomous driving, law enforcement, and military. The European Union has published a major study (European Parliament Resolution, 2019) on the importance of the interpretability of AI in its decisions. How does the model work, and why does it make the decision it makes, are key questions that would help ensure accountability of the AI in its decision.

BIAS IN THE DATA

Bias in data leads to unfairness and discrimination. There have been several cases where object recognition tasks do not work properly when it comes to identifying various minorities. Ethnic/racial/gender stereotyped minorities, as well as poor/low-income earners, may be identified as suspicious in object recognition tasks and be mistakenly prosecuted. Some ways of solving the problem could be human intervention centred on ethical design of AI systems. The European Parliament (2017) highlights that “because of the data sets and algorithmic systems used when making assessments and predictions at the different stages of data processing, big data may result not only in infringements of the fundamental rights of individuals, but also in differential treatment of and indirect discrimination against groups of people with similar characteristics, particularly with regard to fairness and equality of opportunities.” There are several organizations whose aim is the development of standards for safe and responsible AI research.

The IEEE (Institute of Electrical and Electronics Engineers) P7003 Standard for Algorithmic Bias Considerations (A. Koene, 2018) is a set of guidelines created by the Institute of Electrical and Electronics Engineers (IEEE) to help organizations and individuals identify, mitigate, and prevent biases that may arise in the development and deployment of algorithms. The standard provides a framework for addressing bias in algorithmic decision-making systems, including machine learning models and other AI systems.

Another important organization to mention here is ISO that develops standards and guidelines related to AI (ISO/IEC 23053, 2022). ISO has established a technical committee (TC 307) that focuses on developing standards related to AI and related technologies. ISO TC 307 is responsible for developing international standards for AI terminology, reference architectures, ethical considerations, and data management, among other topics. These standards are intended to provide guidance and best practices for the development and deployment of AI systems and technologies. The involvement of government in the regulation of AI is another key issue that needs to be addressed. Here, one of the biggest concerns would be how much policymakers can understand the nuances of the technology to regulate it. We believe that the necessary expertise to make the right decisions in regulating AI can be achieved through an interdisciplinary approach of AI researchers, legal scholars, and government. An important argument for government involvement is the profound impact that technologies controlled by a limited number of business owners or decision-makers can have on our lives.

THE FUTURE OF VITS

For the future development of ViTs, we see two major steps forward. The first step would be to further optimize the computational resources needed for training and inferencing steps in real use cases. Although transformers have optimized the traditional training process of deep learning models for computer vision tasks, the fine-tuning and inference of visual Transformers is still computationally expensive due to a computationally intensive attention mechanism.

The next necessary step to improve the limited local interpretability of ViTs is the integration of state-of-the-art Explainable AI methods. Vision Transformers, like other deep learning models, can be difficult to interpret, making it difficult to understand why they make certain predictions. ViTs offers amazing performance when it comes to understanding global information in data, but it compromises local information that is essential for certain tasks. This could be the biggest drawback when it comes to trusting the decision of deep learning models.

CONCLUSIONS

In this whitepaper, we provided an overview of Vision Transformers in terms of their commercial viability, architecture and ethical aspects. We also presented our work at Reply, where we have successfully implemented Vision Transformers for real-world use cases such as object detection for object detection in industrial sites. We strongly believe that considering the advantages of a pre- trained ViT on a significantly large dataset, it can produce better results than traditional state-of-the- art CNNs for object detection. With the successful implementation of DINO, we also see the benefits of combining transformer models with a few labelled examples to achieve object detection. Achieving the same results with way less data is a powerful new feature, because most computer vision projects spent the majority of their resources on dataset creation.

With our own implementations and use-cases, we prove that the combination offers scalability that eliminates the cost of manually labelling a dataset, the computational cost of pretraining, and does not compromise the accuracy of the object detection tasks.
In addition, we have identified the key areas where we believe ViTs would possibly be of great benefit. We are seeking more real-world use cases that would benefit from the advantages we have identified. We are committed to using AI in a fair and respectful way, not only for the purpose of technological progress, but more importantly for the benefit of our society, through extensive research into the ethics and regulation of AI.

REFERENCES

A. Koene, L. D. (2018). IEEE P7003TM Standard for Algorithmic Bias Considerations. IEEE/ACM International Workshop on Software Fairness (FairWare), (pp. 38-41).

Arnab, A. D. (2021). Vivit: A video vision transformer. In Proceedings of the IEEE/CVF international conference on computer vision, (pp. 6836-6846).

Bao, H. D. (2021). Beit: Bert pre-training of image transformers . arXiv preprint arXiv:2106.08254.

Radford, A. (2021). CLIP / Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.

Majumdar (2023). Visual Cortex (VC-1) / Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence? https://eai-vc.github.io/

Boesch, G. (2022). Computer Vision in Energy and Utilities Industry Applications.

Caron, M. T. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9650-9660). IEEE/CVF international conference on computer vision.

Chen, K. (2021). https://kimbochen.github.io/blog/2021/11/09/tesla-ai-day-vision.html#learning-where-to-look.

Dosovitskiy, A. B. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Edlich, A. P. (2019). Driving impact at scale from automation and AI. McKinsey Global Institute, 100. European Parliament Resolution. (2019). Autonomous driving in European transport. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52019IP0005.

futureoflife.org. (2023). Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open- letter/pause-giant-ai-experiments/.

Hochreiter, S. &. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Hütten N., M. R. (2022). Vision Transformer in Industrial Visual Inspection. Applied Sciences 12.23, 11981.

ISO/IEC 23053. (2022). Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). ISO/IEC 23053.

Klingler N. & Boesch G., v. (2019). How Deep Learning with Visuals is Transforming Delivery Logistics.

LeCun, Y. B. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, (pp. 86(11), 2278-2324).

Mishra, P. V. (2021). VT-ADL: A vision transformer network for image anomaly detection and localization. IEEE 30th International Symposium on Industrial Electronics (ISIE).

Nguyen, T. B. (2023). ClimaX: A foundation model for weather and climate. arXiv preprint arXiv:2301.10343.

Pathak, D. K. (2016). Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, (pp. 2536-2544).

Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities.

Journal of Responsible Technology, 4, 100005.

Tarasiou, M. C. (2023). ViTs for SITS: Vision Transformers for Satellite Image Time Series. arXiv preprint arXiv:2301.04944.