Data Annotation Concept based on Self-Supervised Learning for Computer Vision Developments in Manufacturing

Bachelorarbeit

für die
Prüfung zum Bachelor of Science

an der Fakultät für Wirtschaft
im Studiengang Wirtschaftsinformatik
in der Studienrichtung Data Science

an der
DHBW Ravensburg

Matrikelnummer:	1032686
Verfasserin:	Isabel Janez
Wiss. Betreuer:	Prof. Dr. Martin Zaefferer
Ausbildungsbetreuerin:	Anja Fessler
Ausbildungsbetrieb:	ZF Group
Anschrift:	Löwentaler Straße 20
	88046 Friedrichshafen
Abgabedatum:	03.07.2023

Abstract

This thesis presents a concept for improving the data annotation job for computer vision applications in manufacturing. The focus is on process optimization, cost reduction and resource conservation. The proposed concept is based on the masked autoencoder concept as a self-supervised learning approach. The goal is to provide a scalable, widely applicable solution that can be integrated in the machine learning operations lifecycle. Three specific manufacturing datasets are used for concept validation and the result evaluation is calculated by the loss function and image congruence.

The thesis concludes that the proposed data annotation concept saves resources, improves model quality, and enables organizations to scale artificial intelligence, data, analytics, and model development. As a result, organizations benefit from efficiencies, cost reductions, more robust models, transparency, computer vision experience, and expanded deployment capabilities. Implementing a data annotation concept based on the self-supervised learning approach can significantly improve computer vision performance in the manufacturing industry.

Contents

Acronyms

AI	Artificial Intelligence
AL	Active Learning
CI/CD	Continuous Integration, Continuous Delivery
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CV	Computer Vision
DA	Data Annotation
DAC	Data Annotation Concept
DCNN	Deep Convolutional Neural Network
DL	Deep Learning
FCN	Fully Convolutional Network
GAN	Generative Adversarial Network
HED	Holistically-Nested Edge Detection
ML	Machine Learning
MLOps	Machine Learning Operations
MSE	Mean Square Error
NLP	Natural Languages Processing
OKR	Objective and Key Results
R-CNN	Region Based Convolutional Neural Network
RRI	Responsible Research and Innovation
SMSL	Semi-Supervised Learning
SSL	Self-Supervised Learning
SUSL	Unsupervised Learning
SVM	Support Vector Machine
VQA	Visual Question Answering
YOLO	You Only Look Once

List of Figures

List of Tables

1 Introduction

Artificial Intelligence (AI) is a fast-moving field of technology that is used in many areas. Through the technology applications, efficiency, *Accuracy* and decision-making are improved. This improvement primarily takes place in manufacturing, healthcare, finance, and transport (Steidl et al., 2023). Edge AI in particular is a technology that is receiving a special boost in manufacturing. It is the process of executing AI algorithms directly on an end device using sensor data or signals. One of the biggest areas of opportunity in Edge AI is *Computer Vision* (CV), as the edge architecture delivers significant performance improvements and benefits for CV applications.

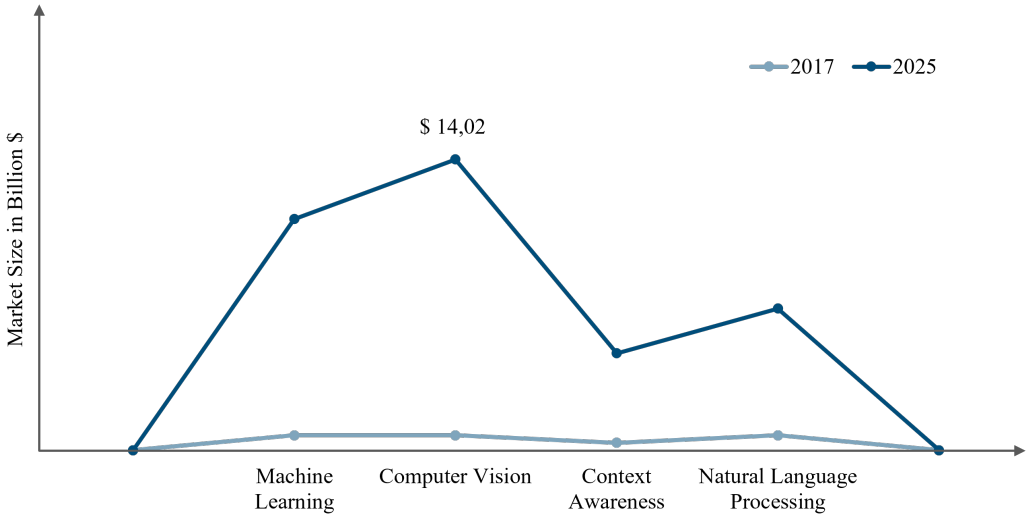


Fig. 1.1: Global AI in Manufacturing Market (based on Priyanka, 2018)

The CV market shown in Fig. *1.1* is anticipated to rule AI in the manufacturing industry in terms of revenue throughout the projected period (Priyanka, 2018). In this context, CV dominates in the area of automated testing and monitoring systems that help improve industrial quality and is also used for inspection, workplace safety, factory automation, and quality control (Nguyen et al., 2022).

Manufacturing benefits greatly from CV since it encourages the automation, digitalization, and intelligence of industrial manufacturing systems (Konstantinidis et al., 2021). The following application areas benefit from CV: product design, modeling, and simulation, planning, and scheduling, the production process, inspection and quality control, assembly, transportation, and disassembly are just a few of the manufacturing-related uses for CV (Zhou et al., 2023).

DAs are essential for the use of CV applications. Various possibilities to deal with DAJ have already been addressed in research for example *Self-Supervised Learning* (SSL), which yielded promising results. Baevski and He showed the use of SSL for the DAJ. Here the manual DAJ is automated and replaced by SSL (Baevski et al., 2022; Kaiming et al., 2021). As a result of the manufacturing sector's digitalization and the continued development of computer-integrated production systems, it can be said that the use of CV as a field of AI holds tremendous potential (Abagiu et al., 2021).

1.1 Motivation

CV is becoming increasingly more important in the industry and especially for manufacturing companies as it enables a variety of applications. For this reason, the ZF Group sees the need to develop a scalable solution in this area. It is also about investigating the fundamentals of visual and data understanding.

This visual quality inspection or manufacturing optimization based on CV depends on one factor, the quality of the training data. The accessibility and *Generalization* of the training data thus depend strongly on the characteristics of the data. Consequently, the ambiguity of *Labels* can be reinforced by unclearly defined DAs.

The accuracy of the systems is determined by the training data characterized by human annotators. This procedure is time-consuming, costly, and prone to errors. This results in a barrier to the development and implementation of CV applications within a commercial setting (Bulten et al., 2020).

To address the DA challenges various methods have been developed to counteract the mentioned difficulties, like *Semi-Supervised Learning* (SMSL) and *Unsupervised Learning* (USL). However, these approaches require a considerable amount of labeled data. This requirement is difficult to ensure in an industrial context due to the need for human resources is greatly reduced (Czimmermann et al., 2020).

In contrast, there is a promising approach to predicting the properties of the data without explicit monitoring. Here we are talking about SSL. Among other things, this could lead to more efficient and cost-effective CV applications in the industrial context (Jing and Tian, 2019).

SSL models make it possible to understand relationships between different input segments. SSL is used more and more in the area of CV (Emam et al., 2021). It is possible to connect CV systems with self-adjustment capabilities with current execution systems to correct flaws in real-time, advancing intelligent system design towards allowing zero-defect production at the human and system level. SSL methods are advancing the potential to transform the area of *Machine Learning* (ML)(Konstantinidis et al., 2021).
Efficiency can be increased by automatically annotating and labeling data. This reduces the need for costly and time-consuming use of human resources. This functionality offers immense advantages, especially in the area of CV (Nguyen et al., 2022).
The potential for the use of AI in manufacturing is becoming ever greater. In order to exploit these potentials and increase the reach, it is necessary to leave proof of concepts and/or single solutions for specific use cases behind and instead pursue the goal of AI product development. After all, the idea behind AI products is this: Scaling AI products is about deploying and leveraging them across multiple systems, processes, or even entire factories. This can be a complex task because of the need to ensure seamless integration of AI algorithms, data pipelines and compute resources. In addition, the scalability of AI products is not limited to their technical aspects but also includes considerations of their adaptability, maintenance, and ongoing improvement.
Machine Learning Operations (MLOps) is a set of practices and tools that focus on streamlining and automating the entire lifecycle of ML models, from development to deployment and maintenance. It helps address the scaling of AI product development in manufacturing (Treveil et al., 2021).

1.2 Problem

As highlighted in the previous section, the use of AI products in manufacturing is of interest. One important aspect that needs to be addressed in the context of AI in manufacturing is the scaling of these AI products.
In addition, the importance of data quality goes beyond the initial deployment of AI products. As manufacturing processes evolve and new data becomes available, ongoing data quality management is critical. Regular assessment and refinement of data pipelines, feature engineering techniques, and model retraining processes are essential to adapt to changing circumstances and ensure optimal performance.

When dealing with a single CV application is large, DA may not be a major issue. However, when the number of CV applications, DA becomes a difficulty and time-consuming. To handle this time-consuming DAJ, especially when it comes to AI Product development, where the scaling aspect plays a crucial role, a Data Annotation Concept (DAC) becomes relevant. The *Data Annotation Job* (DAJ) plays a conclusive role in the development of precise ML models. This involves identifying important elements in the data and adding labels to them, making them understandable or recognizable to machines.
Depending on the activities you wish to complete, there are numerous types of annotations. So far, the DAs in CV applications are created manually by annotators. The manual DAJ has some disadvantages. Depending on the time-consuming and labor-intensive procedure, the costs are increasing. Domain-specific knowledge is required for most ML use cases. This in turn increases the complexity and cost of DAJs. Improper or inconsistent data labeling has a major influence on the management of the ML model. In the worst case, this can lead to unreliable and false assumptions and discoveries (Simonyan and Zisserman, 2014).
In CV, labels are critical to the recognition and categorization of images and videos to identify entities or specific characteristics within the visual data. The common approach of manual *Data Annotation* (DA) is also selected for this. The search for efficient and general effective labeling and DA methods is therefore crucial for improving the accuracy and consistency of ML methods such as CV (Isola et al., 2017).
Ultimately, the DAJ is an important step in the development of ML models. The difficulties mentioned above, as well as the demand for high precision and consistency, cause significant problems. Researching new approaches, such as SSL, is crucial for developing fast and effective DA concepts that are used in a variety of CV applications (Ren et al., 2021).

This leads to the following research questions:

1. Can the performance of CV in the visual quality inspection context in manufacturing be improved through the implementation of a DAC based on the SSL method?
2. How can CV in the visual quality inspection context in manufacturing be improved through the implementation of a DAC based on the SSL method?

1.3 Objective

Based on the problem statement, it becomes clear that DAs have an important role and have an impact on CV development and the quality of its training data. In order to provide a solution approach to the problem, the objective of this thesis is to develop a concept approach using SSL to improve the annotation of data in the industrial context for CV applications. Thereby, the focus is on process optimization and the reduction of costs and resources.

To ensure that the approach is universally applicable, it is incorporated into a conceptual framework. The Data Annotation Concepts (DAC) with its outcome, must ultimately be embedded in the respective area of the holistic ML lifecycle. This should ensure the *Scalability* of the developed concept and increase transparency through better maintainability.

Furthermore, the aim is to achieve improved model performance in the optimization context. The development and use of such a practical concept can lead to higher efficiency and productivity due to an increment in model quality and process optimization. The validation of this thesis will be done based on three datasets.

In order to achieve the objective as well as to answer the research questions, the following topics are addressed in this thesis:

- **Chapter 1**: Gives an overview of the topics of motivation, problem statement, objective, delimitations, and methodology.
- **Chapter 2**: Introduces the theoretical principles of CV, DAJ, SSL, MAE, and MLOps, which gives a fundamental understanding of the topics tangible for the research question discussed in this thesis.
- **Chapter 3**: Highlights the practical application of the theoretical principles. Thereby, the data annotation concept is introduced, with its relevant phases, requirements, and aspects. Furthermore, the concept is being evaluated in this chapter and integrated into the holistic framework of an exemplary AI product called CVT.
- **Chapter 4**: Sums up the topics discussed in the previous chapters.

1.4 Delimitation

The thesis will focus exclusively on CV as one of many fields in the AI environment. Concerning DA, the concept development of a SSL approach using an *Masked Autoencoder* (MAE) for DA in CV in the industrial context. Applications or models outside this domain are not addressed. Other related topics such as SMSL, USL, and other SSL methods are not covered.

Each phase of the ML lifecycle is critical to the success of the model, and a thorough understanding of the process is essential to developing accurate and effective models. However, the focus of the work will be on labeling the data. Only one specific SSL methodology is used. MAEs are used to ensure comparability and reproducibility of the results. The validation is limited to certain metrics, the *Loss* and an image congruence. Other metrics such as the mean squared error are excluded.

Quantitative approaches are pursued, such as use case processing and validation, while qualitative methods, except for literature research, are excluded.

The thesis focuses on the industrial context and refers specifically to the application of CV methods in manufacturing. Other applications outside the industrial context are not considered. With regard to DAs, only image data is considered.

The data from productions comes exclusively from a single database provided by the ZF Group to ensure comparability and consistency of the data. This data is collected and processed within the CVT.

1.5 Methodology

To clarify the procedure, this thesis pursues a quantitative approach that takes place in five successive phases shown in Fig. *1.2*.

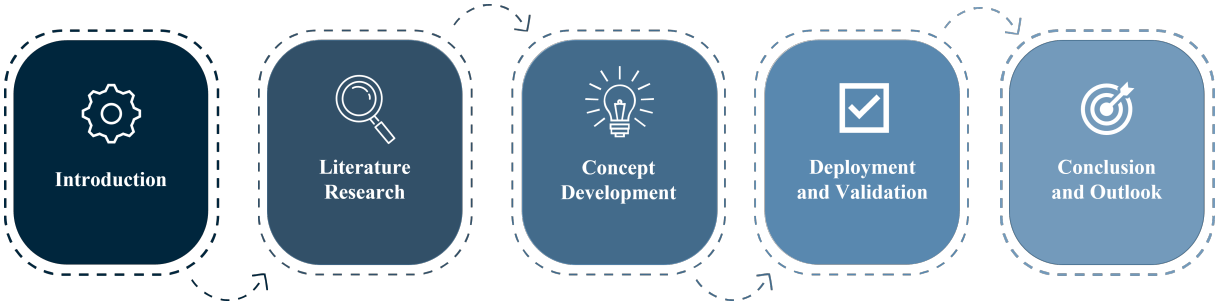


Fig. 1.2: Procedure and Methodology

The first phase describes the general introduction in chapter *1* into the topic of the thesis. It deals with motivation, problem definition, objectives, and delimitations.

In the second phase, an extensive literature search is carried out in chapter *2*. Special attention is paid to the current state of research in the field of SSL and DA in the context of their application in CV.

In phase three, a concept draft is created to solve the DA problematic in section *3.1*. A DA concept is developed based on MAE - SSL method. The focus is on the direct applicability of the concept in CV systems. The aim is to develop an efficient and cost-effective concept for the DAJ that can be used in industrial contexts such as manufacturing.

The concept validation phase takes place in phase four in section *3.2*, which includes the evaluation of the effectiveness and efficiency of the proposed concept on a specific data record from the ZF Computer Vision Tool Box (CVT). Here, the results of the SSL approach will be evaluated based on the loss function and image congruence. By analyzing the results, statements can be made about the quality of the concept and the further course of action in the CVT context can be defined.

Concluding in phase five, elaborated in chapter *4*, a summary of the results and the most important findings is presented at the end of the work. In a separate section, the limitations of the work are discussed and possible approaches for future research are shown. In the critical discussion, the limits of SSL and computer-aided DA are also discussed. The further procedure is defined to implement the developed concept in practice and to test its applicability in various industrial contexts. Finally, an outlook is given on the practical application of the proposed concept in the industrial context.

Conventions used in this thesis

To make terminology more comprehensible, the following typographical conventions are used in this thesis.

Italic

Glossary entries and proper names are formatted in italics to make them visually prominent and easier to distinguish from other parts of the text. This makes it easier to navigate and find specific terms within a document.

`Programming`

Programming content such as code snippets and parameters are presented in special programming font to clearly distinguish them from other text elements. Using this font for code to improve readability and visually grasp code examples.

$Mathematical fonts$

Mathematical content, such as equations, is presented in mathematical fonts to clarify its specific structure and meaning. Using this font makes mathematical expressions clearer and easier to understand, improving readability and interpretation.

2 Theoretical Principles

The basic concepts of CV, SSL, and DAJ are discussed in the next section. Therefore, this section provides the reader with a basic theoretical foundation for CV, SSL, and DAJ and clarifies their importance in many areas. Within the CV context the techniques, and advances, including the burgeoning research area of SSL, are being examined and described. Based on the information on CV, SSL is explored in more detail as a burgeoning field of research in the area of CV.

2.1 Computer Vision

CV also called Machine Vision is a sub-area of AI that focuses on understanding and interpreting visual data. This includes algorithms and models that can analyze images and videos and detect objects, patterns, or anomalies. Some fields of application of CV are autonomous driving, facial recognition, medical imaging, and monitoring. The research area of CV is constantly expanding and will achieve some fascinating breakthroughs in the coming years (Kendall and Gal, 2017).

Relevance of Computer Vision

CV is one of the fundamental technologies utilized in intelligent manufacturing. It has shown to be an excellent replacement for artificial visual inspection (Davies, 2012; Park et al., 2016). Vision is one of the most advanced degrees of human perception. CV is a system that receives and interprets images of actual objects automatically using optical devices and noncontact sensors (Kim and Lee, 2017).

It has become widely employed in industry as a measuring and judgment technology as computer equipment and AI have advanced. CV detection technology has the potential to improve detection efficiency and automation, improve real-time detection performance and accuracy, and minimize human needs, particularly in large-scale repetitive industrial production processes. CV may be used to accomplish automation, intelligence, and precise control since it is a non-contact and nondestructive detecting approach.

Furthermore, CV has a wider range of spectral responses and a greater ability to work in harsh environments for extended periods of time. The use of CV in manufacturing can improve a wide range of industrial activities (Penumuru et al., 2020; Ren et al., 2022).

Opportunities of Computer Vision in Manufacturing

The CV architecture can act as an instructional guide for creating a visual inspection system. For instance, the initial stage in constructing an extremely reflective metal surface visual inspection system was to investigate surface properties. As a result, diffuse bright filed backlight illumination was chosen. The next step was the capture of images using light-sensitive components. Following image acquisition, wavelet smoothing was used for image preprocessing, and Otsu threshold was used to segment the image. Finally, an SVM classifier was created for defect categorization (Xue-wu et al., 2011).

It can support the digitalization and intelligence of manufacturing. In addition to applying to various stages of the entire product lifecycle. CV techniques can also be used for feature detection, recognition, segmentation, and three-dimensional modeling (Zhou et al., 2023).

Additionally, it can be used to forecast the electrical characteristics of photovoltaic modules, which aids in module characterization in manufacturing, Research & Development, and management and operations of power plants. It can assist to boost productivity, lower faults, and enhance product quality (Karimi et al., 2020).

CV tasks include image classification, object detection, semantic segmentation, image captioning, and visual question answering. The tasks entail instructing computers on how to interpret and understand visual information from their environment. By combining data from several sources, the approach of multimodality may be utilized to increase the accuracy of CV models.

Mulitmodality

Deep Learning (DL) has a subset called multimodal DL that deals with the integration and analysis of data from several modalities, such as text, images, video, audio, and sensor data. Better performance on a variety of ML tasks is achieved by multimodal DL, which utilizes the advantages of many modalities to produce a more thorough representation of the data. Mulitmodality combines data from several modalities including images, text, and voice, has recently come to be recognized as a potential method for tackling challenging CV tasks (Potrimba, 2023).

In 2022, vision transformer models gained great importance and established their dominance in the field of CV. These models utilize SSL methods and show immense potential in various tasks. In addition, recent research has placed increasing emphasis on exploring the overlap between the fields of CV and NLP.

The integration of CV and NLP has led to exciting developments, although multimodality is still in its infancy. The combination of visual and textual information has shown remarkable benefits, opening up new opportunities for solving complex problems and improving performance in areas such as captioning, answering visual questions, and visual storytelling.

Overall, advances in Vision Transformers, SSL methods, and the fusion of CV and NLP through multimodality have advanced the field and offers promising avenues for further exploration and innovation (Ayman, 2023).

Application of Computer Vision

Success in a variety of CV tasks have been enabled through deep neural models based on *Convolutional Neural Networks* (CNN), including imagine classification, object detection, semantic segmentation, image captioning, and visual question answering (Selvaraju et al., 2019).

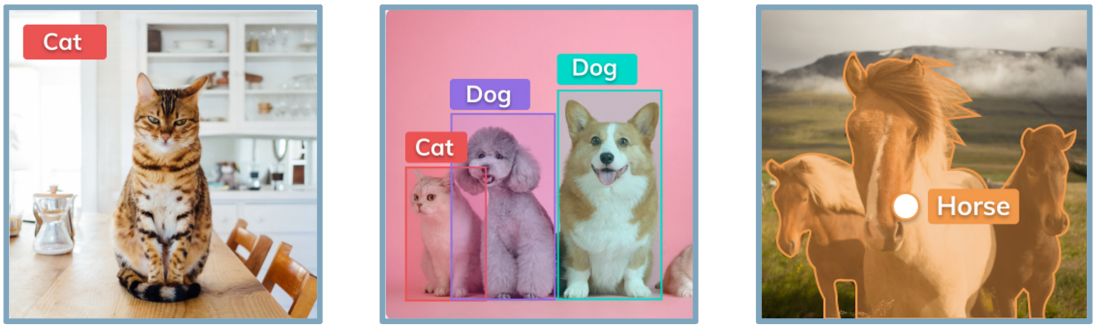


Fig. 2.1: Image Classification, Object Detection and Semantic Segmentation (Rizzoli, 2021)

Fig. *2.1* shows the difference between Image Classification, Object Detection and Semantic Segmentation using a visual example. These methods are explained in more detail below.

Image Classification is the process of assessing and categorizing the contents of an image (Vocaturo, 2021). Some methods for classifying images include AI-based systems, CNNs, and other *Supervised Learning* (SL), USL, and SMSL classification approaches (Shakya, 2020; Sanghvi et al., 2020; Simonyan et al., 2013).
CNNs, *Transfer Learning*, Support Vector Machines (SVM), k-nearest neighbor algorithms, and random forest algorithms are just a few examples of the many image classification approaches that are accessible (Sanghvi et al., 2020). For instance, to build a speedy billing system in the grocery business, CNN-based classifiers are proposed to recognize items by viewing via the camera (Tripathi, 2021; Shakya, 2020).

Object Detection entails locating and recognizing items in an image or video. Thereby, one-stage and two-stage detectors are two of the several methods for object detection. One feed-forward fully CNN, seen in one-stage detectors, immediately supplies the bounding boxes and the object detection. The two-stage frameworks, on the other hand, split the detection process into two stages: the region suggestion stage and the classification step. One-stage detectors have the potential to be quicker and easier while still falling below two-stage detectors in terms of accuracy. They are used over a regular, dense sampling of prospective item locations. After filtering out the majority of negative locations, two-stage detectors provide a sparse collection of candidate proposals that should include all objects. The second step then assigns each candidate location to one of the foreground classes or background. (Lin et al., 2020; Dai et al., 2016)
One of the most widely utilized object detection algorithms are Region-Based Convolutional Neural Networks (R-CNN), Fast R-CNN, Faster R-CNN, Single Shot MultiBox Detector, and You Only Look Once (YOLO).
YOLO is a unified model for object identification that, in a single assessment, predicts bounding boxes and class probabilities from whole images, making it incredibly quick and accurate. When applied to other domains, such as artwork, YOLO beats other detection techniques, such as DPM and R-CNN (Padma et al., 2019; Redmon et al., 2015). YOLO is one of the most popular techniques for recognizing and classifying items that appear on the road in the context of autonomous cars. Microsoft Azure Cloud object detection and Google Tensorflow object detection are two other object detection methods and libraries (Bratulescu et al., 2022; Noman et al., 2019).

Semantic Segmentation includes giving each pixel in an image a name based on its semantic significance. Fully Convolutional Networks (FCNs), DeepLab, and R-CNN are a few techniques for semantic segmentation. With effective inference and learning, FCNs are taught from beginning to end and may provide an output of commensurate magnitude. DeepLab integrates techniques from Deep Convolutional Neural Networks (DCNN) and probabilistic graphical models to improve the localization of object boundaries. It uses atrous convolution and atrous spatial pyramid pooling to segment objects of varying sizes. To locate and segment objects, R-CNN combines region suggestions with CNNs. It may potentially be expanded to the job of semantic segmentation. (Shelhamer et al., 2016; Chen et al., 2016; Girshick et al., 2013)

There are several methods for semantic segmentation, such as Conditional Random Fields (CRF), FCNs, and Holistically-Nested Edge Detection (HED) (Anilkumar and Venugopal, 2021; Jung et al., 2022; Chen et al., 2016).

End-to-end trained FCNs are capable of producing an output of the appropriate magnitude with effective inference and learning. To develop the edges of buildings seen in distant sensing images and improve the bounds of segmentation masks, HED extracts edge features at an *Encoder* of a specific architecture. At the final DCNN layer, CRF is used to combine the responses in order to better localize object boundaries. Due to the little quantity of data included in medical datasets, both 2D and 3D CNNs have been examined for semantic segmentation in medical imaging, however, it is unclear whether one is superior (Crespi et al., 2022).

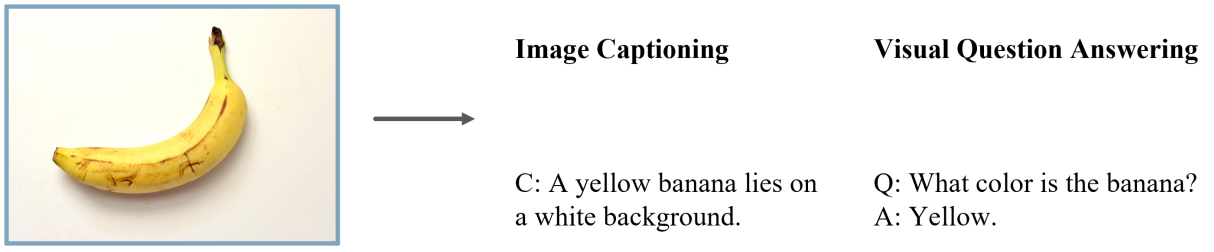


Fig. 2.2: Image Captioning and Visual Question Answering

Fig. *2.2* shows the difference between Image Captioning and Visual Question Answering using a textual example. These methods are explained in more detail below.

Image Captioning is the process of creating a written description for an image. There are several methods for captioning images, including reinforcement learning and attention techniques. Sharma proposes one technique, which is a combined bottom-up and top-down attention mechanism that allows attention to be calculated at the level of objects and other salient image regions (Sharma et al., 2018).
Researchers have also suggested reference-free measurement criteria for captioning images, including CLIPScore (Hessel et al., 2021).
The last reinforcement learning method that has been proven to be successful in improving image captioning systems is self-critical sequence training (Rennie et al., 2016).
Adding CLIP encoding as a prefix to the caption, and including emotions and feelings into the caption production process, improve captioning for low-resource languages utilizing English caption datasets (Mokady et al., 2021; Blandfort et al., 2016).
A model architecture based on Inception-ResNetv2 for image-feature extraction and Transformer for sequence modeling had the highest performance at evaluating several image captioning models using the Conceptual Captions dataset (Sharma et al., 2018).

Visual Question Answering (VQA) is a task that requires both language and visual awareness. It entails answering questions about an image in regular language. There are several methods for VQA, such as co-attention models, multimodal pooling, and attention processes (Anderson et al., 2017; Fukui et al., 2016; Zhou et al., 2019).
A-OKVQA, which relies on common sense and general knowledge to respond to questions (Schwenk et al., 2022), which concentrates on responding to inquiries about remote sensing images (Zheng et al., 2022b), are two current datasets for VQA.

2.2 Data Annotation Job

In order to make data easier to comprehend and analyze for computers, the Data Annotation Job (DAJ) process involves labeling the data with pertinent tags. Data annotators must provide the most precise labels for this data, which might take the form of images, text, video, or audio.

Data labeling is the act of adding markings to videos and images, such as text or objects, to make them traceable and recognizable by CV. This allows AI models to be taught to make accurate predictions using ML techniques. Labeling is the process of adding pertinent tags or information to texts in order to increase their meaning and ability to be understood by robots. Typically, texts and images are labeled, though annotation is now also used for the same purpose, and labeling is done to aid in the training of ML algorithms. There is relatively little difference between DA and data labeling other than the style and kind of content tagging that is used. As a result, they are routinely used interchangeably to create ML training datasets, depending on the AI model and training method (Kumari, 2023).

Relevance of the Data Annotation Job

The DAJ is critical to the development of ML models that can recognize input patterns. Annotated data is essential for training ML algorithms to detect input patterns. The DAJ result is crucial for the development of high-quality ML models (Potter, 2023b).

Historically, independent employees on crowd-work platforms like Amazon Mechanical Turk, Appen, or Clickworker have been reliant on and responsible for the manual DA and labeling tasks (Gray, 2019; Martin et al., 2014). Private annotation businesses, on the other hand, have grown in number and are focused on offering data labeling and annotation services. These are third-party businesses that contract out large amounts of annotation and labeling work to full-time employees. Many thousands of people work as data annotators for annotation companies (Wang et al., 2022a).

According to Settles, *Active Learning* (AL) uses humans as oracles to annotate unlabeled data while maintaining control over the learning process (Settles, 2009). A student asks an oracle (who serves as a teacher) to label a set of chosen instances that are not obvious and that will give information useful to the learning process in the AL ML technique. As a result, the learner becomes more effective at learning while utilizing fewer training

instances. The domains of application of AL are often ones where the cost of annotating data is significant, but these are jobs that people typically excel at, such as image interpretation or natural language processing (Mosqueira-Rey et al., 2023).

Human-in-the-loop is a technique of AL. In the technique, the learning algorithm is controlled by a human expert. This technology is often used in interactive simulation models in aerospace, vehicle handling, and robotics. In such simulations, humans play an essential role by influencing the simulated environment through their actions. The human-in-the-loop technique is an excellent way to ensure the quality of the annotated data while minimizing the cost of annotation (Potter, 2023a).

An ML application requires labeled data sets for the machine to understand the input patterns. To train CV-based ML models, the data must be precisely labeled using appropriate tools and techniques. Different methods of data labeling are used for various applications such as autonomous driving, construction site analysis, and object recognition in satellite and drone images. The accuracy of the results is improved with more labeled data sets, which contributes to a better experience for the end users (Mosqueira-Rey et al., 2023).

Since the effectiveness and accuracy of SL models depend on the type and quantity of annotated data, annotated data is essential to their operation. Machines do not have the same visual capabilities as humans. The varied data kinds are machine-readable through DA. One of the main obstacles to developing precise machine-learning models is finding high-quality annotated data. It helps computers detect and predict future trends utilizing data sets, as well as find and compare specific patterns.

Through the use of DA services, AI can be applied in real-world situations because the outputs will be accurate in the same way that such models are trained. The accuracy will increase the more image-annotated data is utilized to train the ML model. The ML algorithm will learn numerous sorts of factors from the range of data sets used to train it, and it will use this knowledge to use its database to produce the most appropriate outcomes in various circumstances (Karatas, 2023).

Data Annotation Techniques

Depending on the application of the ML and the used data, different DA approaches can be utilized. Several common variants shown in Fig. *2.3* are considered in more detail.

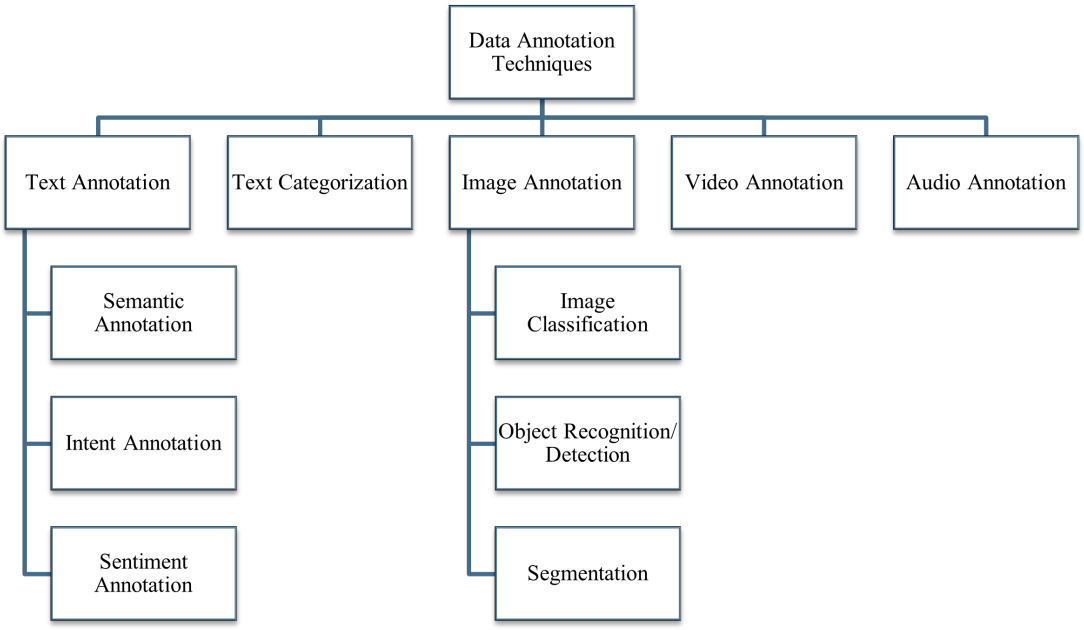


Fig. 2.3: Data Annotation Techniques (based on Karatas, 2023)

Like Fig. *2.3* shows, there are techniques for image and video data, as well as for text and audio data. Since the focus of this thesis is mainly on visual data, only image annotation and video annotation will be discussed in detail. Text annotation and audio annotation are listed only briefly.

Text Annotation is a method for highlighting important information in text data. Text annotation may be done in several ways, both manually and with the use of NLP (Stenetorp et al., 2012).
New methods for text annotation using Wikipedia entities have also been proposed. Techniques for cross-lingual text annotation have been assessed to enhance the annotation of multilingual text data (Makris and Simos, 2014; Zhang et al., 2013).

Audio Annotation is a method for detecting audio data that has significant information. Audio annotation techniques are useful for marking, recording, storing, and transmitting data from audio files (Lin and Nwe, 2021). Audio annotation can be done manually or automatically, and numerous approaches have been developed to increase audio annotation efficiency and accuracy.

NEAL is an open-source audio annotation tool that provides a reactive environment in which users may quickly annotate audio files and edit parameters that affect the related user interface components (Dutta and Zisserman, 2019).
Overall, audio annotation approaches seek to increase the efficiency and accuracy of labeling audio data with important information, which is critical for a variety of applications including audio retrieval and management (Karavellas et al., 2019).

Image Annotation is the technique of tagging images with pertinent information. The effectiveness and precision of image annotation have been enhanced by the development of automatic image annotation tools. An analysis of the many kinds of image annotations, including human, semi-automated, and automatic annotations, has been done. Techniques for context-based image annotations have also been reviewed. The goal of improving image annotation is to bridge the semantic gap between low-level visual data and high-level semantic ideas. Image annotation has emerged as a fundamental research issue in the fields of CV and multimedia.
In order to forecast suitable keywords for a new image, automatic image annotation is utilized, which aids in image retrieval by supplying semantic keywords for search. In general, image annotation approaches work to increase the precision and efficacy of labeling images with pertinent data, which is crucial for several applications including image management and retrieval.
With the help of the prepared annotated images for Image Classification, the machine first learns from annotated images to determine what each image represents. A further variation of image classification is object recognition or detection. It accurately describes the quantities and precise locations of the image's entities. In contrast to image classification, which labels the entire image, object recognition identifies items individually. Using object recognition, specific entities in an image, like a bicycle, tree, or table, are identified individually. Semantic segmentation labels an object in an image based on their shared properties, while instance segmentation identifies and labels each individual entity, and pan optic segmentation combines both semantic and instance segmentation to provide comprehensive labeling of all objects in the image (Pagare and Shinde, 2012).

Video Annotation is the technique of annotating videos with useful information. Video annotation techniques are critical for video content analysis and retrieval, especially as big multimedia archives grow in size. To increase the efficiency and accuracy of video annotation, automatic video annotation systems have been created. Video annotation is a complex procedure that necessitates a huge database, memory, and processing time (Pagare and Shinde, 2012; Randive and Mohan, 2020).
For video annotation, several strategies may be utilized, including manual, automated, and semi-automatic procedures. For 360° movies, interactive annotation techniques have been developed, allowing regular video editing techniques to be used to add content to immersive videos. Overall, video annotation approaches seek to increase the efficiency and accuracy of labeling films with important information, which is required for a variety of applications such as video retrieval and management (Khurana and Chandak, 2013; Meira et al., 2016).

Image data may be labeled in a variety of methods, including structured label(s), image annotations, image segmentations, etc., and it can be concluded in general. More frequently, a free-text report, an expert consensus based on a misinterpretation of the images, or the application of the imaging diagnostic are employed (Hwang et al., 2019).

Image Data Annotations in Manufacturing

There are various types of annotating image or video data shown in Fig. *2.4*. This DA include bounding box annotation, polygon annotation, semantic segmentation, landmark annotation, polyline annotation, and 3D point cloud annotation. Each method of image annotation has its use and application. Bounding boxes are the most commonly used method of image labeling because they can be applied to almost any object. Line annotation creates lines and splines to define boundaries, while polygonal segmentation uses complex polygons to more accurately determine the location and boundaries of an object than bounding boxes. In semantic segmentation, each pixel of an image is assigned a designation based on semantic information, while when landmarks are labeled, dots are created in an image to identify objects. Labeling 3D cubes is similar to bounding boxes, but offers more depth (Potter, 2023b).

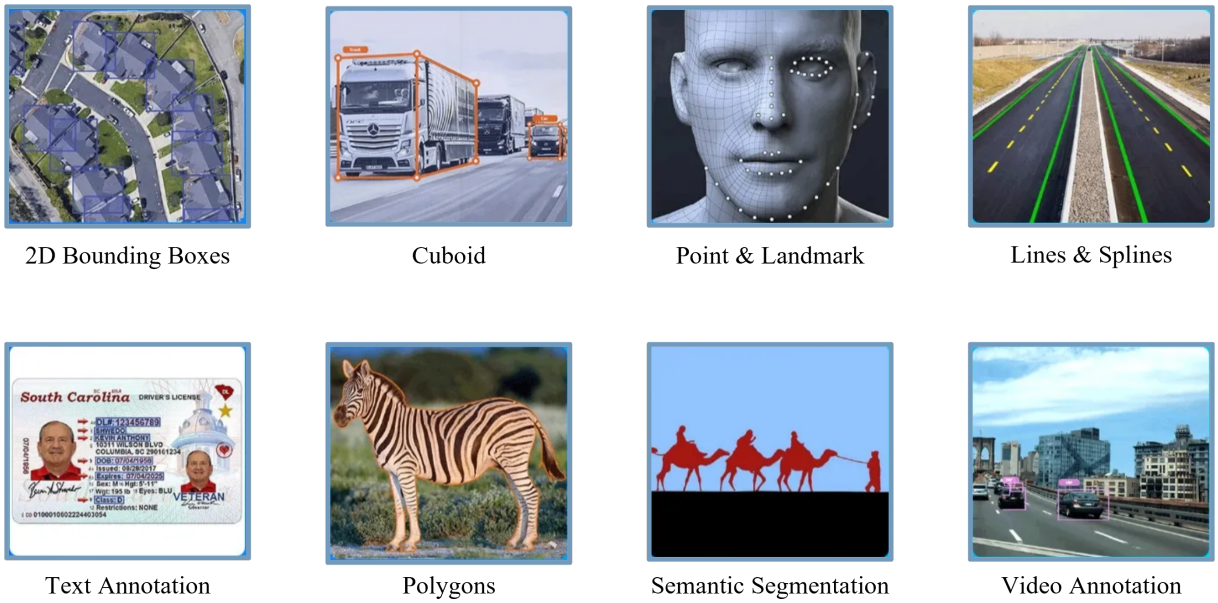


Fig. 2.4: Types of Image Annotations (based on Potter, 2023b)

The process of annotating industrial data, such as production images, maintenance data, safety data, and quality control, is known as industrial DA. Annotating such data accurately may be used to create models capable of identifying irregularities in manufacturing processes and assuring worker safety. It helps enterprises to make use of ML technology to streamline processes, uphold high standards, and keep an edge in an increasingly competitive marketplace (Karatas, 2023).

Tab. *2.1* shows different methodologies for the implementation of the DAJ. They were listed according to their degree of scale. The description and the area of usage were taken into account to provide a uniform overview. In the following, the focus is put on SSL.
In the creation and implementation of the DAC, the best possible scalability is aimed. The Scaling Degree has a value from zero to three.

- 0: no scaling
- 1: low scaling
- 2: medium scaling
- 3: very high scaling

Tab. 2.1: Data Annotation at Scale (based on Lakshmanan et al., 2021)

Method	Description	Common Usage	Scaling Degree (0 - 3)
Human Labeling	The traditional definition of DA is the process of labeling data with human effort, for example using a Folder structure or metadata tables.	All AI applications	0
Active Learning	A modest quantity of labeled data is used to train models in AL, after which the model chooses the most useful examples for expert labeling.	All AI applications	1
Crowdsourcing	The practice of crowdsourcing for DA involves assigning jobs for DA to a sizable group of individuals, generally via an online platform, to acquire high-quality annotated data at scale.	All AI applications	1
Model-assisted labeling	Model-assisted labeling is a method for classifying data more quickly by using ML models.	CV	1
Outsourcing	Companies that require substantial volumes of annotated data for ML and AI applications may find it helpful to outsource the DAJ.	All AI applications	1
Voting	Multiple annotators can submit their labels for a particular data point, and the best label can subsequently be chosen via a voting mechanism.	All AI applications	1
Domain Adaptation	The process of domain adaptation allows ML models that have been trained on one domain to be applied to another domain where the data may have distinct properties.	CV	2
Noisy student	Noisy student describes a technique for developing a model utilizing noisy or insufficient data.	CV	2
Semi-Supervised Learning	The ML technique is known as SMSL involves training a model using both labeled and unlabeled data.	CV, Data Mining	2
Self-Supervised Learning	SSL is a method that lets machines learn from data without human annotation.	NLP, CV	3
Unsupervised Learning	With USL, an ML model is trained on a dataset devoid of any labeled data.	CV	3

2.3 Self-Supervised Learning

The term SSL was initially used in robotics, where labels were applied to training data automatically to exploit the correlations between input signals and sensors (Ohri and Kumar, 2021).

The way newborns educate inspires the idea of SSL. Infants learn by observation, common sense, their environment, and minimal contact. All of these factors contribute to their ability to learn on their own. The environment of newborns becomes a source of supervision for them, which aids in their comprehension of how things operate in the absence of continual monitoring. The same concept is duplicated in the machine through SSL, where the data supervises itself for training the model rather than having annotated labels that advise the network on what is correct or incorrect (Orhan et al., 2020).

SSL is a paradigm of ML where a model creates data labels autonomously when there is unstructured data as input. To train the model in subsequent rounds with backpropagation, similar to any other SL model, the model uses the highly reliable data labels created among the generated data. The only difference is that the data labels used as basic truths in each cycle are changed. NLP, CV, and robotics have all seen promising benefits from SSL. Recent studies have focused on building more efficient SSL strategies such as contrastive learning and transformer models that have reached the state of the art in a variety of tasks (Feng et al., 2023; Ruyi et al., 2023).

Relevance of Self-Supervised Learning for Computer Vision

SSL is considered to be the bridge between SL and USL. By lowering the requirement for human labeling of data, SSL can be useful in the DA task. SSL can assist overcome this issue in CV because labeled data can be expensive and time-consuming to gather, as elaborated in *section 1.2*. Additionally, by enabling models to learn from a bigger and more varied set of data, SSL can increase the accuracy of models (Doersch and Zisserman, 2017).

The application of SSL in CV, particularly in the context of DA, has the potential to overcome these challenges. SSL allows models to learn from a larger and more diverse set of unlabeled data. This not only mitigates the scarcity of labeled data, but also improves the generalization and robustness of the models.

In addition, SSL methods facilitate the creation of pretext tasks that encourage models to

learn useful representations from unlabeled data. Thus, models develop a comprehensive understanding of the underlying structures and patterns in the data, which can then be applied to downstream tasks such as DA in CV.
SSL methods can leverage the wealth of unlabeled data available in many industrial contexts. While this data is not labeled for specific tasks, it can still provide valuable information for training models. By using SSL, the potential of this unlabeled data can be fully exploited, leading to improved performance and accuracy in CV applications (Oleszak, 2023).

SSL is a type of ML in which a model learns to predict data without explicit instruction. It has been employed in a wide range of applications, including voice recognition, video representation learning, federated learning, and recommender systems. SSL can be used to learn meaningful data representations without requiring labeled data, which can be costly and time-consuming to collect. A single model may be trained by merging many SSL tasks to improve performance (Doersch and Zisserman, 2017).

In general, SSL can be divided into *Contrastive Self-Supervised Learning* and *Generative Self-Supervised Learning* as shown in Fig *2.5*.

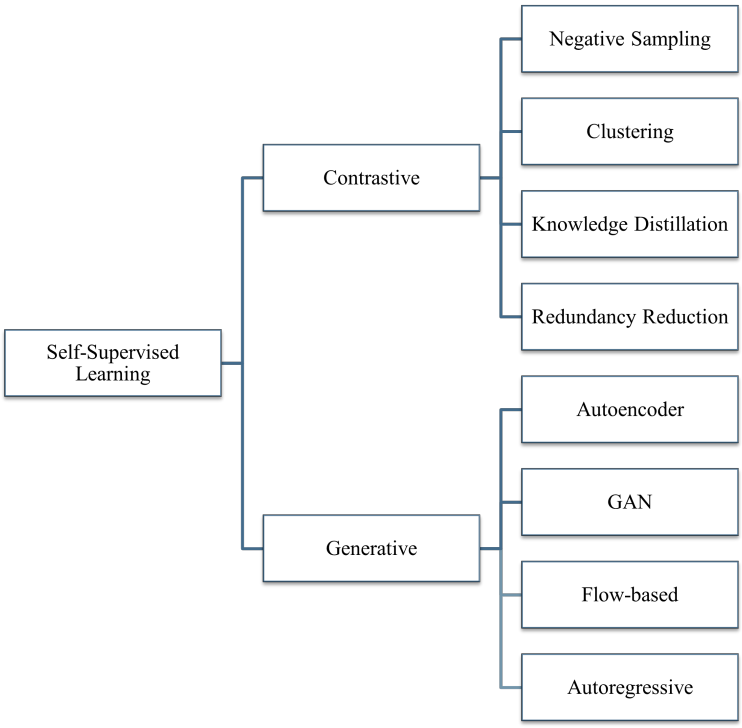


Fig. 2.5: Taxonomy of Self-Supervised Learning (based on Liu et al., 2021a)

Contrastive SSL is a type of SSL method in which a model is trained to perform a classification job made up of unlabeled data. It has lately emerged as one of the most influential learning paradigms in the absence of labels across a wide range of domains, including brain imaging, text, and graphics (Liu et al., 2021a).
Contrastive learning uses several perspectives on the same input to create representations that generalize to numerous downstream circumstances, such as global representations for tasks like classification or local representations for tasks like detection and localization (Ma et al., 2021; Tan et al., 2020)

Contrastive predictive coding is a technique that learns to forecast a data point's future representation based on its prior representation (Liu et al., 2021a). Contrastive SSL has been used in a variety of applications, including CV and NLP (Liu et al., 2021b; Bachman et al., 2019).

In Generative SSL, the objective is to develop a generative model that can create fresh data samples that are comparable to the training data without any further explicit supervision. The generative model learns to represent the underlying distribution of the data by being trained on the input data alone, without any labels. The learned representations may be used for several downstream tasks, including object detection, segmentation, and classification (Liu et al., 2021b; Bachman et al., 2019).
Techniques for Generative SSL include autoregressive, flow-based and auto-encoding models as well as GANs. Autoregressive models are used to generate data by modeling the conditional probability of each data point given the prior ones. Flow-based models use invertible transformations to convert a straightforward distribution into a complicated one. Auto-encoding models acquire the skills necessary to encode input data into a lower-dimensional latent space and then decode it to return it to the original space. To leverage their respective strengths, hybrid generative models mix multiple generative models, such as in NLP and CV (Sutter et al., 2021). GANs combine a generator and a discriminator neural network. While the discriminator tries to differentiate between the created samples and the genuine ones, the generator provides new data samples that are comparable to the training data (Thada et al., 2023). Generative SSL approaches of autoencoders in the CV context will be considered in more detail in section *2.3*.

Application of Self-Supervised Learning

Pretext and downstream tasks are the two divisions of SSL tasks. The pretext uses SL, where labels are created from the data itself, to learn representations. After this learning is finished, the model applies the previously learned representations to the ensuing tasks after this learning is finished (Rani et al., 2023).

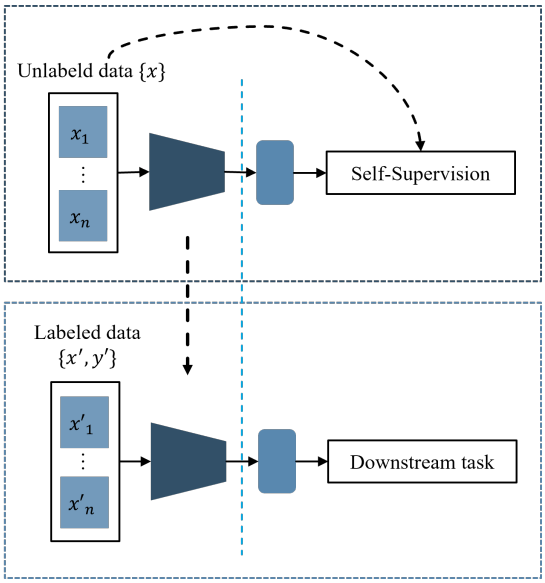


Fig. 2.6: Self-Supervised Learning for Pretext Tasks (based on Wang et al., 2022a)

Pretext tasks, often referred to as secondary tasks, allow the model to pick up important feature representation knowledge that is then applied to subsequent tasks. Feature representation learned in pretext tasks should be computed to guarantee quality in subsequent challenges. Primary tasks known as downstream tasks establish the goal of the model shown in Fig. *2.6*. The main downstream objective is to execute semantic segmentation, action identification, and classification or object detection with inadequate data labels. There are two methods for down streaming: fine-tuning and utilizing a linear classifier. Performance on the downstream task is often better when the domain gap between the SSL pre-training and the downstream task is narrower (Wang et al., 2022a). This procedure can be ensured in the form of a .pth file. The model of the pretext task is saved as a .pth file and embedded in the downstream task. A serialized PyTorch state dictionary is often found in files with the `.pth` suffix. The state of a PyTorch model, comprising the model's weights, biases, and other parameters, is stored in a Python dictionary called a PyTorch state dictionary (Lynn, 2023).

SSL in the context of Computer Vision

SSL and CV are closely linked. Different SSL techniques make it possible to train models without manually annotated data, which is of great importance in CV. Together, they enable the development of algorithms and models that can process visual information from images and videos. Recently, much of the research has focused on developing SSL methods in CV across different applications. The ability to train models with unmarked data solidifies the entire training process and enables the model to learn the underlying semantic functions without introducing label distortion.

Methods of Self-Supervised Learning

In SSL there could be different methods found, like *Colorization*, *Image Rotation*, *Patch Positioning*, Masked Autoencoder. Colorization, Rotation, Patch Positioning were excluded from this thesis for the following reasons. They have limited applicability, lack of robustness, and high accuracy requirements in industrial manufacturing. The rotation of images and the generation of data by *Data Augmentation* can also be a challenge due to limited relevance and diversity. The self-supervised method of colorization can be of limited usefulness due to unpredictable conditions and the dependence on colors. The context dependency of colors in industrial manufacturing requires a more accurate method based on actual color information (Vondrick et al., 2018a; Atsuyuki et al., 2022; Mundhenk et al., 2017).

The present thesis focuses on MAEs. This decision was taken for the following reasons. This methodology ensures the effective processing of large data sets. It is also possible to identify complex contexts and correlations. A further advantage is that the adaptation to different image conditions is very good and can thus be used for a broad spectrum of image data. This results in good expandability and flexibility because individual adjustments can also be made (Kaiming et al., 2021).

The addition of color to a grayscale image is referred to as colorization. There are several techniques for coloring images, including classic reference-based systems that rely on external color images and more modern DL techniques that can color images autonomously. DL methods may convert a grayscale image and sparse, local user hints to an output colorization with a CNN under human supervision (Zhang et al., 2017).

In SSL, colorization can be used to develop meaningful representations without explicit supervision. In one study, for instance, colorization was used as a tool for visual USL, and the networks that were trained on colorization from scratch performed well on other visual tasks. In order to enhance the performance of a single SL task, such as semantic segmentation in applications for autonomous vehicles, colorization can also be employed as an SSL task (Novosel, J. and Viswanath, P. and Arsenali, B., 2019). These approaches can be used in various applications, such as image and video editing, restoration, and colorization of historical images.
Unforeseen circumstances in industrial manufacturing may make colorization difficult, and it may be challenging to gather enough unwritten data for the model's training. While it's crucial to capture visual data with high accuracy, the SSL method of colorization may not always be accurate. Colors can provide significant information, however, their lone use in the creation of color information might result in poor decisions (Vondrick et al., 2018b; Wang et al., 2022a; Treneska et al., 2022).

The act of rotating an object or an image around an axis is referred to as rotation. The capacity to mentally rotate mental images of objects is known as mental rotation. Since objects are frequently distributed with arbitrary orientation in aerial images, rotation in CV can be used to enhance object detection (Shepard and Metzler, 1971)
A novel rotation-based framework has been proposed for arbitrary-oriented text detection in natural scene images. In addition, some mathematical techniques used to approximate the internal structures of rotationally distorted stars are limited by physical assumptions or computational difficulty (Shepard and Metzler, 1971)
Without utilizing any human-labeled annotations, rotation may be utilized as an SSL task to learn spatiotemporal video properties. Additionally, colorization can be a means of visual SSL, where the networks developed through training colorization from scratch are highly generalizable to other visual tasks. In order to clearly express rotation equivariance and rotation invariance and effectively forecast orientation while minimizing the complexity of modeling orientation fluctuations, rotation can also be used in object detection in aerial images (Han et al., 2021). The rotation of images is frequently pointless during industrial production, since crucial details are independent of the direction.

The use of data augmentation to expand the data set may be constrained due to the limited variety of potential perspectives. The self-supervised rotation method may be inaccurate since it just relies on rotation angle predictions and provides no actual information about the object. But in order to make significant decisions based on visual data, high accuracy in the visual data collection is frequently required (Atsuyuki et al., 2022; Gao et al., 2022; Zheng et al., 2022a).

Context prediction, which entails anticipating the context of a given input, is one method of SSL. For instance, by utilizing the natural supervision offered by the data itself, a unique SSL approach for learning graph representations is given, where the global context of each node is made up of all the nodes in the network (Peng et al., 2020).
Given that the relative positions of image components may not be significant, the method of image patch positioning has limited applicability in industrial manufacturing. Additionally, they may be haphazard in the face of difficulties like poor lighting or distortions. In the industrial manufacturing process, high visual data capture accuracy is essential. However, the SSL method of image patch positioning can be inaccurate and provide no real information about the object that can be recognized (Rani et al., 2023; Fuadi et al., 2023; Mundhenk et al., 2017).

Masked Autoencoder Architecture

An MAE is a type of SSL model that is used for unsupervised feature learning. It is a neural network that is trained to reconstruct an input image from a partially masked version of the same image. The model is trained to predict the missing pixels in the masked image, which forces it to learn useful features that can be used for downstream tasks such as image classification. An MAE is a type of neural network used for distribution estimation and generative modeling. It is a modification of the traditional autoencoder that masks the autoencoder's parameters to respect autoregressive constraints, meaning that each input is reconstructed only from previous inputs in a given ordering. This allows the autoencoder outputs to be interpreted as a set of conditional probabilities, and their product, the full joint probability. MAEs have also been used in SSL for vision and beyond (Zhang et al., 2022).

The aim here is to generate a prediction of the pixel values of an unknown patch in the image based on the overall context of the image using encoder decoders as shown in Fig. *2.7*.

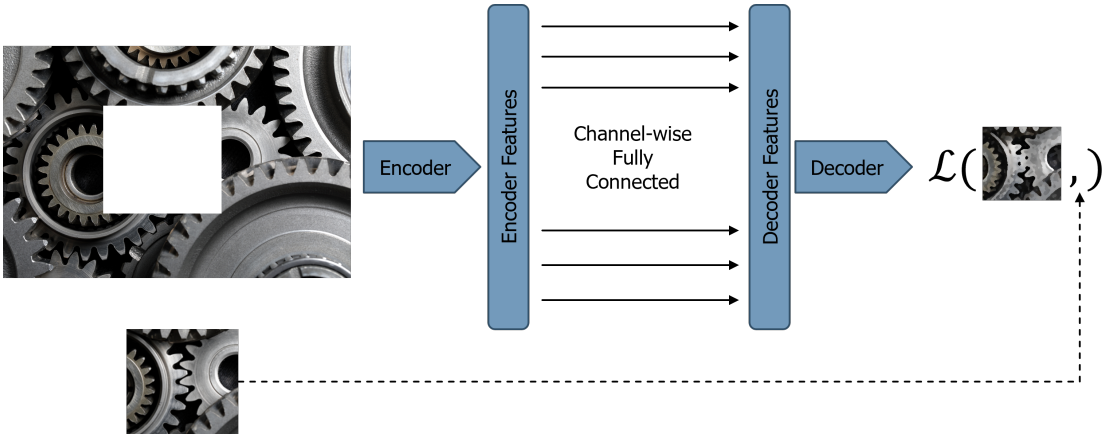


Fig. 2.7: Masked Autoencoder Architecture (based on Yu and Chen, 2022)

To achieve this, an encoder-decoder architecture is trained for a vortex task. Here, the encoder generates a *Latent Features* representation of the input image with hidden areas. The *Decoder* estimates the missing image area based on the reconstruction loss based on the latent feature representation. A channel-wise completely connected layer between the encoder and decoder enables each unit in the decoder to be reproduced over the entire image content. Terms like adversarial loss, joint loss, and reconstruction loss are frequently used in relation to DL and image processing.

The discrepancy between the original image and the rebuilt image is referred to as the reconstruction loss. It serves as a gauge of how effectively the model can recreate the input image. The mean squared error between the original and reconstructed images is commonly used to assess the reconstruction loss.

On the other side, *Generative Adversarial Networks* (GAN) make advantage of adversarial loss. GANs are made up of two networks: a discriminator network that tries to tell the difference between genuine and false images, and a generator network that creates fictitious images. The loss function used to train the discriminator network is adversarial loss. The measure of how effectively the discriminator can discriminate between actual and fraudulent images is commonly a binary cross-entropy loss.

An adversarial loss L_{adv} combined with a reconstruction loss L_{rec} defines a joint loss L_{joint}. It is used to train a model that can produce high-quality images and tell the difference between authentic and false images (Yu and Chen, 2022).

Based on the context of the input image x, the binary mask M corresponding to the distant image region, and the encoder function F, the reconstruction loss is in charge of recognizing the relevant features. Its definition is the normalized masked distance between the encoder's output, $G(E(x))$, and the input image, x.

$$\mathcal{L}_{rec}(G, E, x) = ||x - G(E(x))||_1$$

where G is the generator network, E is the encoder network, x is an input sample, and $|| \cdot ||_1$ denotes the L_1 norm.

On the other hand, the adversarial loss aims to learn the latent space of the input data and to make the output image seem realistic. The discriminator network D uses the external discontinuity in the patched areas and the original context to discriminate between genuine and produced images while the generator network G is conditioned on the input mask.

$$\mathcal{L}_{adv}(G, D) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

where G is the generator network, D is the discriminator network, x is a sample from the real data distribution $p_{data}(x)$, z is a noise vector sampled from a prior distribution $p_z(z)$, $G(z)$ is the generated sample from z, and E denotes the expected value.

$$\mathcal{L}_{joint}(G, D, E, x) = \mathcal{L}_{rec}(G, E, x) + \lambda \mathcal{L}_{adv}(G, D)$$

where $\mathcal{L}_{rec}(G, E, x)$ is the Reconstruction Loss, $\mathcal{L}_{adv}(G, D)$ is the Adversarial Loss, λ is a hyperparameter that controls the relative importance of the two losses.

The SSL method is used for semantic inpainting through auxiliary support and learning from strong feature representations (Baevski et al., 2022). The goal is to achieve a decreasing loss function to increase the accuracy of the following CV task by the pretrained model.

2.4 Machine Learning Operations

The DAJ plays an important role in the ML lifecycle. It enables data enrichment and labeling to train and improve models. As AI has progressed, the motivation to scale the DAJ process has also increased, especially with respect to product development of AI systems. In this regard, several issues arise that accompany the scaling of machine learning operations. To address the requirements, MLOps methods and software have emerged as solutions to reduce the technical debt that can arise from the deployment of ML products. These requirements are also reflected in the definition of MLOps. *"MLOps is a collection of methods and technologies to enhance the efficiency of machine learning model development as well as of operational usage of products based on those ML models."* (Kappel, 2023)

Relevance of MLOps

MLOps also helps solve the model packaging and validation challenge by supporting model portability across a variety of platforms and ensuring model performance meets functional and latency requirements. Finally, MLOps aims to solve the challenge of model deployment and monitoring. Models are released and monitored with confidence to know when they need to be retrained by analyzing signals such as data drift. By addressing these key challenges, MLOps enables organizations to use and manage ML models with confidence to ensure they meet legal requirements and function effectively in production environments. It is an essential practice for any organization looking to unlock the full potential of its ML capabilities (Microsoft, 2023).

Overall, MLOps is a specialty that requires a deep understanding of both ML and *DevOps* practices. By addressing the unique challenges of deploying and managing ML models, MLOps teams can help organizations realize the full potential of their data.

One of ML's biggest challenges is to ensure that models are reproducible and can be versioned over time. As ML becomes increasingly important in industries such as healthcare and finance, it is imperative to maintain asset integrity and maintain access control protocols. MLOps helps address this challenge by enabling teams to certify that model behavior meets regulatory and adversarial standards, and by ensuring that models are verifiable and explainable (Microsoft, 2023).

MLOps as an extension of DevOps

Various software process models and development approaches have previously emerged in the field of software engineering. Waterfall and the Agile Manifesto are two prominent examples (Beck et al., 2001).

MLOps is a methodology, based on DevOps, which improves the collaboration between data scientists and operations professionals. Applying this methodology helps teams to deploy machine learning models in large-scale production environments much faster and with much better results. Likewise, it guarantees automation through continuous integration, continuous delivery, and continuous deployment (CI/CD), enabling rapid, frequent, and dependable releases. It is also intended to guarantee continuous testing, quality assurance, monitoring, logging, and feedback loops (Leite et al., 2020).

Components within MLOps

Data collection, data preparation, model training, model validation, and monitoring are all processes in the ML lifecycle. Within the data preparation step, the DAJ supports the ML model train, which is an important element of the data preparation step. In this context, DAC is to be developed that is embedded in the functioning of MLOps.

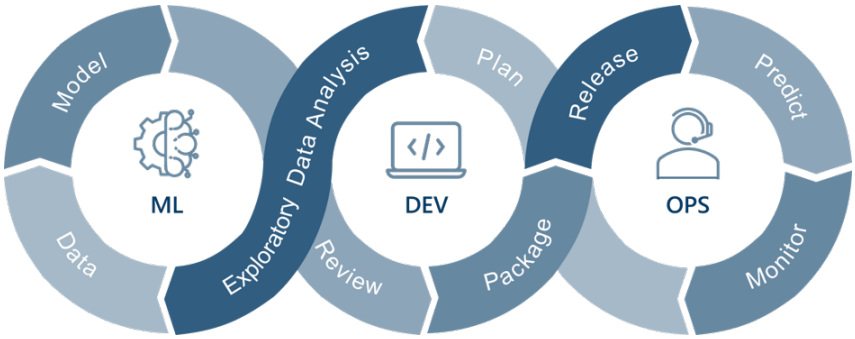


Fig. 2.8: Machine Learning Operations (based on Merritt, 2023)

MLOps is a critical aspect in deploying and managing ML models in production environments, illustrated in Fig. *2.8*. With this approach, every assessment in the ML lifecycle shown in Fig. *2.9* can be effectively tracked, versioned, audited, certified, and reused to ensure that everything runs smoothly and efficiently (Jazmia, 2023).

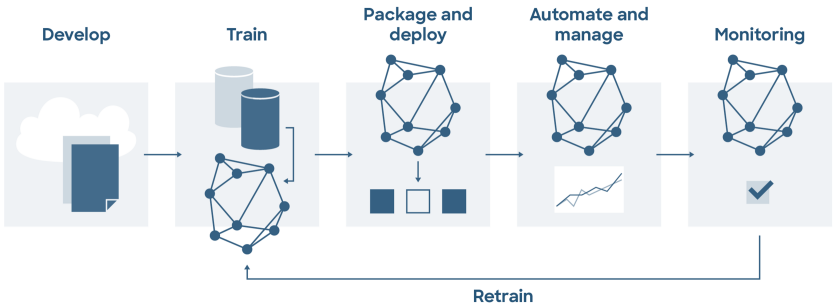


Fig. 2.9: Machine Learning Lifecycle with Azure ML (based on Microsoft, 2023)

The three factors people, process, and platform should be taken into consideration in order to properly implement ML. Individual engineers must integrate their work into a shared repository in terms of people, since this is crucial. Because every update to the code is routinely tested, errors are found more quickly. Innovation is also accelerated through the exchange of code, data, models, and training pipelines.
In terms of process, templates should be made available to hasten the creation of the infrastructure and the models. Efficiency may also be improved by automating the entire procedure from code change to production.

Scaling in the context of MLOps

Finally, it's critical to employ the right platform in order to efficiently and reliably supply functions to clients. Problems may be immediately identified and fixed by monitoring production pipelines, infrastructure, and goods (Kreuzberger et al., 2022).

Tab. 2.2: Maturity Levels in MLOps (based on Kreuzberger et al., 2022)

Maturity Level	Training Process	Release Process	Integration into app
Level 0	Untracked, file is provided for handoff	Manual, hand-off	Manual, heavily DS driven
Level 1	Untracked, file is provided for handoff	Manual, hand-off to SWE	Manual, heavily DS driven, basic integration tests added
Level 2	Tracked, run results and model artifacts captured repeatably	Manual release, clean hand-off process, managed by SWE team	Manual, heavily DS has driven, basic integration tests added
Level 3	Tracked, run results and model artifacts captured repeatably	Automated, CI/CD pipeline set up, everything is version controlled	Semi-automated, unit and integration tests added, still needs human signoff
Level 4	Tracked, run results and model artifacts captured repeatably, retraining set up based on metrics from app	Automated, CI/CD pipeline set up, everything is version controlled, A/B testing has been added	Semi-automated, unit and integration tests added, may need human signoff

Tab. *2.2* shows different maturity levels in the area of MLOps, which differ in terms of the maturity level of the training process, release process, and integration into the app.

Level 0 denotes the lowest maturity level without MLOps at which no tracking of the training process takes place and the model is manually transferred to the development team. Integration into the app is done manually and is driven by the data scientists.
Level 1 describes the maturity level at which the training process is still untracked, and the model is manually transferred to the software development team. However, basic integration tests are carried out.
Level 2 is characterized by the fact that the training process is now tracked and the results and model artifacts are recorded repeatably. The release is done manually with a clean handoff process managed by the team. Integration into the app will continue to be manual and driven by Data Scientists.
Level 3 is characterized by automation of the release process, where everything is versioned and set up in a CI/CD pipeline setup. Integration into the app is semi-automated and includes both unit and integration tests. However, human confirmation is still required.
Level 4 is the highest maturity level at which the training process is performed again based on metrics from the app. The release is done automatically with an established CI/CD pipeline setup that also includes A/B testing. Integration into the app is semi-automated and includes both unit and integration tests. However, human confirmation may be required (Kreuzberger et al., 2022).

Objectives and Key Results within MLOps

Objective and Key Results (OKR) represent a goal-setting process that allows teams and individuals to set rigorous, ambitious goals that lead to quantifiable results. It is a popular management technique used to set goals and monitor progress to foster collaboration and commitment to quantifiable objectives (Sparks, 2023).

OKRs are critical to MLOps because they effectively align goal setting, planning, measurement, and execution across the enterprise. ML engineers need to continuously work on their development to improve their skills and not lose existing competencies.

In a fast-paced development environment like ML, quality must not be compromised when scaling. Therefore, high standards must be maintained, especially when releasing new code. By using OKRs, engineers can ensure that they maintain the desired quality even as they grow rapidly.

Improving the quality of pipelines presents another challenge. There is always room for improvement when companies implement ML. Through OKRs, clear goals can be established. These build confidence in the company's security policy and enable teams to anticipate threats and identify solutions.

A data-driven culture is critical to the success of OKRs. Establishing measurable outcomes provides transparency and objectivity. Aligning the individual goals of MLOps teams with overall business goals can ensure progress and success.

In summary, OKRs are an essential tool for MLOps. They enable ML engineers to develop their skills and ensure quality. With clear objectives and data-driven culture, MLOps teams can effectively align their goals with the overall goals of the business and ensure the success of their projects (Kahansky, 2023; Rehman, 2023).

3 Investigation

This chapter presents a Data Annotation Concept that improves the DA procedure for ML models. Within the concept, SSL with an MAE is used to automatically extract relevant features from images without manual annotation. Based on the real AI product example of the CVT, within this chapter the design, application, and integration of the DAC is described. The DAC design includes data preparation, SSL with MAE, and model validation. The application section demonstrates the use of the DAC in various production processes using sample data sets. The integration section explains how the DAC is implemented in the CVT context. Which means, that the goal is to provide an scalable approach for the manual DAJ that replaces labor-intensive manual annotation and improves the use of ML models.

3.1 Data Annotation Concept Development

In order to achieve the objectives, the concept can be structured in three phases. 1. Data Preparation, 2. SSL Approach and 3. Validation, as shown in Fig. *3.1*. Based on the three phases, the implementation takes place in A. The four Concept Phases can be divided into seven components. These consist of the Concept Phase, Phase Description, Relevant Factors, Procedure, Attention Points, and Phase Result. A detailed overview of each block can be found in Tab. A.2 in the Appendix.

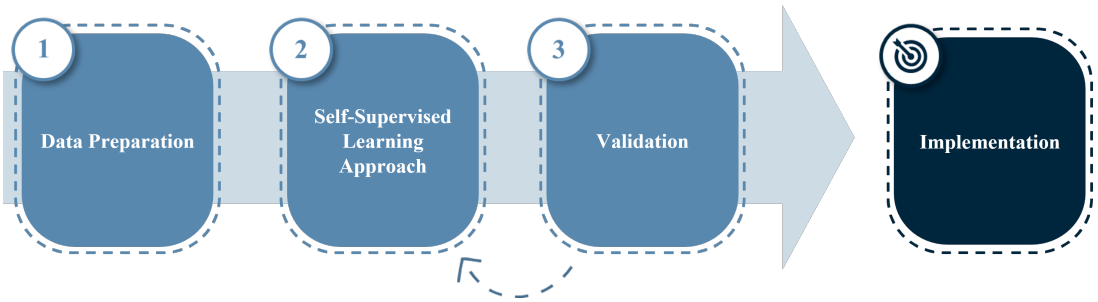


Fig. 3.1: Data Annotation Concept Idea

The concept of SSL using an MAE is aims for discovering features on its own without labeled input. The input image is encoded into a latent space representation using the autoencoder architecture, which is then used to decode it back to the original image. By randomly masking some of the input images, the MAE architecture forces the model to

acquire more reliable features that are independent of the entire image.
This idea aims to train a model that can acquire relevant characteristics without the requirement for manual labeling by making use of a large amount of unlabeled data.
The key benefit of this approach is that it lessons reliance on labeled data, which may be costly and time-consuming to gather. However, the model's performance might be sensitive to hyperparameters like the size of the masking region and the number of hidden layers, and high performance requires a sizable amount of unlabeled data for training.

1. Data Preparation

The first step the Data Preparation Step. This steps is criticial for improve data quality, establishing a consistent baseline, and prepare visual data for effective analysis. Noise cleaning, error corrections, and data standardization does help for more reliable and consistent results.
The visual data without DA, that acts as the concept's fundamental building block is processed and prepared in the first stages during the data preparation phase. The type of data, from natural images to manufacturing images, and the amount of data might vary depending on the application.
The preparation and preprocessing of the labelless visual data is the output of this step, laying the groundwork for additional processing and analysis in the following phases.

2. Self-Supervised Learning Approach

The Self-Supervised Learning Approach phase is necessary to train a model that independently learns features from the prepared data and exhibits improved performance. By providing a well-trained SSL method and parameters, this phase allows the model to be applied in different scenarios and creates opportunities for further improvements and adjustments.
The SSL method as the second step of the concept entails utilizing a particular SSL technique to train a model to learn features on its own. The loss function from the model will be returned after using the SSL technique, showing how well the trained model performed. The `.pth` file containing the well-trained SSL method and all parameter settings is the phase's output. This file can be applied in the future or used as a foundation for additional fine-tuning or downstream operations.

3. Validation

The Validation phase is necessary to verify the performance and efficiency of the trained model. By evaluating the loss function and assessing feature extraction and pattern recognition, this phase helps to ensure the quality and effectiveness of the model concerning the intended goals and requirements.

The pre-trained model is exported as a `.pth` file for use in other applications during the validation step of the MAE-SSL approach. The link between the loss function and the epochs is assessed in this stage.

Depending on the particular context and optimization objectives, the output of this phase is the evaluation of the loss function to obtain the smallest or highest value of the loss. This assessment aids in assessing the model's performance and efficacy in capturing the necessary features and patterns. Depending on the result of the loss value, it is necessary to go back to phase two. There, the necessary hyperparameters must be adjusted until the training is successful.

A. Implementation

A. Implementation is necessary to integrate the pre-trained model into the CV application and maximize its performance. By adapting to specific tasks and using the features already learned, this phase helps to create an accurate and efficient CV solution.

The pre-trained model must be incorporated into a CV application during the implementation phase. The `.pth` file, which contains the pre-trained model with its designated weights, is used in this stage. The model can be further adjusted for a downstream task or Transfer Learning, such as classification or segmentation, by adding extra layers and training on labeled data.

The pre-trained model is integrated into a CV application as a result of this phase. This entails incorporating the pre-trained model with predetermined weights from the `.pth` file into the CV application, to improve the model accuracy of the CV task.

Now all the components have been processed. Phases 1-3 can be assigned to the *preparation* section. The goal A. in combination with the model preparation fall under the section *doing*.

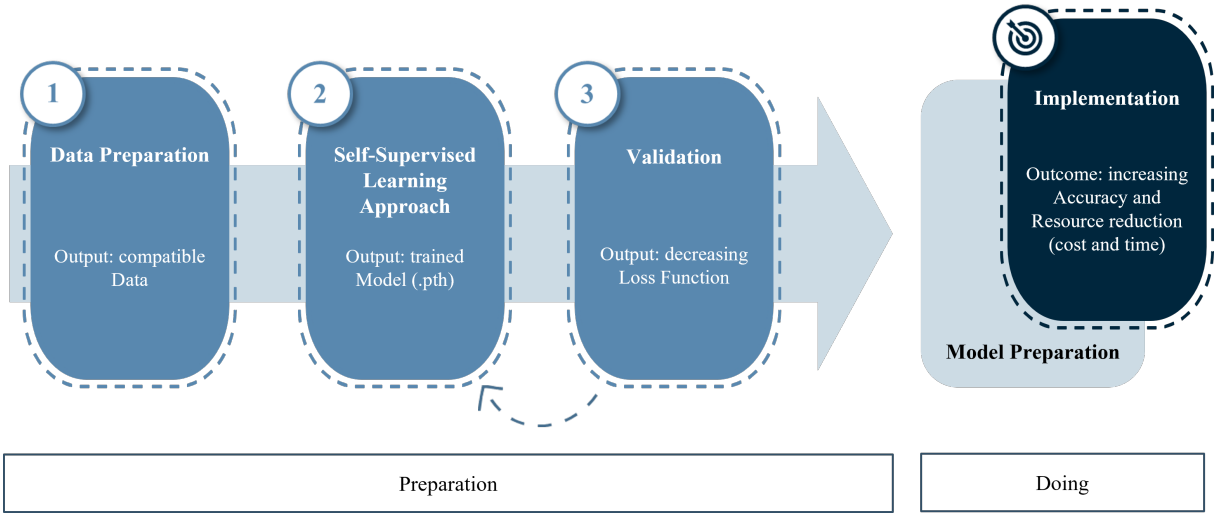


Fig. 3.2: Holistic Data Annotation Concept

From this, the holistic picture of the DAC is generated and is composed as follows. In Fig. *3.2* it is easy to see how the different components build on each other.

Phases 1-3 run through in the form of a process. *Phase 1* with its output is the input for *phase 2*. Depending on the use case the SSL approach implemented in *Phase 2* could be adjusted. In this thesis we will use the MAE approach as described in *section 2.3* described, we will use the MAE approach. The output of a trained model's parameter settings is exported in `.pth` format serves as input for *phase 3*, where the validation of the loss function and image congruence takes place. Up to a certain threshold, the performance of the model is increased with the help of hyperparameter tuning. From a certain threshold, the model can be transferred as output to A. Implementation.

Here, the output of the model preparation serves as input to define the architecture and parameter setting of the following CV downstream task.

This results in the following outcome of the A. Implementation. The pre-trained model could be integrated and with the help of the pre-trained image annotations, it is possible to improve the accuracy of the CV task. This means that the cost and resources for the DAJ in the CV task can be reduced.

3.2 Data Annotation Concept Validation

To validate the concept described in the former chapter, in the following the valdiation of the concept is being explained. Starting with the Data Preparation phase, followed by the SSL approach using the MAE, and finally validation of the loss functions and image congruence. The validation is being done based on three datasets from manufacturing side. The data is being collected within the CVT. The first dataset includes Pump Impeller images, the second dataset includes Sealing Boot Lip images and the third dataset include Brake Caliper images.

1. Data Preparation

In the thesis data preparation stage, the raw data is prepared for the best possible processing and analysis. To ensure high-quality and consistent data, data collection, cleaning, and formatting will be performed. There may be images in the dataset that are of poor quality, in the wrong data formats, or damaged. These must be removed in the cleaning step. The pixel size of the images must also be adapted to the available processing power via the formatting step and reduced if necessary.

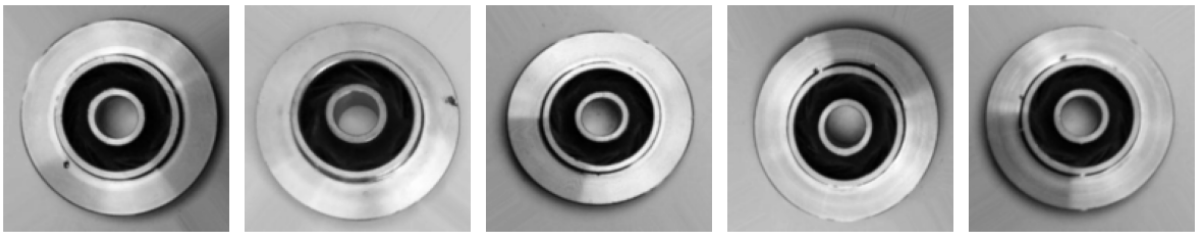


Fig. 3.3: Representative Images from the Pump Impeller Dataset

This dataset focuses on the manufacturing of cast products, specifically the Submersible Pump Impeller. It contains a total of 3137 images showing the impeller in Fig. *3.3*. Of these, 2875 are training images and 262 are test images. All images are in grayscale and have a resolution of 300x300 pixels in a `.jpg` format.

The casting process can cause various defects such as air pockets, pores, burrs, shrinkage defects, defects in the molding material, defects in the casting of the metal, metallurgical defects, and more. These defects are undesirable irregularities that can affect the quality and functionality of the cast products (Dabhi, 2023).

Fig. 3.4: Representative Images from the Sealing Boot Lip Dataset

The present dataset of Sealing Boot Lips contains 748 images, which are in `.png` format. Of these, 638 are training images and 110 are test images. The images shown in Fig. *3.4* were collected over several months and are stored in a database in production. Each image is in grayscale and has a resolution of 1500x1100 pixels.
The purpose of the dataset is quality control for the production of sealing sleeves. The images are used to identify and analyze possible defects or flaws in the sleeves. Various visual features such as cracks, irregularities, or damage are checked to ensure that the manufactured sealing collars meet quality standards.

Fig. 3.5: Representative Images from the Brake Caliper Dataset

The present dataset of Brake calipers consists of 568 images in `.png` format. Of these, 467 are training images and 101 are test images. The images shown in Fig. *3.5* were collected over several months and are stored in a database in manufacturing. Each image is in grayscale and has a resolution of 2448x2048 pixels.
The images in this dataset were captured specifically for quality inspection during the manufacture of brake calipers. They are used to indentifying and analyze potential defects or flaws in the calipers to ensure that the manufactured calipers meet the high-quality standards in the industry. The review of the images takes into account various visual features such as cracks, spalling, signs of wear, and uneven surfaces.

2. SSL Approach - MAE

The design and training process of the MAE model utilized for image reconstruction is described in this section. The MAE model is made to provide high-quality reconstructions by learning a condensed latent representation of the input images.

The PyTorch package is used to build the encoder and decoder components of the MAE model as sequential neural networks. The decoder reconstructs the input images from the latent space after the encoder maps them to a lower-dimensional latent space. The model utilizes fully connected layers for both the masking, encoding, and decoding steps.

The used GPU is the Standard_NC24. This virtual machine provides ample resources for high-performance computing. With 24 cores, 224 GB of memory, and a 1440 GB hard drive, it provides a robust infrastructure to handle demanding workloads. It is also equipped with four NVIDIA Tesla K80 GPUs, known for their parallel processing capabilities and suitability for tasks such as DL and scientific simulations. Several times, an adjustment of the hyperparameters were adjusted to achieve the best possible result. For good comparability, a hyperparameter setting was chosen that produced a relatively good performance for all data sets. This is shown in Tab. 3.1.

Tab. 3.1: Hyperparameter Settings of the Model Training

Parameter	**Value**
`num_epochs`:	100
`batch_size`:	64
`learning_rate`:	0.0001
`latent_dim`:	512
`num_hidden_layers`:	2
`masking_prob`:	0.5
`resized_image_size`:	150x150

In the following the architecture and the model training structure are described concretely. The 'GetDataset' custom dataset class is created in PyTorch as a subclass of the 'Dataset' class. It is responsible for loading and preparing the training data. The class applies transformations to the photos, such as scaling and conversion to a tensor. Additionally, it requires the root directory of the training dataset to be specified. To create

an instance of the custom dataset, the specified root directory and transformations are used. The batch size, shuffle option, and number of workers are set when creating a data loader for the training data.

In the decoder network, which is the inverse of the encoder, hidden layers gradually increase the sampling of the latent vector until the original image size is restored. The decoder employs leaky ReLU activations and linear transformations. The last layer of the decoder uses a leaky ReLU activation followed by another linear transformation to match the input image dimension.

During training, the input data undergoes a masking procedure before being passed to the encoder. The masking probability, a hyperparameter, determines the proportion of features that are set to zero. The encoder and decoder then process the masked data to reconstruct the original image. The reconstruction loss is calculated using the mean squared error (MSE) loss function/L2 between the original image and the reconstructed image.

The MAE model is trained using the Adam optimizer with the specified learning rate. The training is conducted over a certain number of epochs. The training data is loaded using a DataLoader, which handles batching and parallel data loading based on the set batch size. If a GPU is available, it is utilized for training.

Iterating over the training data batches is part of the training loop. The input data is reshaped for each batch, and the masking process is applied. The loss is computed during the forward pass using the MSE loss function/L2 and the reconstructed images. The optimizer adjusts the model parameters based on the gradients obtained through backpropagation. The loss values are recorded for analysis.

To assess the training progress and effectiveness of the MAE model, the loss values are plotted throughout the epochs. Additionally, a sample of randomly selected original, masked, and reconstructed images is displayed. At the end of the training procedure, the trained MAE model is saved by saving its state dictionary to a `.pth` file.

3. Validation

To evaluate the effectiveness and dependability of the MAE-SSL technique across three datasets, the evaluation of loss functions and the image congruence are crucial.

Pump Impeller

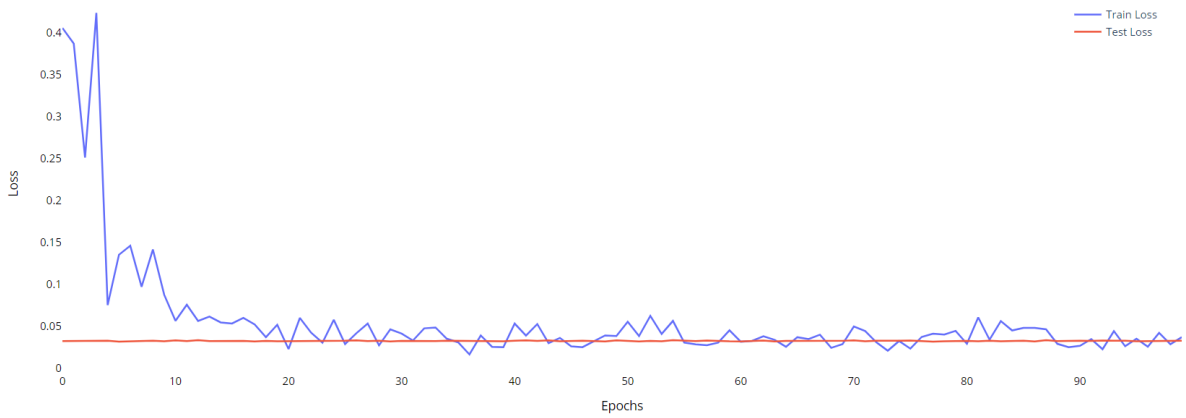


Fig. 3.6: Pump Impeller Dataset - Train and Test Loss Function

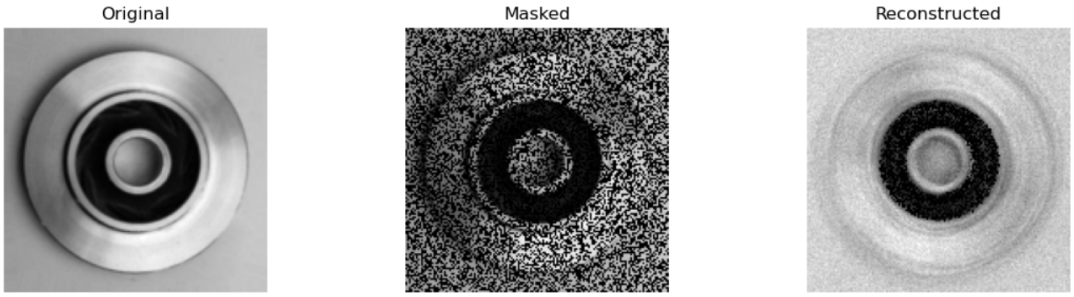


Fig. 3.7: Original, Masked and Reconstructed Pump Impeller Images

In Fig. *3.6* the loss function of the training and the test from the pump impeller data can be seen. The training loss has a decreasing course and settles between 0.050 and 0.020 after about the 20th epoch. In this case, the training loss values range from 0.020 to 0.361. The lowest loss values indicate that the model fits the training data very well and makes accurate predictions. The Standard Deviation of the loss is 0.0684. The test loss function has no decresing course. It stays between 0.032 and 0.034. It shows behavior of overfitting. Fig. *3.7* illustrates the MAE process. For this dataset, the reconstruction of the original image was fine. Thus, there are parallels between the loss function and the image congruence.

Sealing Boot Lip

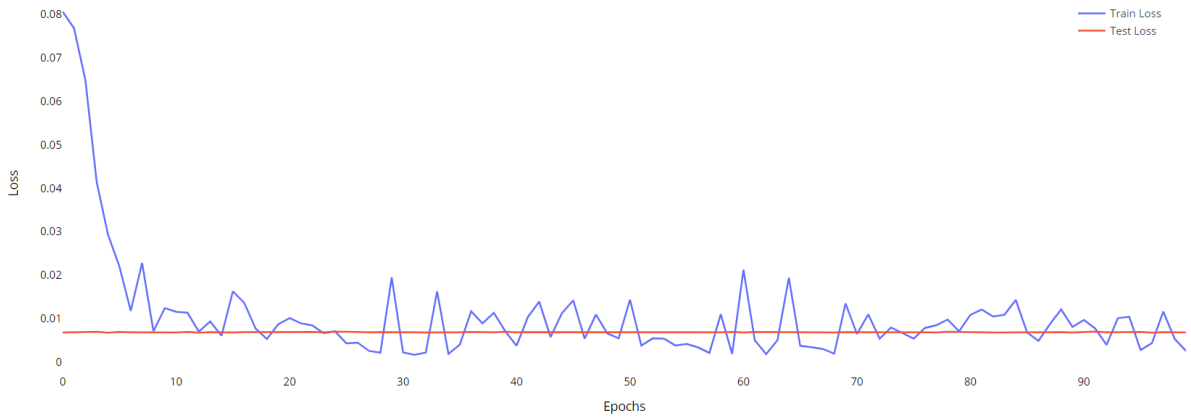


Fig. 3.8: Sealing Boot Lip Dataset - Train and Test Loss Function

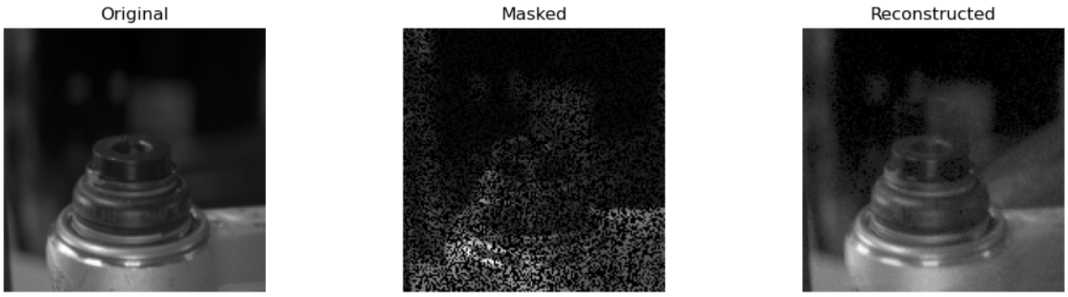


Fig. 3.9: Original, Masked and Reconstructed Sealing Boot Lip Images

In Fig. *3.8* the loss function of the training and the test from the sealing boot lip data can be seen. The training loss has a decreasing course and settles between 0.020 and 0.003 after about the 6th epoch. The training loss values in this example range from 0.002 to 0.077. These values provide information about the discrepancy between the predictions of the model and the actual values of the training data. The Standard Deviation of loss is 0.0128. These values are much lower than those of the Pump Impeller dataset. There is a similarity, however, in the course of the test loss. Here, too, there is a fluctuating course, but between lower values of 0.0065 and 0.009. Signs of overfitting are also noticeable here. Fig. *3.9* illustrates the MAE process. For this data set, the reconstruction of the original image was good, as can also be seen from the shape of the loss function.

Brake Caliper

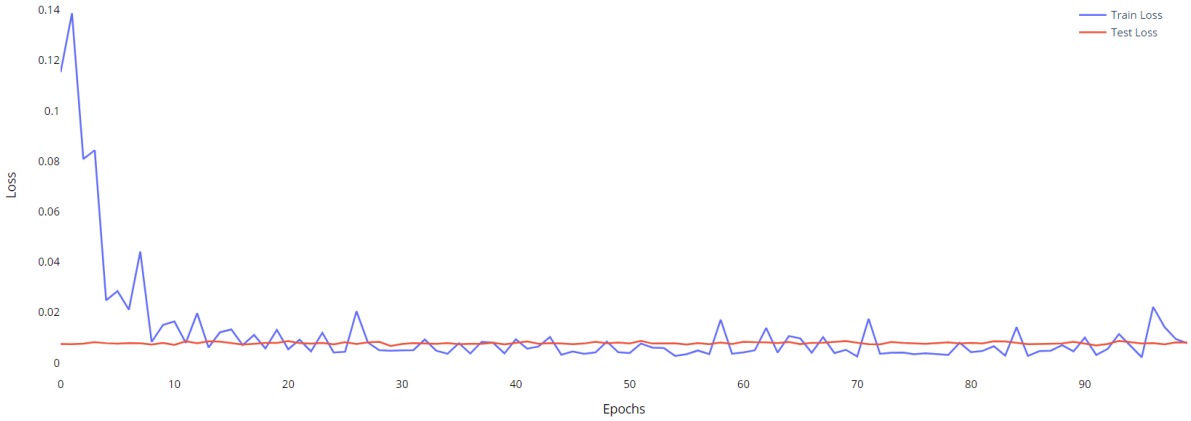


Fig. 3.10: Brake Caliper Dataset - Train and Test Loss Function

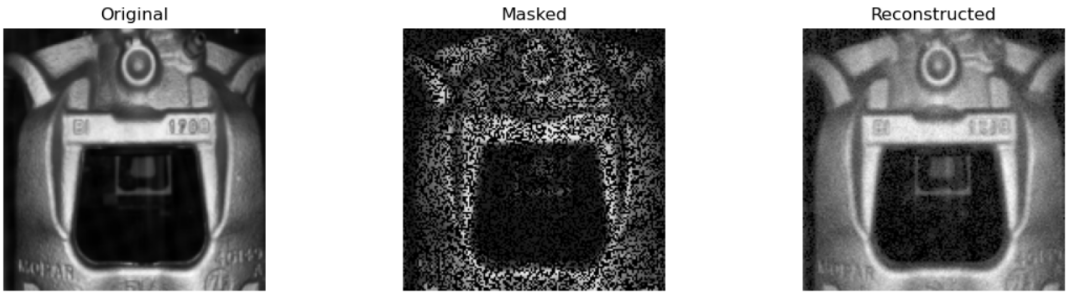


Fig. 3.11: Original, Masked and Reconstructed Brake Caliper Images

In Fig. *3.10* the loss function of the training and the test from the brake caliper data can be seen. The training loss has a decreasing course and settles between 0.014 and 0.004 after about the 10th epoch. The training loss values in this data set range from 0.003 to 0.145, and serve as a measure of the difference between the predictions of the model and the actual values of the training data. The Standard Deviation of loss is 0.0205. This dataset shows a similar result to the Sealing Boot Lip dataset. The loss values are in a similar range. The course of the test loss is almost identical. The values vary between 0.0067 and 0.007. Fig. *3.11* illustrates the MAE process. For this data set, the reconstruction of the original image was very good.

In general, it can be said that the training results turned out well. The reconstructions deliver good results. This is evident not only from the clear loss function course, which does not get worse, but also from the image congruence. It also turns out that the data set from Kaggle (Pump Impeller) has performed the worst. Correlations with the

number of pixels were found. The number of pixels of the Kaggle dataset is at least five times higher than the other two. The Standard Deviation of 0.0684 is much worse than the other two data sets. Thus, a minimum pixel count of 1000x1000 should be given.
All three data sets show similar behavior with respect to training and testing. In places the tendency for *Overfitting* increases, as can be seen in the figures. Here, in Fig. *3.6* there are fewer signs of overfitting.
Also, noticeable is the influence of the amount of data. The Kaggle dataset is almost three times as large as the other two. With the amount of data comes the variety of images. The more similar the images are to each other, the more general the reconstruction images become, since there are fewer diverse features to learn.

Threshold Detection

To achieve an optimal result during the training, it is important to keep an eye on overfitting. Often it is useful to define stopping criteria to implement early stopping. In this thesis, the threshold value is determined for each data set to determine the stopping value. A quantile selection is used. The 0.75 quantile is applied in this instance. Thus, 75 % must lie below the value to be determined (Pen, 2014).

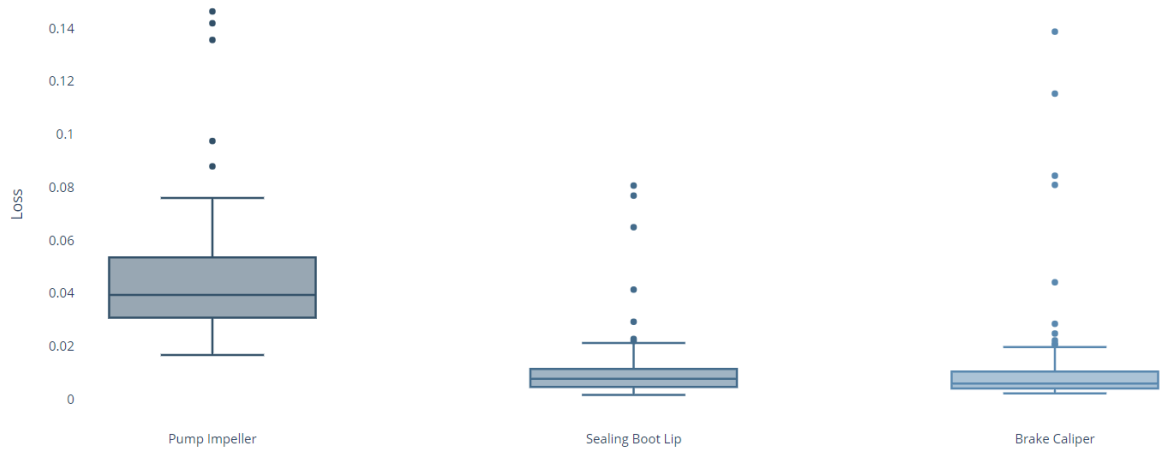


Fig. 3.12: Quantile Representation for Threshold Detection

As can be seen in Fig. *3.12*, the loss values of the three data sets were plotted in boxplot form to visually highlight the quantiles as well. Considering the individual loss values per data set, the following threshold values of the Pump Impeller: 0.054, the Sealing Boot Lip: 0.011, and the Brake Caliper: 0.010. are obtained, as shown in the graph.

These determined thresholds can now be used as a stopping criterion for further training with a larger amount of data or similar image data. Similar image data can, for example, include different vehicle brands for the brake caliper. Also, multiple camera perspectives can be added to provide more information.

3.3 Data Annotation Concept Implementation

As highlighted in the former chapters, the application of AI in manufacturing is becoming more present and essential in the industry. This is why companies in different sectors see the importance of goal setting and strategy development associated with Data, Analytics, and AI (Wennker, 2020).

When it comes to the implementation of the DAC, the strategic and technical implementation must be considered. Firstly, the strategic context is shown, and then the technical one. At the strategic level, the concept is embedded in the OKRs. On a technical level, the focus is on object-oriented programming and generalizability.

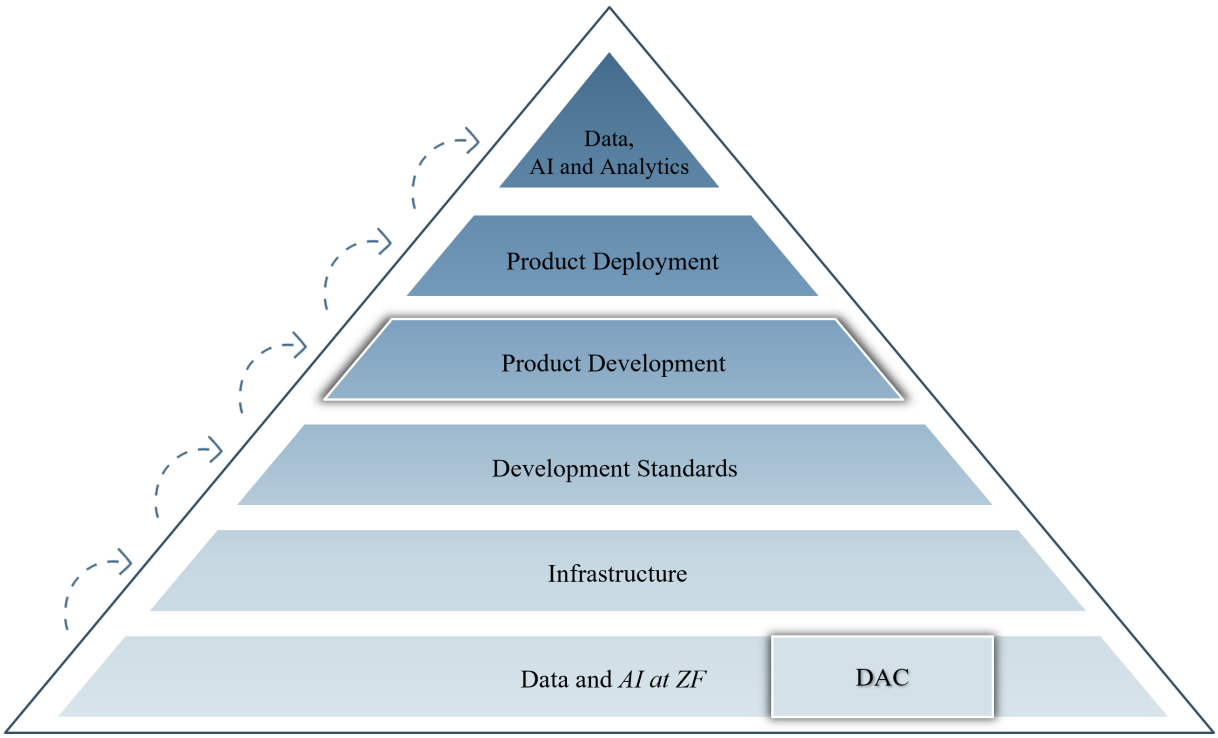


Fig. 3.13: Breakdown of the OKRs in the Data, Analytics, and AI Field of Industry

The ZF Group has defined the following top OKRs in the area of data, AI and analytics: *Scale and industrialize Data, Analytics, and AI solutions to accelerate digital intelligence*. Fig. *3.13* shows the holistic picture of the interaction of the individual OKRs and clarifies their significance through the levels that build on each other. From the above OKR, the following OKRs of the pyramid are derived.

AI Quality Inspection at scale is the focus to meet the top OKR. To fullfill this the AI product development of the CVT comes into play. Then the CVT is a scaleable is AI Visual Quality Inspection Software for manufacturing. The application refers to delivering a standardized, automated, and reusable approach.

The CVT gives all ZF factories the ability to integrate scalable CV technology into their manufacturing processes. This solution is based on ZF's strategic Digital Manufacturing Platform and Advanced Analytics Platform platforms, guaranteeing the best internal synergy and utilizing the current infrastructure of the product deployment. The ability to preserve, adapt, and reuse previously completed work, including a repository of pretrained AI models for quality inspection cases, is made possible by CVT by eliminating reliance on outside sources. This results in the subordinate OKR of the ZF Group: *CVT has seen 40 deployments in 2023 across ZF plants.*

Within the product development, the ecosystem of the CVT allows the availability of quick and reusable AI solutions for various Quality Inspection demands by utilizing and conserving gathered data. By enabling internal scalability and cost-effectiveness, this capacity enables enterprises to lessen their dependency on independent vendors and their standalone products.

Organizations may promote a self-sufficient and autonomous approach to quality inspection with the help of this solution. Businesses may better control their Quality Inspection requirements and create the conditions for long-term development and success by reducing their reliance on outside services. This leads to the following CVT OKR: *Guaranteed traceability, enable image archive and savings in indirect labor costs.*

At the development standards level, the use of the CV ecosystem enables faster detection of defects, an increase in overall line throughput and Overall Equipment Effectiveness values, and automatic adjustment of tolerance limits. This results in a reduction of pseudo defects that lead to unnecessary downtime. In addition, the system enables the potential detection of defective tools to optimize additional line stops. Compared to manual inspec-

tions, the defect detection rate increases significantly. This leads to the following OKR: *Performance Increase, reduce downtimes and 100% quality inspection.* In the area of infrastructure, the use of the CV ecosystem enables a reduction in manual efforts, resulting in lower costs for indirect labor and callbacks. This delivers the desired quality and avoids problems on the customer side. In addition, the risk of problems on the customer side or during assembly is reduced. The resulting OKR is the following: *Savings in time and cost, fulfill customer requirements, and reduce callback cost.*

The basis for these aspects is composed of the Data and *AI at ZF*. An important component of data quality is whether the data are appropriate for the model and fit the intended use. Data must be devoid of missing data, outliers, and disruptions in order to be useful for modeling. To guarantee a high-quality level, they must also comply with the specifications and be error-free. Data heterogeneity resulting from various formats, structures, and quality concerns can influence the quality, particularly if the underlying data model has a poor structure. Data quality is evaluated using metrics including consistency, consistency, validity, uniqueness, correctness, and completeness. ML places a high value on high data quality because the model's success directly depends on it. Only data of the highest caliber and the most pertinent types can yield accurate findings.

Within *AI at ZF* object-oriented programming provides a solid foundation for the development of ML models. Coding standards and the PyTorch framework can be used to create efficient and well-documented models. Data must be carefully prepared, and the selection of the right algorithm and tuning of hyperparameters is critical. Evaluation metrics, overfitting prevention, and model interpretation play an important role. In conclusion, model deployment in a productive environment is essential. This leads to the following OKR: *Object-oriented programming and PyTorch for efficient ML models and deployment.*

Looking at the technical side of the implementation, it can be seen that data and its preprocessing, form the foundation for any AI application. When embedded in the CVT context, the outcome of the DAC can be seen as a subordinate OKR of resources and capacities, and data and plays into the staging as follows.

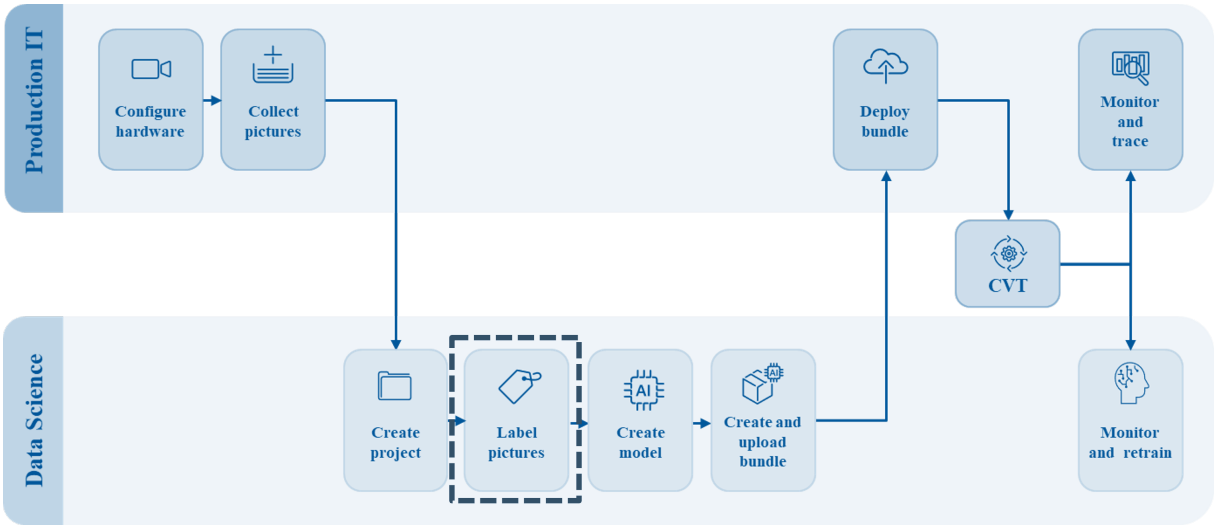


Fig. 3.14: Computer Vision Toolbox Application - User Journey

The DAC plays into the overall dynamics of the pyramid by achieving the outcome explained above, in the form of a DAC-OKR. A closer look at the CVT application lanes in Fig. 3.14 shows that the DAC is classified in the *Data Science* Lane under *Label pictures*.

In the area of Data Science, labeling of images and camera parameter consultations are performed. In addition, AI models are created, evaluated, and adapted to the specific requirements of the factory.

This enables the optimal use of AI models in manufacturing. Thus, the technical implementation of the concept fits into the given environment. Among other things, strong generalizability is enabled by the life of *AI at ZF*. The DAC could be included as a module in the existing libraries or used as an external notebook implementation.

Now the question arises of how the DAC affects the already described levels of OKRs. The DAC has a significant impact on OKRs. It saves resources through efficient DA and automation, streamlines the process, and automates repetitive tasks. This increases overall efficiency and reduces costs. Model quality improves through accurate annotation, resulting in more robust and higher-performing models. Traceability enables transparency and validation of results. The DAC offers companies the opportunity to gain experience in CV and develop functional models for different domains and locations. It also enables the scaling of AI, data, analytics, and model development in the industry. Overall, the company benefits from the positive impact of the DAC on various levels of OKRs.

4 Summary and Outlook

This chapter summarizes the previous chapters' findings and provides an outlook based on the chapters to follow. Limitations and critical perspectives are discussed to provide a comprehensive picture. Based on this, the research questions are answered and explained in more detail.

4.1 Conclusion

The use of AI in the manufacturing context is increasing. CV for predictive maintenance, quality inspection, among others. is becoming more and more present and offers a lot of potentials as discussed in chapter *2*. However, as for any AI application, data remains the foundation. The selection, provision, and processing of data plays an immense role and account for more than 80% of ML development. The DAJ is a part of this and therefore essential for model performance. SSL is a new approach in this area, but it is already very effective in improving the DAJ. MAEs, which are already researched, are a general method in this particular setting and can be applied to a variety of scenarios, unlike other SSL methods, which is why they were used for the investigation of the DAC in this thesis. The structure of the DAC is kept intuitive and can and should be adapted to the respective manufacturing context. With the phase-specific instructions, a kind of manual is available to enable the application of the DAC step by step.

In the course of the thesis, individual loss values per data set are considered, and thresholds for the Pump Impeller (0.054), Sealing Boot Lip (0.011), and Brake Caliper (0.010) are determined. The results of the Brake Caliper dataset perform best. Different factors like the size of the data set, number of pixels, among other aspects. affect the result. The DAC's testing can be evaluated positively in the approach and the results.

For the DAC implementation it is important to look at the classification and implementation of the DAC from strategic and technical perspective. From the strategic perspective, defining OKRs supports the scaling of AI, data, analytics, and model development in the industry. With having the different aspects of the DAC considered in the various levels of OKRs, companies have a clear direction to go.

The added value of the thesis is a reflection of its innovativeness. So far SSL has only been used in the field of NLP. There is a lot of potential for research when using the new approach SSL in the field of CV and specifically referring to the DAJ. It is also important to better involve various areas in the industry and to connect the levels of production, strategy and research.

This now leads to answering the research questions posed, which are processed based on the results:

1. Can the performance of CV in the visual quality inspection context in manufacturing be improved through the implementation of a DAC based on the SSL method?

To answer this question, the performance-reducing factors from section *1.2* can be used: time-consuming, labor-intensity, cost increasing.
With regards to the time-consuming, it could be shown that SSL has a significant positive impact through speeding up the training process, improving data understanding, accelerating model development, and facilitating model scaling (Silva, 2020; Dilmegani, 2023; Benčević et al., 2022). Furthermore, it was also confirmed that it offers a promising method to improve models by providing effective representations for downstream tasks without labelling, reducing the labelling effort, improving robustness (Newton, 2023; Dilmegani, 2023; Vijayrania, 2023).
In terms of the cost factor affecting performance, SSL helps to reduce the cost of training and deploying models, reduce the effort required for labelling, detect errors and improve model performance (Hendrycks et al., 2019; Bengar et al., 2021; Rani et al., 2023).

The labor-intensity has been limited to data preprocessing and model development. Likewise, no human bias could be incorporated into the DAJ. Due to the universal application of SSL to any application the performance of various CV solutions in the visual quality inspection context in manufacturing can be improved through the implementation of a DAC based on the SSL method. Therefore no specific DA methods are required for tracking CV applications. Whereas especially, MAEs have generalizable properties, as shown in section *2.3*.

In the present use case within the ZF Group, up to 80 % of the labeling costs to be incurred could be saved by using SSL. These would need to be weighed against the computational costs incurred. Looking ahead, costs are also saved with regard to possible errors. SSL eliminates manual, possibly negligent annotation errors and the model automatically determines meaningful features.
The results above summarized in Tab. *A.1* show that the first research question can be answered positively. The performance of CV in manufacturing can be improved through the implementation of a DAC based on the MAE - SSL method. Nevertheless, further investigation is needed to produce the existing potential for improvement.

The second research question deals with the question of how an improvement of the quality inspection is possible:

2. How can CV in the visual quality inspection context in manufacturing be improved through the implementation of a DAC based on the SSL method?

Defining OKRs and identifying levels of goal setting is critical to effective goal achievement. As shown in Fig. *3.13*, the lowest level forms the foundation on which the upper levels are built. This is where data and the DAC play a significant role.
Data takes up a large part of this foundation. By optimizing the factors of time, cost, and labor intensity, a solid foundation is created that enables OKRs to be achieved. This is done by minimizing time, reducing costs, and reducing labor intensity. This enables more efficient and effective goal achievement.
This optimization has a positive impact on the achievement of the higher-level OKRs. Especially in manufacturing quality control, developing better models on a sound basis leads to improvement. Even the smallest optimizations have a big impact on the whole picture. A solid data strategy and a well-thought-out DAC play a crucial role in supporting the goals at all levels of the organization.

4.2 Critical Discussion

In performing this thesis, some limitations were identified that may affect the results and conclusions. These are critically examined in the following.
The use of color coding in DA can appear subjective and lead to differences in perception. This can lead to potential challenges in the consistency of the annotations. Color coding can be interpreted differently from person to person. What appears to be a positive meaning to one person may have a negative connotation to another. This can lead to misunderstandings and affect the consistency of DA. It is important to clearly communicate context and ensure that all stakeholders have a common understanding of color coding. Instead of exclusive color coding, other visual elements such as symbols, shapes, or text can be used to indicate comments. This can reduce perceptual differences and allow for more consistent interpretation. This leads to the following questions: How can it be ensured that color coding is interpreted consistently? And what alternative visual elements could be used to indicate comments?

Due to limited computational resources, processing a large number of pixels can lead to performance issues. This can affect the speed of DA and model training. This may affect the scalability and efficiency of the overall approach.
To minimize the performance problems associated with processing large numbers of pixels, efficient use of computing resources is required. This can be achieved by using dimensionality reduction algorithms or prioritizing relevant pixels. Instead of processing each pixel separately, data aggregation methods can be used to group multiple pixels and treat them as unified entities. This can help reduce the computational load and improve the speed of data annotation without losing essential information. By using specialized hardware, complex pixel computations can be parallelized and accelerated, leading to improved scalability and efficiency.
Processing large volumes of pixels can present both technical and conceptual challenges. Efficient use of resources, development of scalable solutions, and careful selection of relevant information are critical to improve performance and ensure quality results. The following questions arise:What strategies can be used to make more efficient use of computing resources? And how can algorithms be optimized to improve the scalability of the overall approach?

The selection of image data can be subjective and depend on the individual preferences of the annotator. This can lead to some variation and fuzziness in the annotation labels. The same applies to the reconstruction comparison.
Blurring can occur in DA of image data because different annotators may have different interpretations and perspectives. What seems obvious or important to one person may be less significant to another. Image data selection and DA are highly dependent on context. The intended purpose of the annotation, the desired analysis outcome, or the specific application may influence the selection of image data, leading to variations. There may be a lack of standardized criteria or guidelines for image data selection. Developing and implementing standardized criteria for image data selection can help ensure consistency and comparability in annotations. In the context of the DAJ following question arises: How can standardized criteria or guidelines for selecting and annotating image data be developed? And how can domain experts be involved in the process without compromising DA consistency?

The availability and selection of appropriate image files can be a limiting factor. It is important to ensure that the data are representative and diverse enough to ensure adequate training of the model.
It is important to ensure that image files are of high quality to enable accurate and meaningful DA. Poor image quality, distortion, or blurring can affect the accuracy of the DA and cause distortion. Selection of appropriate image files requires that the data be representative of the application scenario or problem. The data should cover different aspects, features, or categories to ensure adequate training of the model. Limited diversity can lead to limitations in the model's ability to fit and potentially miss certain data patterns or features. Consideration of different variations and variances in image files is therefore critical. Privacy and legal considerations must also be taken into account when selecting image files. It is important to ensure that the image files selected comply with applicable privacy regulations and do not contain confidential or sensitive information. The following questions are at the forefront in relation to this topic: How can it be ensured that the selected image files are of high quality? And how can data protection and legal aspects be taken into account?

Although the loss function is used as an evaluation metric, it may not always be 100% informative. There is a possibility of overfitting as shown in the loss functions in chapter *3*, where the model overlearns the training data, and performance on new, unknown data may be degraded. The decision to use the MAE method is based on its general applicability to most use cases. However, there may be scenarios where other SSL methods may be more appropriate. Selecting the right method requires a comprehensive evaluation and consideration of the project's specific requirements.
The definition of OKRs may not be sufficiently well-defined. This can lead to challenges in evaluating the success of the DAC and affect the interpretation of the results. Careful selection of the loss function, avoidance of overfitting, clear definition of OKRs, and appropriate evaluation of results are critical to assessing and improving the success of the DAC.

There are several uses for CV, as explored in chapter *2*. This discipline is fast-growing. CV systems can be expensive to implement, which can be a limitation for small and medium-sized businesses. Effective implementation of algorithms is one of the challenges of using CV in manufacturing. There is a need for benchmarks to evaluate the performance of CV algorithms in manufacturing. Despite these challenges, CV is playing an important role in advancing the informatization, digitization, and intelligence of industrial manufacturing systems (Zhou et al., 2021).
The DAJ is one of the main difficulties within CV. It can take a lot of time, effort, and money to manually annotate data with the help of human specialists who provide the data with knowledge that is grounded in reality.

Additionally, the procedure is somewhat subjective because several annotators may have different interpretations or prejudices. This subjectivity may add mistakes and inconsistencies to the labeled data, which will have an impact on how well the trained models function. Additionally, the manual annotation has a limited capacity to scale, especially when working with huge datasets or frequently updated data (Javaid, 2023).
Some of the images that are inevitably noisy, or improperly tagged, are those that are retrieved while creating a dataset using a search engine. An important research direction that would make the process of creating datasets easier is the design of robust algorithms for training on noisy data collected from the internet (Albert et al., 2021).

All in all, the question arose why the SSL approach for the DAJ for CV was not tackled earlier. SSL has been around for years in the NLP area, but it only gained popularity in use for DAJ for CV from the early 2020s. This led to the problem of the distribution of time and capacity for industry research. This will continue to slow down the progress of innovation in the industrial context.

The accelerated innovation race is necessary to remain competitive in a rapidly changing world. It is important to maintain public trust in the industry through innovations that generate societal value as well as economic returns. Responsible Research and Innovation (RRI) has recently emerged as a new concept that can drive an understanding of the industry's responsibility to society and the environment. The motivation of companies to adopt RRI, the status of implementation of concrete RRI practices, the role of stakeholders in responsible innovation processes, and drivers and barriers to the further diffusion of RRI in the industry are relevant to companies of different sizes and sectors. Collaboration between industry and academia in RRI requires a clear definition of success from both sides of the collaboration (Martinuzzi et al., 2018; Amini et al., 2020).

Therefore, it would be helpful to set OKRs in the area of Research and Development, which serve as KPIs to be able to monitor the process and drive progress in different research areas of the industry in a targeted manner.

SSL for CV is still in the beginning phase, so there are not yet a general number of papers available that cover the use of SSL. To make the DAJ effective, SSL needs a lot of unlabeled data. For some activities, SSL is not appropriate (Zhang et al., 2021b; Tsutsui et al., 2021).

Although SSL might lessen the labeling load, oversight is still necessary. In some circumstances, it might not perform as well as SL. It is a tried-and-true technique with excellent detection accuracy, but it costs a lot to integrate and needs a lot of labeled training data. When labeled data is sparse or unavailable, SSL can be used instead, although its performance might not be as excellent as SL (Yang et al., 2022).

SSL can be used to pre-train a model on a large quantity of unlabeled data when there is only a small amount of labeled data available. The model may then be adjusted to perform better using the scant-labeled data. To acquire meaningful representations of unstructured data, such as text, images, and audio, SSL can be utilized. To help SL models perform better, SSL can produce augmented data.

SSL requires massive computing power to train models on large data sets. Computational power is typically required on a similar scale as for neural networks or small base models, since a model takes raw data and labels it without human input. The accuracy of SSL is inherently lower than SL or other approaches. Without human input in the form of labels, annotations, and training data with reliable results, the initial accuracy score is low (Hvilshøj, 2023).

In the context of the MAE-SSL approach, it might be difficult to choose suitable masks that maintain the key properties and encourage the model to acquire useful representations. Masking techniques that are inaccurate or inadequate might result in subpar performance and restrict the model's ability to convey the appropriate visual notions. Depending on the quantity of the dataset, the computational difficulty of training MAEs might drastically rise, rendering it unworkable for some applications. Finding the ideal configuration may need lengthy testing. MAEs' performance can be influenced by the network designs and hyperparameters used (Zhang and Shen, 2022; Ly et al., 2022).
Autoencoders are trained on a specific dataset, and their performance may not generalize well to new data that is different from the training set. Autoencoders may not be able to learn complex representations of the input data, especially if the data is highly nonlinear or has complex dependencies. Autoencoders may not perform well when there is missing or noisy data, as they try to reconstruct the input data exactly. The learned representations in autoencoders may be difficult to interpret, especially if they are high-dimensional (Wang et al., 2022b; Lu et al., 2022; Barwey et al., 2023).
For this reason, it must be critically considered to what extent the use in sensitive areas such as production is applicable concerning the reusability of the DAC.

4.3 Outlook

For the further implementation of SSL for the DAJ, it is important to define the quality deficits in customized solutions. Different areas should be taken into account. The logo, color scheme, quality standards, customer preferences, etc. It is important to define that the DAC is not to be seen as a set concept, but as a holistic fundament that can be extended and tailored. This means that metrics and the SSL methodology are interchangeable.

It might be interesting to investigate and compare other SSL methods to determine which method is best suited for specific use cases. This will allow for a more comprehensive evaluation and optimization of the DAC.
Expanding the data palette by using different image datasets can help improve the robustness and generality of the DAC. By exploring different image types, resolutions, and scenarios, the limitations of the DAC can be expanded, and its performance evaluated in different contexts. There the best fitting area of application can be extracted.
Integration of the DAC into various CV downstream tasks can be explored to evaluate the effectiveness of the concept in different application domains. By applying the DAC to specific tasks such as object detection, segmentation, or classification, its performance, and flexibility can be further investigated. A more detailed long-term study can be conducted to observe the evolution of the OKR in the context of the DAC. This will allow continuous evaluation and improvement of the DAC over a longer period of time.
CV can improve quality control, detect defects, increase inventory management efficiency, optimize the supply chain, and improve workplace safety. CV has the potential to improve efficiency, productivity, and accuracy, and reduce costs.
There are many uses for CV in the manufacturing industry, and it may revolutionize production and quality control to increase flexibility and efficiency (Clark, 2023).
Automation of quality inspection during production is one of CV's most crucial applications in manufacturing. In the industrial industry, upholding quality standards is crucial. The use of CV technology is significantly more successful than human observations in identifying changes in manufacturing equipment, even though one may accomplish this manually by enlisting the help of quality control professionals. Even in tiny machine parts, CV technologies have been utilized to detect flaws in real-time. This allows for the timely discovery and repair of the components (Boesch, 2023).
In order to monitor machinery and identify possible issues before they arise, CV can be employed. This may save maintenance expenses and downtime. CV may be used to keep an eye on the assembly line and spot any potential problems. This might increase effectiveness and cut waste. Inventory levels and the movement of commodities during the production process may be tracked using CV. This can facilitate waste reduction and enhance supply chain management (Ahramovich, 2023).

The market for CVs is anticipated to expand rapidly over the next few years as a result of factors like rising Internet of Things adoption, rising acceptance of automation in retail and industry, the emergence of autonomous cars, and the rising need for security and surveillance systems (Future Market Insights Global ; Ltd., 2023).

In addition to having great potential as independent learning processes, SSL methods may also be used as downstream tasks in subsequent ML processes. SSL techniques like contradictory learning and predictive coding may be used to extract useful representations from massive amounts of unlabeled data. These representations can then be used as a springboard for more specialized tasks, such as CV applications.
A general benefit of SSL as a downstream task is that it facilitates data processing by developing a fundamental understanding of the underlying structures and characteristics. Through the use of SSL, models may be trained using a wider variety of data, improving their robustness and generalization capabilities (Gavrilova and Markov, 2023; Gleave et al., 2023).
Future developments in CV applications and more exploration of their potential in the manufacturing sector are anticipated. In addition, it is anticipated that SSL will be integrated into the CV-learning process as a downstream task to enhance the performance of CV modules and enable the use of expensive but high-quality data (Hvilshøj, 2023).
SSL and USL are two topics that LeCun has been considering and discussing for years. Meta's Yann LeCun thinks that the future of AI will be heavily reliant on DL and artificial neural networks. He especially supports SSL, a kind of ML that requires less human input and direction during the training of neural networks. LeCun thinks that building the kind of reliable world models necessary for human-level AI will involve SSL using these kinds of high-level abstractions (Balestriero et al., 2023; Dickson, 2023).

LeCun points out the positive impact of SSL on human-level AI. From his point of view, it was *"the amplification of human intelligence, the fact that every human could do more stuff, be more productive, more creative, spend more time on fulfilling activities, which is the history of technological evolution."* (Dickson, 2022)

Glossary

Accuracy

Model accuracy in ML is the capacity of a ML model to produce accurate predictions or classifications on fresh, unexplored data. It is a statistic that counts how many cases out of all the instances were properly predicted or categorised. p. 1

Active Learning

In order to annotate or label samples from a big dataset, active learning is a ML approach that includes choosing the most instructive and pertinent examples (Tong and Koller, 2002). p. 15

Artificial Intelligence

The goal of AI is to build intelligent robots that can carry out activities like speech recognition, decision-making, and language translation that ordinarily need human intellect (Russell et al., 2010). p. 1

Audio Annotation

To make audio data comprehensible to machines, audio annotation involves adding metadata or labels. Identifying and naming certain sounds or elements in the audio, such as speech, music, or background noise, is known as audio annotation (Dutta and Zisserman, 2019). p. 17

Colorization

The technique of adding color to grayscale or black and white photographs is known as image colorization. The purpose of picture colorization is to turn a monochrome image into a realistic and aesthetically attractive color rendition (Vondrick et al., 2018a). p. 26

Computer Vision

CV is a subfield of AI and computer science that focuses on giving robots the ability to analyze and comprehend visual input from the outside world, such as images and videos (Kendall and Gal, 2017). p. 1

Contrastive Self-Supervised Learning

In Contrastive SSL, pairs of samples that are similar and dissimilar are contrasted to train the model to learn meaningful representations of data (Liu et al., 2021a). p. 23

Convolutional Neural Network

Convolutional neural networks are a subset of deep neural networks that are often employed in CV and image recognition tasks (Rahman et al., 2020). p. 11

Data Annotation

Data annotation is the process of labeling or tagging data with relevant information or metadata to make it more useful for ML algorithms. This involves adding labels or annotations to raw data, such as images, text, or audio, to provide context and meaning for the ML model to learn from (Rainer et al., 2022). p. 4

Data Annotation Job

Labeling or annotating data is referred to as a data annotation job, usually for ML and AI applications. To make raw data, such as photographs, movies, or text, comprehensible to computers, data annotation entails adding metadata or tags (Potter, 2023b). p. 4

Data Augmentation

A method used in data science and ML to expand the size and variety of a dataset by generating additional samples from the original set using operations like rotation, cropping, and noise addition(Fu et al., 2021). p. 26

Decoder

An artificial neural network's decoder is a component that reconstructs data from an encoded or compressed form. A decoder uses the encoded representation of input data as input in a neural network architecture and converts it into a higher-dimensional representation that is more similar to the original input data (Ye et al., 2023). p. 29

DevOps

DevOps is a set of practices and methodologies used to improve collaboration and communication between software development teams and IT operations teams (Ebert et al., 2016). p. 31

Encoder

A component of an artificial neural network called an encoder is utilized to extract features from input data. An encoder in a neural network design converts input data into a lower-dimensional representation that is more manageable and faster to

process (Ye et al., 2023). p. 13

Generalization

The capacity of a model or algorithm to perform effectively on fresh, untried data in addition to the data it was trained on is known as generalization (Zhang et al., 2021a). p. 2

Generative Adversarial Network

A certain kind of neural network design called GANs may produce new data that is comparable to a given dataset (Thada et al., 2023). p. 29

Generative Self-Supervised Learning

Generative SSL is a type of SSL where the model is trained to generate new data that is similar to the training data (Liu et al., 2021a). p. 23

Human-in-the-loop

The phrase human in the loop refers to incorporating human judgment and knowledge into the pipeline of data processing or decision-making in ML and AI (Potter, 2023a). p. 16

Image Annotation

To make images comprehensible to machines, image annotation involves adding information or labels to the images. Identifying and categorizing particular elements or objects in an image, such as objects, areas, or qualities, is known as image annotation (Pagare and Shinde, 2012). p. 18

Image Captioning

Creating a textual description of an image is the task of image captioning, a CV and natural language processing technology (Sharma et al., 2018). p. 14

Image Classification

In CV and ML, the process of classifying images into distinct groups or classes based on their attributes is known as image classification (Simonyan et al., 2013). p. 12

Image Rotation

The act of rotating a picture about its center point is known as rotation. In order to edit or change pictures, this procedure is frequently employed in CV and image processing applications (Atsuyuki et al., 2022). p. 26

Label

An output or target variable that a model is taught to predict is referred to as a label in ML and data science (Potter, 2023b). p. 2

Latent Features

Latent features are traits or aspects of a dataset that may be deduced from data patterns even when they are not readily visible. ML algorithms, which utilize mathematical methods to find underlying patterns in data and represent them in a lower-dimensional space, are often used to learn them (Luss et al., 2021). p. 29

Loss

In ML, loss is a metric that measures the difference between the predicted output of a ML model and the actual output. It represents the error or cost associated with the model's predictions, and is used to train the model by updating its parameters to minimize the loss (Wang et al., 2020). p. 6

Machine Learning

ML is a type of AI that enables machines to automatically learn and improve from experience, without being explicitly programmed. It involves developing algorithms and statistical models that can analyze and make predictions or decisions based on patterns and relationships found within data (Murphy, 2012). p. 3

Machine Learning Operations

MLOps is a set of best practices and techniques used to operationalize ML models and ensure their efficient deployment, scalability, and maintenance. MLOps is applying principles from DevOps, agile development, and data engineering to ML projects (Microsoft, 2023). p. 3

Masked Autoencoder

A Masked Autoencoder (MAE) is a particular kind of autoencoder neural network that is trained to reconstruct input data after using a masking function to selectively conceal certain of the input properties. A lower-dimensional representation of the input data is created after the masked input has been processed through an encoder network. The original input is then recreated using this lower-dimensional representation, while the masked features are recreated using the learnt model parameters (Zhang et al., 2022). p. 6

Object Detection

A CV approach called object detection includes finding and classifying items in an image or video by their location and identification (Selvaraju et al., 2019). p. 12

Overfitting

Overfitting is a common problem in ML where a model is trained too well on the training data, resulting in poor performance on new, unseen data (Huesmann et al., 2021). p. 47

Patch Positioning

Depending on their spatial connection with other image patches or areas, context placement, also known as patch positioning, is a technique used in CV and image processing to extract picture patches or regions of interest (Peng et al., 2020). p. 26

Scalability

Scalability refers to the ability of a system, process, or technology to handle increasing amounts of work, data, or traffic, without sacrificing performance, reliability, or maintainability (Sun et al., 2020). p. 5

Self-Supervised Learning

In the method of ML known as SSL, a model learns from the data on its own without the use of human labeling or supervision. SSL uses the data itself to produce labels or targets for the model to learn from, as opposed to SL, where the data is labeled with a specified output variable, and USL, where the data is not labeled (Zhou et al., 2022). p. 2

Semantic Segmentation

A CV approach called semantic segmentation includes giving a class label to each pixel in an image based on the semantic meaning of the item or region that pixel is supposed to represent (Shelhamer et al., 2016). p. 13

Semi-Supervised Learning

A ML approach called SMSL includes training a model using both labeled and unlabeled data. SMSL integrates both types of data, in contrast to SL, where the model is trained exclusively on labeled data, and USL, where the model is trained primarily on unlabeled data (Liu et al., 2022). p. 2

Supervised Learning

A model is trained on labeled data using SL, which aims to predict an output or target variable using brand-new, untainted data. The model learns to map the input characteristics and their associated output values together from the labeled data, which consists of both (Hardt et al., 2016). p. 12

Text Annotation

In order to make raw text data comprehensible to computers, text annotation is the act of adding information or labels to the text. Text annotation entails locating and labeling particular elements or characteristics of the text, such as named entities, grammatical constructions, sentiment, or topical groups (Stenetorp et al., 2012). p. 17

Transfer Learning

Transfer learning is a ML approach that includes using the information and understanding learned from one assignment to enhance the performance of a separate but connected activity (Fuzhen et al., 2019). p. 12

Unsupervised Learning

A model is trained on unlabeled data without a defined output or target variable using USL, a kind of ML. Without any prior understanding of what the data represents, USL aims to find patterns, structures, or correlations in the data (Hinton, 1999). p. 2

Video Annotation

The technique of adding information or labels to movies in order to make them comprehensible to computers is known as video annotation. Identifying and annotating particular areas, activities, or objects inside a video is known as video annotation (Pagare and Shinde, 2012). p. 19

Visual Question Answering

Answering questions about images is a part of the CV and natural language processing activity known as visual question answering (Anderson et al., 2017). p. 14

A Appendix

Tab. A.1: Effects of the DAC on the Performance of CV in Manufacturing

Performance-reducing Factor	**Effect of the DAC**	**Evaluation**
Time-consuming	- reduced time building a ML model - reduced data labeling time - increased scalability	SSL has a significant positive impact by speeding up the training process, improving data understanding, accelerating model development, and facilitating model scaling.
Labor-intensity	- increased effectivity - increase fairness of the model - increase model robustness	SSL offers a promising method for improving models by providing effective representations for downstream tasks without labels, reducing labeling overhead, improving robustness and uncertainty, and helping to promote model fairness.
Cost Increasing	- reduced data labeling costs - reduced training costs - reduced errors - increased data efficiency	SSL helps reduce the cost of training and deploying models, reduce labeling efforts, detect errors, and improve model performance.

Tab. A.2: Phases of the Data Annotation Concept (based on Kaiming et al., 2021; Baevski et al., 2022)

Concept Phase	**Phase Description**	**Relevant Factors**	**Parameter**	**Procedure**	**Attention Points**	**Phase Result**
1. Data Preparation	The labelless visual data that acts as the concept's fundamental building block is processed. It is prepared in the first stages during the data preparation phase. The type of data, from natural images to manufacturing images, and the amount of data might vary depending on the application.	- Format: Same data format, in which the data will be processed and stored - Data: The quantity and size of the data - Storage: The storage policies for the data - Instance: Individual instances or samples of the data are referred to as instances - Pixel: Individual elements or pixels that make up an image are referred to as pixels - Sector: Particular industries or areas that the data falls under	/	1. Establish what kind of image data the downstream task needs. -> It's vital to understand that there are no labels required at this stage. 2. Determine where the image data came from. Choose the websites, databases, or repositories from which the data will be gathered. -> Be aware of possible data storage guidelines. 3. Verify if there are any regulations to follow regarding data storage. 4. After finding the source of the image data, download it and store it in a suitable format (JPEG, PNG, etc.). 5. The image data must be processed. Various methods, such as scaling or normalizing, may be used for this. To increase the dataset size, in some circumstances, data augmentation techniques might be used.	- The objective variable is not mentioned, implying that this stage is concentrated on working with labelless visual data, maybe for SSL. - It is not indicated how much image data is needed for a SSL strategy to be successful. A thumb value for orientation could be a number of more than 1000 images. - The image data's format (JPEG, PNG, etc.), which must be determined, is not indicated. If any specific preprocessing methods should be used to improve the model's performance, it is unclear what they should be. - In batch processing, it is important to choose the appropriate batch size, shuffle data before training, ensure consistent data normalization, consider batch balancing for unevenly distributed data, and implement an appropriate system when batch processing is ongoing.	**Output**: Prepared and preprocessed compatible labelless visual data for further processing and analysis

Concept Phase	**Phase Description**	**Relevant Factors**	**Parameter**	**Procedure**	**Attention Points**	**Phase Result**
2. Self-Supervised Learning Approach	It entails utilizing a particular SSL technique to train a model to learn features on its own. The loss function from the model will be returned after using the SSL technique, showing how well the trained model performed.	- GPU/TPU: For effective computing during training, the GPU or TPU can be used	- `num_epochs`: The quantity of training iterations or epochs - `batch_size`: The quantity of samples processed in each training iteration - `learning_rate`: The speed at which a model's parameters are changed while being trained - `latent_dim`: The representation's latent space's dimensionality - `num_hidden_layers`: The SSL method's architecture's total number of hidden layers - `masking_prob`: The likelihood that input data for the SSL method will be masked or corrupted	1. Build the SSL method's architecture first: Design the network structure, layering, and connections for the specified SSL approach. 2. Decide on the batch size, latent dimensions, hidden layer count, learning rate, and masking probability. -> The SSL method's behavior and training procedure are controlled by these factors. 3. To train the model using the SSL technique, use the unlabeled image data. -> Without relying on labeled data, the model develops the ability to extract significant features. 4. Include all parameter settings and save the pre-trained model as a.pth file for later usage. To enable reuse and replication of the trained model's setup, save it in a .pth file format together with the parameters of the selected SSL method.	- To get good results, the SSL technique often needs a sizeable amount of unlabeled data for training. To enable the model to learn meaningful representations, there must be a minimum amount data. - The properties of the provided image data determine the ideal architecture for the SSL technique. Effective feature learning requires careful architecture selection. - The difficulty of the task and the quantity of accessible data can affect how many training epochs are required to reach the target level of performance. - Regularization methods, like dropout or weight decay, can be used to avoid overfitting and enhance the trained model's capacity to generalize.	**Output**: The .pth file containing the well-trained SSL method and all parameter settings is the phase's output. This file can be applied in the future or used as a foundation for additional fine-tuning or downstream operations of the CV task.

Concept Phase	Phase Description	Relevant Factors	Parameter	Procedure	Attention Points	Phase Result
3. Validation	The pre-trained model is exported as a `.pth` file for use in other applications during the validation step of the MAE-SSL approach. The link between the Loss function and the epochs is assessed in this stage.	- Loss: The outcome of the training and evaluation phases of the Loss function. - Epochs: The total number of training iterations or epochs. - Images: Original, Masked and Reconstructed Images.	/	1. Perform the Loss calculation and visualize the Loss function for each epoch. In order to see the trend, calculate the value of the Loss function at each training period and plot it over the epochs. 2. Secondly, explain the Loss function. Analyze the Loss function's actions during the training process. This analysis sheds light on the model's learning process and performance while also assisting in the understanding of how the Loss function evolves over time. 3. Retrain the model with an alternative Hyperparameter setting, if desired. The model can be retrained if necessary using various hyperparameter values. By examining various hyperparameter combinations, this stage enables experimentation and performance optimization of the model.	- The performance of the model might not fully represent all aspects of the work, and the ideal values for these measures might vary depending on the particular application. Additional analysis beyond the Loss function alone might be necessary. - The ideal threshold for the Loss function to consider the SSL technique effective is not defined, and finding it would call for further background information and domain-specific expertise. It has to be adapted to the specific use case to receive individual result.	**Output**: Evaluated Loss function and achieved the minimum maximum value of the loss

Concept Phase	Phase Description	Relevant Factors	Parameter	Procedure	Attention Points	Phase Result
A. Implementation	The `.pth` file with it's parameters is being implemented into a CV application. The `.pth` file, which contains the pre-trained model with its designated weights, is used in this stage. The model can be further adjusted for a downstream task or TL, such classification or segmentation, by adding extra layers and training on labeled data.	- `.pth` file - Accuracy: A performance metric used to assess the model's efficiency in handling the downstream task. - Epochs: The total number of training iterations or epochs.	/	1. The first step is to load the pre-trained model from the .pth file. The .pth file should contain the model settings and its assigned weights. 2. Increase the model's layer count to accommodate the subsequent task. Add new layers to the pre-trained model or change its architecture to suit the needs of the particular downstream task, such as a CV task like classification or segmentation. 3. Assess how well the model performed the downstream task. Use the modified model to do the intended CV task, then assess its effectiveness based on accuracy or other pertinent metrics. This evaluation sheds light on how successfully the pre-trained model, supplemented with extra layers, executes the particular downstream task. 4. Keep track of all the hyperparameters you used for training. As you go through the training and implementation phase, keep track of the hyperparameters you utilized. When applying the model to new data or doing additional fine-tuning, this documentation is critical for future reference and enables reviewing and maybe modifying the hyperparameters to obtain optimal performance.	- It's vital to remember that if the downstream task demands very specialized features that were not learnt during the MAE-SSL phase, the pre-trained model may not perform as well as a model trained on labeled data. - The pre-trained model's fine-tuning on labeled data might be a time- and resource-consuming operation. - Since the best architecture for the downstream task is not predetermined, it must take into account its unique requirements and characteristics.	**Outcome**: Implemented pre-trained model into a CV application, involves integrating the pre-trained model (stored in the .pth file) with specified weights into the CV application, to improve the model accuracy of the CV task

References

[Abagiu et al. 2021] Abagiu, M. M. ; Cojocaru, D. ; Manta, F. L. et al.: Detection of a Surface Defect on an Engine Block Using Computer Vision. In: *Proceedings of the 2021 22nd International Carpathian Control Conference (ICCC)*. Piscataway, NJ : IEEE, 2021, pp. 1–5. – ISBN 978-1-7281-8609-2

[Ahramovich 2023] Ahramovich, A.: Computer Vision in Manufacturing: 9 Use Cases, Solution Framework, Challenges & Best Practices. In: *itransition* (2023). – URL https://www.itransition.com/computer-vision/manufacturing. – Access Date: 01.06.2023

[Albert et al. 2021] Albert, P. ; Ortego, D. ; Arazo, E. et al.: Addressing out-of-distribution label noise in webly-labelled data. In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2021), pp. 2393–2402

[Amini et al. 2020] Amini, L. ; Chen, C. ; Cox, D. D. et al.: Experiences and Insights for Collaborative Industry-Academic Research in Artificial Intelligence. In: *AI Mag.* 41 (2020), pp. 70–81

[Anderson et al. 2017] Anderson, P. ; He, X. ; Buehler, C. et al.: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In: *arXiv preprint arXiv:1707.07998* (2017), pp. 1–15

[Anilkumar and Venugopal 2021] Anilkumar, P. ; Venugopal, P.: A Survey on Semantic Segmentation of Aerial Images using Deep Learning Techniques. In: *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE, 2021, pp. 1–7. – ISBN 978-1-6654-2691-6

[Atsuyuki et al. 2022] Atsuyuki, M. ; Qing, Y. ; Daiki, I. et al.: Rethinking Rotation in Self-Supervised Contrastive Learning: Adaptive Positive or Negative Data Augmentation. In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022), pp. 2808–2817

[Ayman 2023] Ayman, M. W.: Recent computer vision applications for pavement distress and condition assessment. In: *Automation in Construction* 146 (2023), pp. 104664. – ISSN 0926-5805

[Bachman et al. 2019] Bachman, P. ; Hjelm, R. D. ; Buchwalter, W.: Learning Representations by Maximizing Mutual Information across Views. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA : Curran Associates Inc., 2019

[Baevski et al. 2022] Baevski, A. ; Hsu, W. ; Xu, Q et al.: data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In: *arXiv preprint arXiv:2202.03555* (2022), pp. 1–15

[Balestriero et al. 2023] Balestriero, R. ; Ibrahim, M. ; Sobal, V. et al.: A Cookbook of Self-Supervised Learning. In: *arXiv preprint arXiv:2304.11210* (2023), pp. 3; 44

[Barwey et al. 2023] Barwey, S. ; Shankar, V. ; Viswanathan, V. et al.: Multiscale Graph Neural Network Autoencoders for Interpretable Scientific Machine Learning. In: *arXiv preprint arXiv:2302.06186* (2023), pp. 1–30

[Beck et al. 2001] Beck, K. ; Beedle, M. ; Bennekum, A. van et al.: *Manifesto for Agile Software Development*. 2001

[Bengar et al. 2021] Bengar, J. Z. ; Weijer, J. van de ; Twardowski, B. et al.: Reducing Label Effort: Self-Supervised meets Active Learning. In: *arXiv preprint arXiv:2108.11458* (2021), pp. 1–9

[Benčević et al. 2022] Benčević, M. ; Habijan, M. ; Galić, I. et al.: Self-Supervised Learning as a Means To Reduce the Need for Labeled Data in Medical Image Analysis. In: *arXiv preprint arXiv:220600344* (2022), pp. 1–5

[Blandfort et al. 2016] Blandfort, P. ; Karayil, T. ; Borth, D. et al.: Introducing Concept And Syntax Transition Networks for Image Captioning. In: Kender, John R. (Ed.) ; Smith, John R. (Ed.) ; Luo, Jiebo (Ed.) ; Boll, Susanne (Ed.) ; Hsu, Winston (Ed.): *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. New York, NY, USA : ACM, 2016, pp. 385–388. – ISBN 9781450343596

[Boesch 2023] Boesch, G.: Computer Vision In Manufacturing – The Most Popular Applications in 2023. In: *viso.ai* (2023). – URL https://viso.ai/applications/computer-vision-in-manufacturing/. – Access Date: 01.06.2023

[Bratulescu et al. 2022] Bratulescu, R. ; Vatasoiu, R. ; Sucic, G. et al.: Object Detection in Autonomous Vehicles. In: *2022 25th International Symposium on Wireless Personal Multimedia Communications (WPMC)*. IEEE, 2022, pp. 375–380. – ISBN 978-1-6654-7318-7

[Bulten et al. 2020] Bulten, W. ; Pinckaers, H. ; Boven, H. van et al.: Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. In: *The Lancet. Oncology* 21 (2020), Nr. 2, pp. 233–241. – ISSN 1470-2045

[Chen et al. 2016] Chen, L. ; Papandreou, G. ; Kokkinos, I. et al.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. In: *arXiv preprint arXiv:1606.00915* (2016)

[Clark 2023] Clark, B.: Computer Vision Transforms Production and Quality Control. In: *Machine Design* (2023). – URL https://www.machinedesign.com/automation-iiot/blog/21261964/bright-machines-computer-vision-transforms-production-and-quality-control. – Access Date: 01.06.2023

[Crespi et al. 2022] Crespi, L. ; Loiacono, D. ; Sartori, P.: Are 3D better than 2D Convolutional Neural Networks for Medical Imaging Semantic Segmentation? In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8. – ISBN 978-1-7281-8671-9

[Czimmermann et al. 2020] Czimmermann, T. ; Ciuti, G. ; Milazzo, M. et al.: Visual-Based Defect Detection and Classification Approaches for Industrial Applications — A Survey. In: *Sensors* 20 (2020), Nr. 5, pp. 1–25. – ISSN 1424-8220

[Dabhi 2023] Dabhi, R.: *Casting product image data for quality inspection.* 2023. – URL https://www.kaggle.com/datasets/ravirajsinh45/real-life-industrial-dataset-of-casting-product?resource=download&select=casting_data. – Access Date: 24.05.2023

[Dai et al. 2016] Dai, J. ; Li, Y. ; He, K. et al.: R-FCN: Object Detection via Region-based Fully Convolutional Networks. In: *arXiv preprint arXiv:1605.06409* (2016)

[Davies 2012] Davies, E.R.: *Computer and Machine Vision.* Waltham, MA : Elsevier, 2012. – ISBN 9780123869081

[Dickson 2022] Dickson, B.: *Meta's Yann LeCun on his vision for human-level AI.* 2022. – URL https://bdtechtalks.com/2022/03/07/yann-lecun-ai-self-supervised-learning/. – Access Date: 14.06.2023

[Dickson 2023] Dickson, B.: Meta's Yann LeCun is betting on self-supervised learning to unlock human-compatible AI. In: *The Next Web* (2023). – URL https://thenextweb.com/news/metas-yann-lecun-is-betting-on-self-supervised-learning-to-unlock-human-compatible-ai. – Access Date: 01.06.2023

[Dilmegani 2023] Dilmegani, C.: Self-Supervised Learning: Benefits & Uses in 2023. In: *AIMultiple* (2023). – URL https://research.aimultiple.com/self-supervised-learning/. – Access Date: 02.06.2023

[Doersch and Zisserman 2017] Doersch, C. ; Zisserman, A.: Multi-task Self-Supervised Visual Learning. In: *arXiv preprint arXiv:1708.07860* (2017)

[Dutta and Zisserman 2019] Dutta, A. ; Zisserman, A.: The VIA Annotation Software for Images, Audio and Video. In: Amsaleg, Laurent (Ed.): *Proceedings of the 27th ACM International Conference on Multimedia*. New York,NY,United States : Association for Computing Machinery, 2019 (ACM Digital Library), pp. 2276–2279. – ISBN 9781450368896

[Ebert et al. 2016] Ebert, C. ; Gallardo, G. ; Hernantes, J. et al.: DevOps. In: *IEEE Software* 33 (2016), Nr. 3, pp. 94–100. – ISSN 0740-7459

[Emam et al. 2021] Emam, Z. ; Kondrich, A. ; Harrison, S. et al.: On The State of Data In Computer Vision: Human Annotations Remain Indispensable for Developing Deep Learning Models. In: *arXiv preprint arXiv:2108.00114* (2021), pp. 1–12

[Feng et al. 2023] Feng, T. ; Dong, A. ; Yeh, C et al.: Superb SLT 2022: Challenge on Generalization and Efficiency of Self-Supervised Speech Representation Learning. In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. 2023, pp. 1096–1103

[Fu et al. 2021] Fu, K. ; Lin, J. ; Ke, D. et al.: A Full Text-Dependent End to End Mispronunciation Detection and Diagnosis with Easy Data Augmentation Techniques. In: *arXiv preprint arXiv:2104.08428* (2021), pp. 1–5

[Fuadi et al. 2023] Fuadi, E. H. ; Ruslim, A. R. ; Wardhana, P. W. K. et al.: Gated Self-supervised Learning For Improving Supervised Learning. In: *arXiv preprint arXiv:2301.05865* (2023)

[Fukui et al. 2016] Fukui, A. ; Park, D. H. ; Yang, D. et al.: Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In: *arXiv preprint arXiv:1606.01847* (2016)

[Future Market Insights Global ; Ltd. 2023] Future Market Insights Global ; Ltd., Consulting P.: Global Computer Vision Market is likely to reach a worth of US$ 26.11 Billion, at a CAGR of 7.3% by the forecast period ending 2033 | Future Market Insights, Inc. In: *Future Market Insights Global and Consulting Pvt. Ltd.* (2023). – URL https://www.futuremarketinsights.com/reports/computer-vision-market. – Access Date: 01.06.2023

[Fuzhen et al. 2019] Fuzhen, Z. ; Zhiyuan, Q. ; Keyu, D. et al.: A Comprehensive Survey on Transfer Learning. In: *Proceedings of the IEEE* 109 (2019), pp. 43–76

[Gao et al. 2022] Gao, W. ; Wu, M. ; Lam, S. et al.: Decoupled self-supervised label augmentation for fully-supervised image classification. In: *Knowledge-Based Systems* 235 (2022), pp. 107605. – ISSN 0950-7051

[Gavrilova and Markov 2023] Gavrilova, Y. ; Markov, I.: Is Self-Supervised Learning the Future of AI? In: *Serokell* (2023). – URL https://serokell.io/blog/is-self-supervised-learning-future-of-ai. – Access Date: 01.06.2023

[Girshick et al. 2013] Girshick, R. ; Donahue, J. ; Darrell, T. et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *arXiv preprint arXiv:1311.2524* (2013)

[Gleave et al. 2023] Gleave, W. ; Downes, J. ; Nakada, D.: *Launch self-supervised training jobs in the cloud with AWS Parallel Cluster*. 2023. – URL https://aws.amazon.com/de/blogs/hpc/launch-self-supervised-training-jobs-in-the-cloud-with-aws-parallelcluster/. – Access Date: 01.06.2023

[Gray 2019] Gray, M. L.: *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. 1st ed. Sydney : HarperCollins Publishers, 2019. – ISBN 9781328566287

[Han et al. 2021] Han, J. ; Ding, J. ; Xue, N. et al.: ReDet: A Rotation-equivariant Detector for Aerial Object Detection. In: *arXiv preprint arXiv:2103.07733* (2021)

[Hardt et al. 2016] Hardt, M. ; Price, E. ; Srebro, N: Equality of Opportunity in Supervised Learning. In: *arXiv preprint arXiv:1610.02413* (2016), pp. 1–22

[Hendrycks et al. 2019] Hendrycks, D. ; Mazeika, M. ; Kadavath, D.: Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In: *arXiv preprint arXiv:1906.12340* (2019), pp. 1–13

[Hessel et al. 2021] Hessel, J. ; Holtzman, A. ; Forbes, M. et al.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: *Conference on Empirical Methods in Natural Language Processing*, 2021

[Hinton 1999] Hinton, G.: *Unsupervised learning: Foundations of neural computation*. Cambridge, Mass. : MIT Press, 1999 (A Bradford book). – ISBN 026258168X

[Huesmann et al. 2021] Huesmann, K. ; Rodriguez, L. G. ; Linsen, L. et al.: The Impact of Activation Sparsity on Overfitting in Convolutional Neural Networks. In: *arXiv preprint arXiv:2104.06153* (2021)

[Hvilshøj 2023] Hvilshøj, F.: *Self-supervised Learning Explained*. 2023. – URL https://encord.com/blog/self-supervised-learning/. – Access Date: 31.05.2023

[Hwang et al. 2019] Hwang, E. J. ; Park, S. ; Jin, K. et al.: Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. In: *JAMA network open* 2 (2019), Nr. 3, pp. e191095

[Isola et al. 2017] Isola, P. ; Zhu, J. ; Zhou, T. ; Efros, A. A.: Image-to-Image Translation with Conditional Adversarial Networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5967–5976

[Javaid 2023] Javaid, S.: Top 4 Computer Vision Challenges & Solutions in 2023. In: *AIMultiple* (2023). – URL https://research.aimultiple.com/computer-vision-challenges/. – Access Date: 25.05.2023

[Jazmia 2023] Jazmia, H.: MLOps: A Primer for Policymakers on a New Frontier in Machine Learning. In: *arXiv preprint arXiv:2301.05775* (2023), pp. 1–20

[Jing and Tian 2019] Jing, L. ; Tian, Y.: Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2019), pp. 4037–4058

[Jung et al. 2022] Jung, H. ; Choi, H. ; Kang, M.: Boundary Enhancement Semantic Segmentation for Building Extraction From Remote Sensed Image. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–12. – ISSN 0196-2892

[Kahansky 2023] Kahansky, N.: What are OKRs? (and why you need to know about them). In: *Hypercontext* (2023). – URL https://hypercontext.com/blog/work-goals/what-are-okrs. – Access Date: 15.06.2023

[Kaiming et al. 2021] Kaiming, H. ; Xinlei, C. ; Saining, X. et al.: Masked Autoencoders Are Scalable Vision Learners. In: *arXiv preprint arXiv:2111.06377* (2021), pp. 4

[Kappel 2023] Kappel, N.: Warum MLOps? Die 5 Herausforderungen der ML-Produktentwicklung. In: *[at] data.ai blog* (2023). – URL https://www.alexanderthamm.com/en/blog/mlops-5-challenges-of-ml-product-development/. – Access Date: 16.06.2023

[Karatas 2023] Karatas, G.: Data Annotation in 2023: Why it matters & Top 8 Best Practices. In: *AIMultiple* (2023). – URL https://research.aimultiple.com/data-annotation/. – Access Date: 18.04.2023

[Karavellas et al. 2019] Karavellas, T. ; Prameswari, A. ; Inel, O. et al.: Local Crowdsourcing for Annotating Audio: the Elevator Annotator platform. In: *Human Computation* 6 (2019), pp. 1–11

[Karimi et al. 2020] Karimi, A. M. ; Fada, J. S. ; Parrilla, N. A. et al.: Generalized and Mechanistic PV Module Performance Prediction From Computer Vision and Machine Learning on Electroluminescence Images. In: *IEEE Journal of Photovoltaics* 10 (2020), Nr. 3, pp. 878–887

[Kendall and Gal 2017] Kendall, A. ; Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: *arXiv preprint arXiv:1703.04977* (2017), pp. 1–11

[Khurana and Chandak 2013] Khurana, K. ; Chandak, M. B.: Study of Various Video Annotation Techniques. In: *International Journal of Advanced Research in Computer and Communication Engineering* (2013), pp. 1–6

[Kim and Lee 2017] Kim, J. ; Lee, S.: Deep Learning of Human Visual Sensitivity in Image Quality Assessment Framework. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1969–1977

[Konstantinidis et al. 2021] Konstantinidis, F. K. ; Mouroutsos, S. G. ; Gasteratos, A.: The Role of Machine Vision in Industry 4.0: an automotive manufacturing perspective. In: *IST 2021*. Piscataway, NJ, USA : IEEE, 2021, pp. 1–6. – ISBN 978-1-7281-7371-9

[Kreuzberger et al. 2022] Kreuzberger, D. ; Kühl, N. ; Hirschl, S.: Machine Learning Operations (MLOps): Overview, Definition, and Architecture. In: *IEEE Access* 11 (2022), pp. 31866–31879

[Kumari 2023] Kumari, K.: How Data Annotation Differs from Labeling: Including various tools for annotating and labeling data. In: *Heartbeat* (2023)

[Lakshmanan et al. 2021] Lakshmanan, V. ; Görner, M. ; Gillard, R.: *Practical machine learning for computer vision: End-to-end machine learning for images*. First edition, second release. Bejing and Boston and Farnham and Sebastopol and Tokyo : O'Reilly, 2021. – 189–197 p. – ISBN 9781098102364

[Leite et al. 2020] Leite, L. ; Rocha, C. ; Kon, F. et al.: A Survey of DevOps Concepts and Challenges. In: *ACM Computing Surveys* 52 (2020), Nr. 6, pp. 1–35. – ISSN 0360-0300

[Lin and Nwe 2021] Lin, K. Z. ; Nwe, K. H.: Audio Annotation on Myanmar Traditional Boxing Video by Enhancing DT. In: *Journal of Advances in Information Technology* 12 (2021), Nr. 2, pp. 107–112. – ISSN 17982340

[Lin et al. 2020] Lin, T. ; Goyal, P. ; Girshick, R. et al.: Focal Loss for Dense Object Detection. In: *IEEE transactions on pattern analysis and machine intelligence* 42 (2020), Nr. 2, pp. 318–327

[Liu et al. 2021a] Liu, B. ; Ravikumar, P. ; Risteski, A.: Contrastive learning of strong-mixing continuous-time stochastic processes. In: *International Conference on Artificial Intelligence and Statistics*, 2021

[Liu et al. 2022] Liu, L. ; Guo, C. ; Xiang, Y. et al.: A Semisupervised Learning Framework for Recognition and Classification of Defects in Transient Thermography Detection. In: *IEEE Transactions on Industrial Informatics* 18 (2022), Nr. 4, pp. 2632–2640

[Liu et al. 2021b] Liu, X. ; Zhang, F. ; Hou, Z. et al.: Self-supervised Learning: Generative or Contrastive. In: *IEEE Transactions on Knowledge and Data Engineering* (2021), pp. 1. – ISSN 1041-4347

[Lu et al. 2022] Lu, H. ; Fei, N. ; Huo, Y. et al.: COTS: Collaborative Two-Stream Vision-Language Pre-Training Model for Cross-Modal Retrieval. In: *arXiv preprint arXiv:2204.07441* (2022), pp. 1–10

[Luss et al. 2021] Luss, R. ; Chen, P. ; Dhurandhar, A. et al.: Leveraging Latent Features for Local Explanations. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021), pp. 1–27

[Ly et al. 2022] Ly, S. T. ; Lin, B. ; Vo, H. Q. et al.: Multiplexed Immunofluorescence Brain Image Analysis Using Self-Supervised Dual-Loss Adaptive Masked Autoencoder. In: *arXiv preprint arXiv:2205.05194* (2022), pp. 1–17

[Lynn 2023] Lynn ; Medium (Ed.): *What exactly is the .pth file?. This article will give you a general. . .* . 2023. – URL https://medium.com/@lyl1617670866/what-exactly-is-the-pth-file-9a487044a36b. – Access Date: 24.05.2023

[Ma et al. 2021] Ma, S. ; Zhaoyang, Z. ; McDuff, D. et al.: Contrastive Learning of Global-Local Video Representations. In: *35th Conference on Neural Information Processing Systems (NeurIPS 2021)* (2021), pp. 1–16

[Makris and Simos 2014] Makris, C. ; Simos, M. A.: Novel Techniques for Text Annotation with Wikipedia Entities. In: Hancock, Edwin (Ed.) ; Bayro Corrochano, Eduardo (Ed.): *Progress in pattern recognition, image analysis, computer vision, and applications* Bd. 8827. Cham : Springer, 2014, pp. 508–518. – ISBN 978-3-319-12567-1

[Martin et al. 2014] Martin, D. ; Hanrahan, B. V. ; O'Neill, J. et al.: Being a turker. In: Fussell, Susan (Ed.) ; Lutters, Wayne (Ed.) ; Morris, Meredith R. (Ed.) ; Reddy, Madhu (Ed.): *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. New York, NY, USA : ACM, 2014, pp. 224–235. – ISBN 9781450325400

[Martinuzzi et al. 2018] Martinuzzi, A. ; Blok, V. ; Brem, A. et al.: Responsible Research and Innovation in Industry—Challenges, Insights and Perspectives. In: *Sustainability* 10 (2018), pp. 702

[Meira et al. 2016] Meira, J. ; Marques, J. ; Jacob, J. et al.: Video annotation for immersive journalism using masking techniques. In: Bessa, Maximino (Ed.) ; Gonçalves, Daniel (Ed.): *2016 23° Encontro Português de Computação Gráfica e Interação (EPCGI)*. Piscataway, NJ : IEEE, 2016, pp. 1–7. – ISBN 978-1-5090-5387-2

[Merritt 2023] Merritt, R.: *What is MLOps?* 2023. – URL https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/. – Access Date: 13.04.2023

[Microsoft 2023] Microsoft: *Machine Learning Operations-Framework (MLOps) zum Hochskalieren des Machine Learning-Lebenszyklus mit Azure Machine Learning*. 2023. – URL https://learn.microsoft.com/en-us/azure/architecture/example-scenario/mlops/mlops-technical-paper. – Access Date: 13.04.2023

[Mokady et al. 2021] Mokady, R. ; Hertz, A. ; Bermano, A.: ClipCap: CLIP Prefix for Image Captioning. In: *arXiv preprint arXiv:2111.09734* (2021)

[Mosqueira-Rey et al. 2023] Mosqueira-Rey, E. ; Hernández-Pereira, E. ; Alonso-Ríos, D. et al.: Human-in-the-loop machine learning: a state of the art. In: *Artificial Intelligence Review* 56 (2023), Nr. 4, pp. 3005–3054. – ISSN 0269-2821

[Mundhenk et al. 2017] Mundhenk, T. N. ; Ho, D. ; Chen, B.Y.: Improvements to Context Based Self-Supervised Learning. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), pp. 9339–9348

[Murphy 2012] Murphy, K.: Machine learning - a probabilistic perspective. In: *Adaptive computation and machine learning series*. 2012

[Newton 2023] Newton, E.: You Need to Know the Pros and Cons of Self-Supervised Learning. In: *IT Chronicles* (2023). – URL https://itchronicles.com/artificial-intelligence/you-need-to-know-the-pros-and-cons-of-self-supervised-learning/. – Access Date: 02.06.2023

[Nguyen et al. 2022] Nguyen, T. ; Jump, A. ; Casey, D.: Emerging Tech Emerging Tech Impact Radar: 2023. In: *Gartner Research* (2022), pp. 3–4. – URL https://www.gartner.com/en/articles/4-emerging-technologies-you-need-to-know-about

[Noman et al. 2019] Noman, M. ; Stankovic, V. ; Tawfik, A.: Object Detection Techniques: Overview and Performance Comparison. In: *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2019, pp. 1–5. – ISBN 978-1-7281-5341-4

[Novosel, J. and Viswanath, P. and Arsenali, B. 2019] Novosel, J. and Viswanath, P. and Arsenali, B.: Boosting semantic segmentation with multi-task self-supervised

learning for autonomous driving applications. In: *Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (2019), pp. 1–11

[Ohri and Kumar 2021] Ohri, K. ; Kumar, M.: Review on self-supervised image recognition using deep neural networks. In: *Knowledge-Based Systems* 224 (2021), pp. 107090. – ISSN 0950-7051

[Oleszak 2023] Oleszak, M.: Self-Supervised Learning in Computer Vision | Towards Data Science. In: *Towards Data Science* (2023), pp. 1–18

[Orhan et al. 2020] Orhan, A. E. ; Gupta, V. V. ; Lake, B. M.: Self-supervised learning through the eyes of a child. In: Larochelle, H. (Ed.) ; Ranzato, M. (Ed.) ; Hadsell, R. (Ed.) ; Balcan, M.F. (Ed.) ; Lin, H. (Ed.): *Advances in Neural Information Processing Systems* Bd. 33, Curran Associates, Inc., 2020, pp. 9960–9971

[Padma et al. 2019] Padma, P. ; Srinivasan, S. ; Ghayathri, J. et al.: A Decisive Object Detection using Deep Learning Techniques. In: *International Journal of Innovative Technology and Exploring Engineering* 9 (2019), Nr. 1S, pp. 414–417

[Pagare and Shinde 2012] Pagare, R. ; Shinde, A.: A Study on Image Annotation Techniques. In: *International Journal of Computer Applications* 37 (2012), Nr. 6, pp. 42–45

[Park et al. 2016] Park, J. ; Kwon, B. ; Park, J. et al.: Machine learning-based imaging system for surface defect inspection. In: *International Journal of Precision Engineering and Manufacturing-Green Technology* 3 (2016), Nr. 3, pp. 303–310. – ISSN 2288-6206

[Pen 2014] Pen, S. E.: Demand Effects of Changes of Income Distribution and Domestic Demand Issues——Based on Quantile Regression of Pseudo Panel Threshold Model. In: *Journal of Shanxi University of Finance and Economics* (2014), pp. 1–11

[Peng et al. 2020] Peng, Z. ; Dong, Y. ; Luo, M. et al.: Self-Supervised Graph Representation Learning via Global Context Prediction. In: *arXiv preprint arXiv:2003.01604* (2020)

[Penumuru et al. 2020] Penumuru, D. P. ; Muthuswamy, S. ; Karumbu, P.: Identification and classification of materials using machine vision and machine learning in the context of industry 4.0. In: *Journal of Intelligent Manufacturing* 31 (2020), Nr. 5, pp. 1229–1241. – ISSN 0956-5515

[Potrimba 2023] Potrimba, P.: Multimodal Models and Computer Vision: A Deep Dive. In: *Roboflow Blog* (2023). – URL https://blog.roboflow.com/multimodal-models/. – Access Date: 13.06.2023

[Potter 2023a] Potter, R.: How Data Annotation Companies Manually Label their Data? In: *ANOLYTICS* (2023). – URL https://medium.com/anolytics/how-data-annotation-companies-manually-label-their-data-b94d6afcca0b. – Access Date: 20.04.2023

[Potter 2023b] Potter, R.: What is Data Annotation and What are its Advantages? In: *ANOLYTICS* (2023). – URL https://medium.com/anolytics/what-is-data-annotation-and-what-are-its-advantages-95766213351e. – Access Date: 18.04.2023

[Priyanka 2018] Priyanka, K.: *Artificial Intelligence in Manufacturing Market by Deployment (Cloud and On-Premise), Technology (Machine Learning, Computer Vision, Context Awareness, and Natural Language Processing), Application (Material Movement, Predictive Maintenance & Machinery Inspection, Production Planning, Field Service, and Quality Control & Reclamation), and Industry (Semiconductor & Electronics, Energy & Power, Pharmaceutical, Automobile, Heavy Metal & Machine Manufacturing, and Others): Global Opportunity Analysis and Industry Forecast, 2018 - 2025*. 2018

[Rahman et al. 2020] Rahman, T. ; Chowdhury, M. E. H. ; Khandakar, A. et al.: Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray. In: *Applied Sciences* 10 (2020), Nr. 9. – URL https://www.mdpi.com/2076-3417/10/9/3233. – ISSN 2076-3417

[Rainer et al. 2022] Rainer, J. ; Vicini, A. ; Salzer, L. et al.: A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. In: *Metabolites* 12 (2022), Nr. 2. – ISSN 2218-1989

[Randive and Mohan 2020] Randive, K. ; Mohan, R.: A State-of-Art Review on Automatic Video Annotation Techniques. In: Abraham, Ajith (Ed.): *Intelligent Systems Design and Applications* Bd. 940. Cham : Springer International Publishing AG, 2020, pp. 1060–1069. – ISBN 978-3-030-16656-4

[Rani et al. 2023] Rani, V. ; Nabi, S. T. ; Kumar, M. et al.: Self-supervised Learning: A Succinct Review. In: *Archives of computational methods in engineering : state of the art reviews* 30 (2023), Nr. 4, pp. 2761–2775

[Redmon et al. 2015] Redmon, J. ; Divvala, S. ; Girshick, R. et al.: You Only Look Once: Unified, Real-Time Object Detection. In: *arXiv preprint arXiv:1506.02640* (2015)

[Rehman 2023] Rehman, S.: How To Implement an OKR Strategy for Your Organization | airfocus. In: *airfocus* (2023). – URL https://airfocus.com/product-learn/how-to-implement-okr-strategy-for-organization/. – Access Date: 15.06.2023

[Ren et al. 2021] Ren, P. ; Xiao, Y. ; Chang, X. et al.: A Survey of Deep Active Learning. In: *ACM Comput. Surv.* 54 (2021), Nr. 9, pp. 1–40. – ISSN 0360-0300

[Ren et al. 2022] Ren, Z. ; Fang, F. ; Yan, N. et al.: State of the Art in Defect Detection Based on Machine Vision. In: *International Journal of Precision Engineering and Manufacturing-Green Technology* 9 (2022), Nr. 2, pp. 661–691. – ISSN 2288-6206

[Rennie et al. 2016] Rennie, S. J. ; Marcheret, E. ; Mroueh, Y. et al.: Self-critical Sequence Training for Image Captioning. In: *arXiv preprint arXiv:1612.00563* (2016)

[Rizzoli 2021] Rizzoli, A.: *The Ultimate Guide to Object Detection*. 2021. – URL https://www.v7labs.com/blog/object-detection-guide. – Access Date: 16.06.2023

[Russell et al. 2010] Russell, S. J. ; Norvig, P. ; Davis, E.: *Artificial intelligence: A modern approach*. 3rd ed. Upper Saddle River : Prentice Hall, 2010 (Prentice Hall series in artificial intelligence)

[Ruyi et al. 2023] Ruyi, J. ; Jiaying, L. ; Libo, Z.: Siamese self-supervised learning for fine-grained visual classification. In: *Computer Vision and Image Understanding* 229 (2023), pp. 103658. – ISSN 1077-3142

[Sanghvi et al. 2020] Sanghvi, K. ; Aralkar, A. ; Sanghvi, S. et al.: A Survey on Image Classification Techniques. In: *SSRN Electronic Journal* (2020), pp. 1–5

[Schwenk et al. 2022] Schwenk, D. ; Khandelwal, A. ; Clark, C. et al.: A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. In: *arXiv preprint arXiv:2206.01718* (2022)

[Selvaraju et al. 2019] Selvaraju, R. R. ; Cogswell, M. ; Das, A. et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 128 (2019), Nr. 2, pp. 1

[Settles 2009] Settles, B.: Active Learning Literature Survey. In: *Journal of Biomedical Science and Engineering* 3 (2009), Nr. 10

[Shakya 2020] Shakya, S.: Analysis of Artificial Intelligence based Image Classification Techniques. In: *Journal of Innovative Image Processing* 2 (2020), Nr. 1, pp. 44–54

[Sharma et al. 2018] Sharma, P. ; Ding, N. ; Goodman, S. et al.: Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In: Gurevych, Iryna (Ed.) ; Miyao, Yusuke (Ed.): *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2018, pp. 2556–2565

[Shelhamer et al. 2016] Shelhamer, E. ; Long, J. ; Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. In: *arXiv preprint arXiv:1411.4038* (2016)

[Shepard and Metzler 1971] Shepard, R. N. ; Metzler, J.: Mental rotation of three-dimensional objects. In: *Science (New York, N.Y.)* 171 (1971), Nr. 3972, pp. 701–703. – ISSN 0036-8075

[Silva 2020] Silva, T. S.: Self-Supervised Learning and the Quest for Reducing Labeled Data in Deep Learning. In: *https://sthalles.github.io* (2020). – URL https://sthalles.github.io/self-supervised-learning/

[Simonyan et al. 2013] Simonyan, K. ; Vedaldi, A. ; Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: *CoRR* abs/1312.6034 (2013)

[Simonyan and Zisserman 2014] Simonyan, K. ; Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. In: Ghahramani, Z. (Ed.) ; Welling, M. (Ed.) ; Cortes, C. (Ed.) ; Lawrence, N. (Ed.) ; Weinberger, K.Q. (Ed.): *Advances in Neural Information Processing Systems* Bd. 27. Curran Associates, Inc., 2014

[Sparks 2023] Sparks, R.: *OKRs: the ultimate guide to objectives and key results*. 2023. – URL https://www.atlassian.com/agile/agile-at-scale/okr. – Access Date: 02.06.2023

[Steidl et al. 2023] Steidl, M. ; Felderer, M. ; Ramler, R.: The pipeline for the continuous development of artificial intelligence models—Current state of research and practice. In: *Journal of Systems and Software* 199 (2023), pp. 1–26. – ISSN 0164-1212

[Stenetorp et al. 2012] Stenetorp, P. ; Pyysalo, S. ; Topic, G. et al.: brat: a Web-based Tool for NLP-Assisted Text Annotation. In: *Conference of the European Chapter of the Association for Computational Linguistics*, 2012

[Sun et al. 2020] Sun, P. ; Kretzschmar, H. ; Dotiwalla, X. et al.: Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In: *arXiv preprint arXiv:1912.04838* (2020), pp. 1–9

[Sutter et al. 2021] Sutter, T. M. ; Daunhawer, I. ; Vogt, J. E.: Generalized Multimodal ELBO. In: *arXiv preprint arXiv:2105.02470* (2021)

[Tan et al. 2020] Tan, J. ; Song, C. ; Boularias, A.: A Self-supervised Learning System for Object Detection in Videos Using Random Walks on Graphs. In: *arXiv preprint arXiv:2011.05459* (2020)

[Thada et al. 2023] Thada, V. ; Shrivastava, U. ; Sharma, J. et al.: A Primer on Generative Adversarial Networks. In: *International Journal of Innovative Research in Computer Science & Technology* 8 (2023), Nr. 3, pp. 2–3

[Tong and Koller 2002] Tong, S. ; Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification. In: *J. Mach. Learn. Res.* 2 (2002), pp. 45–66

[Treneska et al. 2022] Treneska, S. ; Zdravevski, E. ; Pires, I. et al.: GAN-Based Image Colorization for Self-Supervised Visual Feature Learning. In: *Sensors (Basel, Switzerland)* 22 (2022), Nr. 4, pp. 1–17

[Treveil et al. 2021] Treveil, M. ; Omont, N. ; Stenac, C. et al.: *MLOps – Kernkonzepte im Überblick: Machine-Learning-Prozesse im Unternehmen nachhaltig automatisieren und skalieren.* Heidelberg : O'Reilly, 2021 (Animals). – ISBN 9783960105817

[Tripathi 2021] Tripathi, M.: Analysis of Convolutional Neural Network based Image Classification Techniques. In: *Journal of Innovative Image Processing* 3 (2021), Nr. 2, pp. 100–117

[Tsutsui et al. 2021] Tsutsui, S. ; Desai, R. ; Ridgeway, K.: How You Move Your Head Tells What You Do: Self-supervised Video Representation Learning with Egocentric Cameras and IMU Sensors. In: *arXiv preprint arXiv:2110.01680* (2021), pp. 1–4

[Vijayrania 2023] Vijayrania, N.: Self-Supervised Learning Methods for Computer Vision. In: *Towards Data Science* (2023). – URL https://towardsdatascience.com/self-supervised-learning-methods-for-computer-vision-c25ec10a91bd. – Access Date: 02.06.2023

[Vocaturo 2021] Vocaturo, E.: Image Classification Techniques. In: Rani, Geeta (Ed.) ; Tiwari, Pradeep K. (Ed.): *Handbook of Research on Disease Prediction Through Data Analytics and Machine Learning.* IGI Global, 2021 (Advances in Medical Diagnosis, Treatment, and Care), pp. 22–49. – ISBN 9781799827429

[Vondrick et al. 2018a] Vondrick, C. ; Shrivastava, A. ; Fathi, A. et al.: Tracking Emerges by Colorizing Videos. In: *arXiv preprint arXiv:1806.09594* (2018), pp. 3

[Vondrick et al. 2018b] Vondrick, C. ; Shrivastava, A. ; Fathi, A. et al.: Tracking Emerges by Colorizing Videos. In: Ferrari, Vittorio (Ed.) ; Hebert, Martial (Ed.) ; Sminchisescu, Cristian (Ed.) ; Weiss, Yair (Ed.): *Computer vision - ECCV 2018* Bd. 11217. Cham : Springer, 2018, pp. 402–419. – ISBN 978-3-030-01260-1

[Wang et al. 2020] Wang, Q. ; Ma, Y. ; Zhao, K. et al.: A Comprehensive Survey of Loss Functions in Machine Learning. In: *Annals of Data Science* (2020), pp. 1–26

[Wang et al. 2022a] Wang, Q. ; Ma, y. ; Zhao, K. et al.: A Comprehensive Survey of Loss Functions in Machine Learning. In: *Annals of Data Science* 9 (2022), Nr. 2, pp. 187–212. – ISSN 2198-5804

[Wang et al. 2022b] Wang, Z. ; Feng, Z. ; Li, Y. et al.: BatmanNet: Bi-branch Masked Graph Transformer Autoencoder for Molecular Representation. In: *arXiv preprint arXiv:2211.13979* (2022), pp. 1–11

[Wennker 2020] Wennker, P.: *Künstliche Intelligenz in der Praxis: Anwendungen in Unternehmen und Branchen: KI wettbewerbs- und zukunftsorientiert einsetzen.* Wiesbaden, Germany and Heidelberg : Springer Gabler, 2020. – ISBN 978-3-658-30479-9

[Xue-wu et al. 2011] Xue-wu, Z. ; Yan-qiong, D. ; Yan-yun, L. et al.: A vision inspection system for the surface defects of strongly reflected metal based on multi-class SVM. In: *Expert Systems with Applications* 38 (2011), Nr. 5, pp. 5930–5939. – ISSN 09574174

[Yang et al. 2022] Yang, F. ; Wu, K. ; Zhang, S. et al.: Class-Aware Contrastive Semi-Supervised Learning. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14401–14410

[Ye et al. 2023] Ye, T. ; Chen, S. ; Liu, Y. et al.: Towards Real-Time High-Definition Image Snow Removal: Efficient Pyramid Network with Asymmetrical Encoder-Decoder Architecture. In: *Computer Vision – ACCV 2022.* Cham : Springer Nature Switzerland, 2023, pp. 37–51. – ISBN 978-3-031-26313-2

[Yu and Chen 2022] Yu, H. ; Chen, L.: A Joint Loss based Deep Learning Model for Enhanced Image Inpainting. In: Fortino, Giancarlo (Ed.): *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing/Intl Conf on Pervasive Intelligence and Computing/Intl Conf on Cloud and Big Data Computing/Intl Conf on Cyber Science and Technology Congress.* Piscataway, NJ : IEEE, 2022, pp. 1–6. – ISBN 978-1-6654-6297-6

[Zhang et al. 2021a] Zhang, C. ; Bengio, S. ; Hardt, M. et al.: Understanding Deep Learning (Still) Requires Rethinking Generalization. In: *Commun. ACM* 64 (2021), feb, Nr. 3, pp. 107–115. – ISSN 0001-0782

[Zhang et al. 2022] Zhang, C. ; Zhang, C. ; Song, J. et al.: A Survey on Masked Autoencoder for Self-supervised Learning in Vision and Beyond. In: *arXiv preprint arXiv:2208.00173* (2022)

[Zhang and Shen 2022] Zhang, K. ; Shen, Z.: i-MAE: Are Latent Representations in Masked Autoencoders Linearly Separable? In: *arXiv preprint arXiv:2210.11470* (2022), pp. 1–18

[Zhang et al. 2021b] Zhang, L. ; Amgad, M. ; Cooper, L. A. D.: A Histopathology Study Comparing Contrastive Semi-Supervised and Fully Supervised Learning. In: *arXiv preprint arXiv:2111.05882* abs/2111.05882 (2021), pp. 1–7

[Zhang et al. 2013] Zhang, L. ; Rettinger, A. ; Färber, M. et al.: A Comparative Evaluation of Cross-Lingual Text Annotation Techniques. In: Forner, Pamela (Ed.) ; Müller, Henning (Ed.) ; Paredes, Roberto (Ed.) ; Rosso, Paolo (Ed.) ; Stein, Benno (Ed.): *Information access evaluation* Bd. 8138. Berlin and Heidelberg : Springer, 2013, pp. 124–135. – ISBN 978-3-642-40801-4

[Zhang et al. 2017] Zhang, R. ; Zhu, J. ; Isola, P. et al.: Real-Time User-Guided Image Colorization with Learned Deep Priors. In: *arXiv preprint arXiv:1705.02999* (2017)

[Zheng et al. 2022a] Zheng, L. ; Puy, G. ; Riccietti, E. et al.: Self-supervised learning with rotation-invariant kernels. In: *The Eleventh International Conference on Learning Representations* (2022), pp. 1–27

[Zheng et al. 2022b] Zheng, X. ; Wang, B. ; Du, X. et al.: Mutual Attention Inception Network for Remote Sensing Visual Question Answering. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–14. – ISSN 0196-2892

[Zhou et al. 2021] Zhou, L. ; Zhang, L. ; Konz, N.: Computer Vision Techniques in Manufacturing. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 53 (2021), pp. 105–117

[Zhou et al. 2023] Zhou, L. ; Zhang, L. ; Konz, N.: Computer Vision Techniques in Manufacturing. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 53 (2023), Nr. 1, pp. 105–117. – ISSN 2168-2216

[Zhou et al. 2022] Zhou, P. ; Zhou, Y. ; Si, C. et al.: Mugs: A Multi-Granular Self-Supervised Learning Framework. In: *arXiv preprint arXiv:2203.14415* (2022)

[Zhou et al. 2019] Zhou, Y. ; Jun, Y. ; Yuhao, C. et al.: Deep Modular Co-Attention Networks for Visual Question Answering. In: *arXiv preprint arXiv:1906.10770* (2019)

Eigenständigkeitserklärung

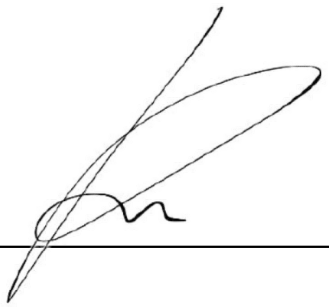
Ich versichere hiermit, dass ich meine Bachelorarbeit mit dem Thema

Data Annotation Concept
based on Self-Supervised Learning
for Computer Vision Developments
in Manufacturing

selbständig verfasst und keine anderen als die angegebenen
Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte
elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ravensburg, 29.06.2023

Ort, Datum



Unterschrift