Are Foundation Models Ready for Industrial Defect Recognition? A Reality Check on Real-World Data

Simon Baeuerle[1,3,*], Pratik Khanna[1,3,*], Nils Friederich[1,2], Angelo Jovin Yamachui Sitcheu[1], Damir Shakirov[4], Andreas Steimer[4], Ralf Mikut[1]

[1] Institute for Automation and Applied Informatics (IAI)
Karlsruhe Institute of Technology
[2] Institute of Biological and Chemical Systems (IBCS)
Karlsruhe Institute of Technology
[3] Mobility Electronics, Robert Bosch GmbH
[4] Bosch Center for Artificial Intelligence, Robert Bosch GmbH
*shared first, e-mail: simon.baeuerle@kit.edu

Abstract

Foundation Models (FMs) have shown impressive performance on various text and image processing tasks. They can generalize across domains and datasets in a zero-shot setting. This could make them suitable for automated quality inspection during series manufacturing, where various types of images are being evaluated for many different products. Replacing tedious labeling tasks with a simple text prompt to describe anomalies and utilizing the same models across many products would save significant efforts during model setup and implementation. This is a strong advantage over supervised Artificial Intelligence (AI) models, which are trained for individual applications and require labeled training data. We test multiple recent FMs on both custom real-world industrial image data and public image data. We show that all of those models fail on our real-world data, while the very same models perform well on public benchmark

datasets.

1 Introduction

Defect pattern recognition is carried out during quality inspection in automotive series manufacturing. During image-based quality inspection, images are recorded for a very wide range of manufacturing processes. Beyond color images, advanced inspection technologies yield, e.g., depth images, x-ray images or Scanning Acoustic Tomography (SAT) images. The recording procedure is set up to capture defects that are specific to the respective manufacturing process. This ensures high product quality and prevents defective products from being delivered to customers. Sometimes, further expensive processing steps of a defective product can also be saved. Due to the high volume of manufactured parts, an automated defect recognition procedure can significantly reduce manual inspection efforts. Here, Artificial Neural Networks (ANNs) are increasingly used besides classic image processing techniques [1, 2]. However, a significant drawback of ANN-based classifiers is their reliance on large amounts of manually labeled training data. While this effort can be mitigated by unsupervised AI approaches [3], it remains a considerable hurdle. At the same time, simpler methods, such as a direct comparison to a reference image, are often insufficient. Real-world challenges like manufacturing tolerances, image registration errors, or varying brightness prevent such an approach from reliably detecting defects.

These limitations of both supervised models and simple heuristics motivate the exploration of a new class of powerful, pre-trained models. On various text and image processing tasks in other domains, FMs have recently shown impressive performance [4–6]. For example, the Segment Anything Model (SAM) and Contrastive Language-Image Pre-training (CLIP) generalize very well across many domains and datasets in a zero-shot setting, significantly reducing labeling efforts. While CLIP accepts small text inputs, Large Vision and Language Models (LVLMs) like Gemini [11] are more powerful at the simultaneous processing of image and text input. Domain experts can formulate text prompts without

significant effort. Promising ideas for prompting during quality inspection are, e.g., a description of the normal state of the product or a description of the visual or physical properties of the defects. These reduced labeling efforts, combined with the additional opportunity to integrate domain expert knowledge via text input, would enable easier scaling across several products, i.e., with significantly lower efforts for each product. The seemingly clear advantages over State Of The Art (SOTA) approaches motivate the question of how suitable FMs are for image-based quality inspection tasks. To close these gaps for industrial use cases, we analyze the applicability of various recent FMs on real-world industrial data in this work. The achieved performance is compared to the performance of the same models on a public dataset.

The main objective during quality inspection is the distinction of defective and defect-free products. A classification model can perform such an inspection task with minimal setup. However, it might not cover all desired functionalities fully: In some cases, the classification of an AI model is re-checked by a human operator. Furthermore, a high level of explainability is preferred during model monitoring. A model that outputs a full segmentation mask as opposed to only a single class makes both manual re-checking and model monitoring easier. As such, we include both classification and segmentation models in our study.

The remainder of this work is structured as follows: The following section contains an overview of related work regarding e.g. FMs and SOTA pipelines for image segmentation. The datasets that we use for benchmarking are introduced in Section 3. The setup of our experiments is described in Section 4 and respective results are shown in Section 5. Section 6 contains our interpretation of results and potential implications.

2 Related Work

Several studies have investigated FMs for computer vision. This section contains an overview of the most relevant models and studies in related applications.

Radford et al. [6] propose the CLIP approach, which jointly trains an image encoder and a text encoder. It is trained on a dataset, which consists of 400 million pairs of images and corresponding text descriptions. CLIP can perform zero-shot inference on unseen objects, i.e., it can be applied to new datasets without any finetuning. It outperforms a SOTA supervised approach on multiple datasets [6].

The SAM [5] consists of an image encoder, a prompt encoder and a mask decoder. The image encoder uses a Vision Transformer (ViT) [7], which is pre-trained using the Masked Autoencoder approach [8]. The prompt encoder encodes prompts like points, boxes, text or masks, whereas the prompt encoder for text uses the text encoder from CLIP. The mask decoder creates masks based on the image embeddings and the prompt embeddings. The SAM model is trained on the SA-1B dataset, which contains over one billion annotated masks on 11 million images. This covers a wide range of different objects, locations and scenarios. Images of people, buildings, vehicles, animals, and other elements from everyday life around the world are well represented.

Li et al. [9] propose CLIPSurgery, which introduces small adaptations to the CLIP architecture. They remove redundant features and modify the attention mechanism to link semantically similar regions better. This significantly improves the model's explainability.

Liu et al. [10] introduce GroundingDINO, which can detect objects based on text input. The feature enhancer includes cross-attention between text and image information. The language-guided query selection selects features that match the input text. The cross-modality decoder fuses the text modality with the image modality for the generation of the output regions. It is pretrained on large datasets. The prediction head outputs multiple object bounding boxes along with a text and a similarity score. A box threshold can be set to only include bounding boxes with a minimum score. A text threshold can be set to additionally filter for bounding boxes that match the given input text prompt well.

LVLMs are models that accept multiple modalities such as text, images, audio and/or video as input. They can be utilized in a wide range of applications. Gemini 2.5 Pro [11] can be directly applied to defect classification tasks, since

it can process a text prompt along with an image. It has furthermore shown strong reasoning capabilities.

Zhang et al. [12] propose different pipelines of multiple foundation models, e.g., GroundingDINO + SAM, SAM + CLIP or CLIPSurgery + SAM. GroundingDINO and CLIP can effectively capture semantic features and provide visual prompts for subsequent instance segmentation by SAM. Zhang et al. analyze the different pipelines on aerial images as taken by, e.g., an Unmanned Aerial Vehicle.

Cao et al. [13] propose the framework Segment Any Anomaly + (SAA+) for zero-shot anomaly segmentation. This includes GroundingDINO to propose abnormal regions which are then fed into a second model, such as SAM to refine the abnormal regions. They specifically study defect detection and analyze performance on the public industrial dataset *MVTec AD*.

Xu et al. [14] combine human expert knowledge with the capabilities of Visual-Language-Foundation Models. They study different prompts for the task of anomaly detection in images. This includes simple prompts that only query for any defect or anomaly and more advanced prompts that provide, e.g., more detailed information about the shown object or the expected defects.

The Text2Seg approach contains multiple of the most relevant FM models. They have been successfully tested on an image segmentation task. SAA+ follows a similar approach to Text2Seg and its authors even include the task of defect recognition into their study. As outlined above, the Gemini model is widely applicable and has shown promising capabilities. This raises high expectations towards the application of those models during industrial defect recognition. While there is a wide range of further models that would also be interesting in this context, we limit the scope of our study to these models.

3 Datasets

During this research, we utilized three different image datasets: our custom real-world industrial dataset *IndustrialSAT*, the public industrial dataset *MVTec*

AD and the general public image dataset *Oxford-IIIT-Pet*. The main focus of our study is the real-world data. The public datasets serve as a reference.

3.1 Oxford-IIIT-Pet

Figure 1: Exemplary images from the dataset *Oxford-IIIT-Pet* [15]

The dataset *Oxford-IIIT-Pet* [15] contains images of various pet animals in different everyday scenarios. The animals are recorded both outdoors and indoors in different scales, pose and lighting. Such animal images are found across the internet and are well-represented in the SA-1B dataset. During this work, only cats and dogs are utilized. The dog images are treated as defect-free images ("OK"), whereas cat images are treated as defective ("NOK"). During testing, we utilize 140 dog images and 60 cat images. Examples are shown in Figure 1.

3.2 MVTec AD

The *MVTec AD* dataset [16] is utilized during the benchmarking of various models for industrial anomaly detection [17, 18]. In this work, we analyze only object categories that are visually related to our domain of electronic packaging. This includes the categories carpet, grid, leather, tile and wood with a resulting test dataset of 253 defect-free images ("OK") and 76 defective images ("NOK").

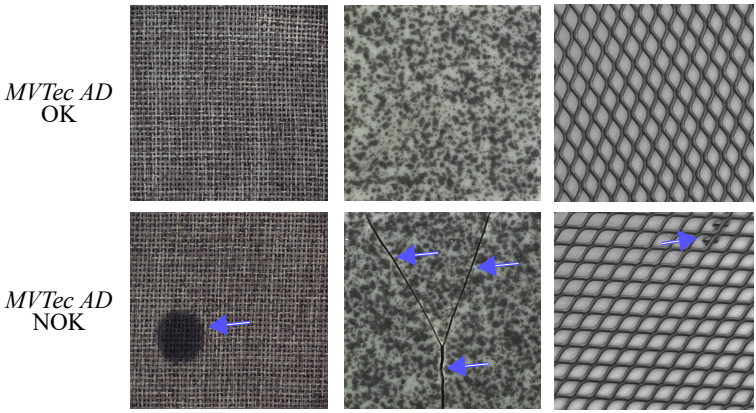

Figure 2: Exemplary images from the dataset *MVTec AD* [16] from the classes carpet (left), tile (center) and grid (right). Top row: defect-free images. Bottom row: defective images.

No distinction is made between different defect types. Examples from the categories carpet, tile and grid are shown in Figure 2.

3.3 IndustrialSAT

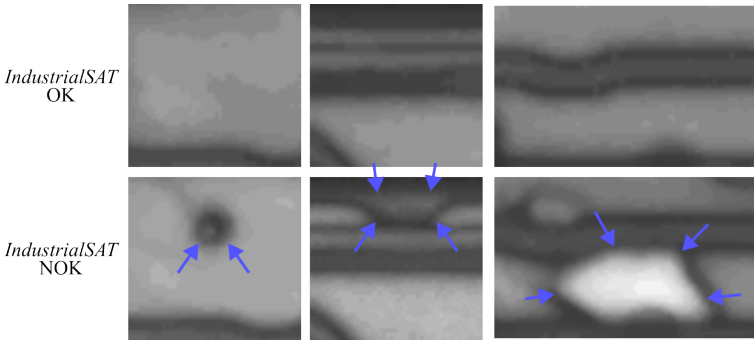

Figure 3: Exemplary images from the dataset *IndustrialSAT*. Crop of defective images (bottom row) and corresponding regions without defect (top row).

Our custom real-world industrial dataset *IndustrialSAT* is a dataset of greyscale images. They are recorded with SAT [19,20] as follows. The electronic package

is submerged in water. A toolhead with an ultrasonic sender and receiver is moved to a position above the package. An ultrasonic wave is sent out and the signal reflection is recorded. This yields a time-series of data points for this position. From the time-series, the signal value at a certain time value is extracted with special postprocessing methods. This outputs a single value for this position of the toolhead. The procedure is repeated for multiple toolhead positions. The toolhead positions correspond to the pixels of the resulting image. The extracted signal values correspond to the grey values of the resulting image. SAT is used for quality inspection during electronic packaging of products in the field of, e.g., sensors, electronic control units or power electronics. These products are highly relevant for electric and autonomous vehicles. Examples of defect types that could occur during electronic packaging are cracks, voids or delaminations between or within the different layers of an electronic package. Our test dataset, *IndustrialSAT*, contains 231 defect-free images and 32 images with one of the defect types: void, crack or delamination. Exemplary defects on SAT images are shown in the bottom row of Figure 3. The shown images are cropped down to the immediate area around the defects. The top row of Figure 3 shows the same crop position of a non-defective product. To evaluate our model, we assigned the defective label to images with defects of any size. No distinction is made with respect to different defect types during model evaluation.

During a preliminary study, we trained a classification model on the dataset *IndustrialSAT*. This was built by concatenating a pre-trained feature extractor based on a ResNet model [21] with a multi-layer perceptron. This classifier reached an F2-score of 0.82 on a stratified five-fold cross-validation dataset split. This proves that the defects on this dataset (such as visualized in Figure 3) can be recognized by a Machine Learning (ML) model.

The SAT images during manufacturing look similar to each other at first sight. During another preliminary study, we tested a simple approach: we aligned all images with respect to a predefined reference image. Then, we calculated the difference on the pixel level with respect to the reference image. However, this was not sufficient to detect defects such as voids or delaminations. The variation during the SAT recording, along with the variation in mechanical tolerances, was too large.

4 Methodology

We test multiple FMs both on the task of segmentation and the task of classification on the above datasets. Our experiments are set up as follows.

4.1 Segmentation

For the task of segmentation, we use the Intersection over Union (IOU) as a metric to assess the segmentation output:

$$IoU = \frac{|A \cap B|}{|A \cup B|}, \tag{1}$$

with A being the set of pixels predicted positive by the model and B being the set of pixels marked positive in the ground-truth. The IOU metric yields values between zero and one, whereas one corresponds to an optimum result. This metric penalizes especially the slip of small defects (=false negatives): The correctly detected defective area (intersection of defective prediction and ground truth) is divided by the union of the defective prediction and the ground truth. A miss of a small defect, such as a void, will thus return low metric values. Such false negatives are especially critical since defective products would be delivered to customers.

Three pipelines are chosen from Text2Seg for evaluation: GroundingDINO + SAM, SAM + CLIP and CLIPSurgery + SAM. The Git repository of the CLIPSurgery paper is used both for the CLIP Surgery model and the CLIP model. The original Git repository by IDEA research [24] is used for the GroundingDINO model. All of our Text2Seg pipelines use the huge backbone of SAM. The base VIT is used for both CLIP and CLIP Surgery with weights *CS-ViT-B/16* [22] for CLIP Surgery and weights *ViT-B/16* [23] for CLIP. The checkpoint *groundingdino swint ogc* [24] is used for the GroundingDINO model, along with a box threshold of 0.35 and a text threshold of 0.25. Minor issues such as the handling of empty segmentation masks, are resolved to enable an automated end-to-end evaluation in all cases. The utilized text prompts are short and simple, e.g., “defect” or “cat”.

For SAA+, we use GroundingDINO with weights *groundingdino swint ogc* as the region proposal network and a SAM model with the huge backbone as region refiner. The box threshold is set to 0.1 and the text threshold to 0.1. While the authors of Text2Seg evaluate their approach on the task of semantic segmentation in remote sensing images as taken, e.g., by an Unmanned Aerial Vehicle (UAV), the authors of SAA+ focus specifically on the segmentation of anomalies and also utilize the *MVTec AD* dataset during evaluation.

Since initial tests for the segmentation models have shown insufficient performance on both the real-world *IndustrialSAT* dataset and the public industrial *MVTec AD* dataset, we extend the evaluation for those models to include the more general public dataset *Oxford-IIIT-Pet*. This is done to validate the correct setup of our software pipelines and to gain additional insights.

The chosen model pipelines represent a diverse setup of multiple prominent SOTA FMs. The models are all executed in their inference mode and are tested on the defective images of all three datasets that were introduced in the previous section. The reported IoU metric values are averaged over all defective images. This yields an insight into how well defects can be recognized in the different datasets.

4.2 Classification

For the task of classification, we analyze the LVLM model Gemini 2.5 Pro. It features a high token limit, which enables long prompts. Furthermore, it can process multi-modal input, i.e., it can receive prompts consisting of both written text and images. The model's output includes a single class label and a corresponding reasoning string. The reasoning string is more verbose and gives a deeper insight into how the model has made a decision, whereas the class label can be used during automated postprocessing. For each dataset, a reference image is defined. This reference is used to automatically align all images, ensuring the same position and orientation of the product in all images.

The testing is carried out using the Google Cloud Platform, which offers a convenient access to the model via an Application Programming Interface

(API). It is carried out using both defect-free and defective images from the industrial datasets *IndustrialSAT* and *MVTec AD*. Initial experiments have shown promising performance on industrial data. Thus, an assessment on the general public dataset *Oxford-IIIT-Pet* is omitted, since the aim of this paper is to assess model performance on real-world industrial data.

The F2-score is used as a metric during evaluation:

$$F_2 = \frac{5 \cdot \text{precision} \cdot \text{recall}}{4 \cdot \text{precision} + \text{recall}} \tag{2}$$

Precision is the ratio of true positives (defects detected as defects) divided by the sum of true and false positives. Recall is the ratio of true positives divided by the sum of true positives and false negatives. As compared to F1-score, the F2-score puts a higher weight on recall than on precision. This penalizes false negatives (slips of defective products) more heavily compared to false positives (wrong alerts for defect-free products) [26]. This is desired, since it is better to re-check a suspicious product then to deliver a defective product to the customer.

When querying the Gemini model, a test image with an unknown state is sent to the model along with a prompt as described below. The prompt may include a predefined reference image. Two types of prompts were defined for the Gemini model. The basic prompt consists of a) a simple text prompt that queries for any anomalies or defects and b) an exemplary defect-free image. The refined prompt additionally includes specific information about the physical structure of e.g. the electronic package in *IndustrialSAT* or the object properties in *MVTec AD* and information about the visual appearance of the respective defects. This follows the conceptual approach of varying information depth in prompts by Xu et al. [14]. On *MVTec AD*, the prompts are as follows. The basic prompt for *MVTec AD* is “Please determine whether the image contains anomalies or defects. If yes, give a specific reason” - similar to the naive prompt studied by Xu et al. [14]. The refined prompt for *MVTec AD* is “Please determine whether the last image given about object contains anomalies or defects. If so, please provide a specific reason. Normally, the image given should depict a clear and identifiable object. It may have defects such as broken/bented parts, contaminations, threads, color stains, cuts, holes, scratches, liquids, glue, folds,

pokes, oil or glue strips.". The formulation of the prompts for *IndustrialSAT* is made in a close alignment with process engineers, who are responsible for the respective packaging processes. These prompts include specifics of the respective product and thus may not be published in full detail.

5 Results

This section contains the benchmarking results of both the segmentation and classification models on the different datasets as introduced before.

5.1 Segmentation

The IoU metric values for the segmentation models are reported in Table 1. Essentially, the defects in the SAT images cannot be detected in any of the tested pipelines. Visual examples are shown in Figure 4 for each model to give deeper insights. The pipeline based on GroundingDINO + SAM seems to be very sensitive and segments large defect-free regions of the image. For the other pipelines, the defects are missed entirely. In some cases, the pipeline CLIPSurgery + SAM segments geometric features of sub-components instead of defective regions. For the *MVTec AD* dataset, metric values are rather low. A visual depiction of results in Figure 5 shows that some defects can be detected accurately. For the public dataset *Oxford-IIIT-Pet*, the IoU metric results are promising, with IoU scores ranging from 0.65 to 0.80. A visual depiction of results in Figure 6 shows that the cat can be segmented very well by all models. We conducted a thorough inspection of various images and report that the model's segmentation results align with the depicted cat even more closely than the ground truth masks. Since the cat is configured to represent defective regions, this validates the correct setup of all pipelines for the segmentation task. Furthermore, we report a correct segmentation of other animals (e.g. dogs). Only IoU scores larger than 0.60 are considered for a comparison between models (indicated in Table 1 by bold print), since lower values are too low to offer practical value.

Table 1: IoU scores on all datasets for various FMs pipelines. None of the methods can detect defects on our internal dataset *IndustrialSAT*. Only SAA+ works somehow reasonable on the public *MVTec AD* dataset. All segmentation models work well on the public dataset *Oxford-IIIT-Pet*.

Classes and datasets	**GroundingDINO + SAM**	**SAM + CLIP**	**CLIPSurgery + SAM**	**SAA+**
Defect types void, crack and delamination in dataset *IndustrialSAT*	0.00	0.00	0.00	0.00
Various defects such as cracks, holes, scratches in dataset *MVTec AD*	0.13	0.05	0.19	0.52
Animal cat in dataset *Oxford-IIIT-Pet*	**0.80**	0.65	0.78	0.77

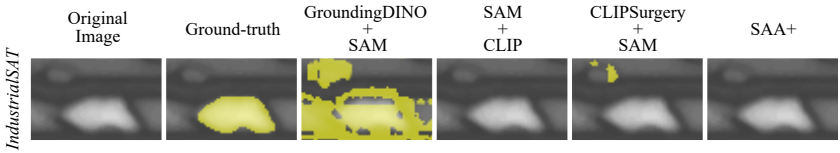

Figure 4: Exemplary segmentation results on dataset *IndustrialSAT*. A yellow color overlay is used to indicate defective regions as defined in ground-truth data or as predicted by the respective model. All models fail to detect the defects in most images.

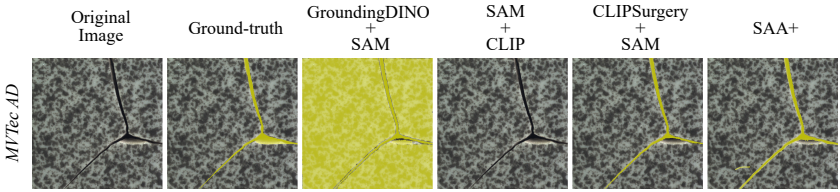

Figure 5: Exemplary segmentation results on dataset *MVTec AD* [16]. A yellow color overlay is used to indicate defective regions as defined in ground-truth data or as predicted by the respective model. In some cases, defects can be accurately recognized. In other cases, the models fail to recognize the defects.

Figure 6: Exemplary segmentation results on dataset *Oxford-IIIT-Pet* [15]. A yellow color overlay is used to indicate defective regions as defined in ground-truth data or as predicted by the respective model. Animals such as the cat shown in this image can be segmented very accurately by all tested models.

5.2 Classification

Results for the classification use case are shown in Table 2. The F2-score values are far too low to be suitable for industrial applications during quality inspection. However, the model can recognize some defects accurately, both with the basic and refined prompt strategy. Also, we report that the reasoning string output looks promising and shows that the LVLM model was able to recognize the package build-up to a certain extent.

The Gemini model performs significantly better on the public industrial dataset *MVTec AD* as compared to our internal industrial dataset *IndustrialSAT*. Interestingly, the highest metric results were achieved with simple prompts, which consist of a naive text prompt for defects along with a defect-free reference image. The maximum F2-score of 0.37 achieved by the Gemini model on *IndustrialSAT* is far below the F2-score of 0.82 reached with a supervised AI model in a preliminary study.

In analogy to the *Oxford-IIIT-Pet* dataset on the segmentation task, the good performance on the *MVTec AD* dataset validates the correct setup of our evaluation pipeline for the classification task.

6 Discussion

The three tested datasets can be ordered according to their visual similarity to datasets like SA-1B that are used to train the FMs: *Oxford-IIIT-Pet* is in a very

Table 2: Classification results (F2-score) on both industrial datasets for the Gemini 2.5 Pro LVLM model. While the model performs well on the *MVTec AD* dataset, its performance on SAT imaging data is insufficient for industrial-quality inspection.

Dataset	**Gemini** (basic prompt)	**Gemini** (refined prompt)
IndustrialSAT	0.37	0.31
MVTec AD	0.99	0.92

similar visual domain as SA-1B. *MVTec AD* is more focused on defects and includes industrially relevant parts, but some of its object categories, such as carpets and wood, are likely to be found in, e.g., SA-1B. However, our real-world dataset, *IndustrialSAT*, differs significantly from datasets like SA-1 B. It is not a color image, but rather a greyscale image. The geometric features that are shown resemble mostly geometric primitives, such as rectangular areas and line features. The performance of all the FMs is decreasing along with the visual similarity of the test dataset to training datasets such as SA-1B. The domain gap seems to be significant, especially for the real-world dataset *IndustrialSAT*. This is likely to be the cause of the insufficient performance on our real-world industrial data.

Furthermore, the performance difference on the datasets *MVTec AD* and *Industrial*SAT motivates the question of how well *MVTec AD* can represent advanced imaging procedures such as SAT. The *MVTec AD* dataset contains industrially relevant defects, but it is in a similar visual domain as public datasets. For example, screws and fabric-like materials can be found in everyday items, which are likely to be included in the training datasets of FMs to a certain extent. Defect patterns such as voids and cracks in SAT imaging are not well represented.

The reasoning strings of the Gemini model show a deeper understanding of the product build-up. This may not be fully correct, but it gives an interesting starting point for further studies. If this knowledge is available at least partially for the model, more advanced prompt strategies could leverage this knowledge better. The Gemini model was the only model that could detect some of the defects and a more refined prompting strategy seems to be a promising way to

improve performance.

7 Conclusion and outlook

Industrial quality inspection regularly involves advanced imaging procedures such SAT. In this work, we investigate the use of FMs to improve industrial defect recognition. None of the tested FMs were able to detect defects in our real-world SAT images. The results on other public datasets validate the correct setup of our models. While an application of FMs to such data could be scaled across many manufacturing stations, currently available FMs are not ready for practical application in series manufacturing for advanced inspection technologies such as SAT imaging.

The low performance in *IndustrialSAT* may be due to a large domain gap between the training data of FMs. Future work could thus involve fine-tuning FMs on image data that better matches the real-world industrial domain. Furthermore, more advanced prompt strategies for LVLMs seem to be a promising direction for an in-depth follow-up study.

8 Acknowledgments

The Helmholtz Association funds this project under the research school "Helmholtz Information and Data Science School for Health (HIDSS4Health)", the program "Natural, Artificial and Cognitive Information Processing (NACIP)" and the Initiative and Networking Fund on the HAICORE@KIT partition. This work was funded by the European Union (NextGenerationEU) and the German Federal Ministry for Economic Affairs and Energy (BMWE) based on a decision by the German Bundestag. The authors acknowledge financial support within the SmartMan project (reference number: 13IK033) in the Kopa35c program.

We describe the contributions of Simon Baeuerle (SB), Pratik Khanna (PK), Nils Friederich (NF), Angelo Jovin Yamachui Sitcheu (AJYS), Damir Shakirov

(DS), Andreas Steimer (AS) and Ralf Mikut (RM) according to CRediT [25]: Writing-Original Draft: SB; Writing-Review & Editing: PK, NF, AJYS, DS, AS, RM; Conceptualization: SB, RM; Investigation: PK (within masters thesis [27]); Methodology: SB, PK, NF, AJYS, DS, AS, RM; Software: PK; Supervision: SB, RM; Project administration: SB, RM; Funding Acquisition: SB, RM.

References

[1] Zihan He, Yudong Lian, Yulei Wang and Zhiwei Lu. A comprehensive review of research on surface defect detection of PCBs based on machine vision. *Results in Engineering*, 27:106437, 2025.

[2] Prahar M. Bhatt, Rishi K. Malhan, Pradeep Rajendran, Brual C. Shah, Shantanu Thakar, Yeo Jung Yoon and Satyandra K. Gupta. Image-based surface defect detection using deep learning: a review. *Journal of Computing and Information Science in Engineering*, 21(4):040801, 2021.

[3] Alina Pleli, Simon Baeuerle, Michel Janus, Jonas Barth, Ralf Mikut and Hendrik P. A. Lensch. Iterative Cluster Harvesting for wafer map defect patterns Technical report, arXiv:2404.15436, 2024. URL `http://arxiv.org/abs/2404.15436`.

[4] Rishi Bommasani et al. On the opportunities and risks of foundation models. Technical report, arXiv:2108.07258, August 2021. URL `https://arxiv.org/pdf/2108.07258`.

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. Technical report, arXiv:2304.02643, April 2023. URL `http://arxiv.org/abs/2304.02643`.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning

transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, New Orleans, LA, USA, June 2022. IEEE. doi:10.1109/CVPR52688.2022.01553. URL https://ieeexplore.ieee.org/document/9879206/.

[9] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of Contrastive language-image pre-training. *Pattern Recognition*, 162:111409, June 2025. ISSN 00313203. doi:10.1016/j.patcog.2025.111409. URL https://linkinghub.elsevier.com/retrieve/pii/S003132032500069X.

[10] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15105, pages 38–55. Springer Nature Switzerland, Cham, Switzerland, 2025. Series Title: Lecture Notes in Computer Science.

[11] Google Gemini Team. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and

next generation agentic capabilities. Technical report, Google, 2025. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf.

[12] Jielu Zhang, Zhongliang Zhou, Gengchen Mai, Mengxuan Hu, Zihan Guan, Sheng Li, and Lan Mu. Text2Seg: Zero-shot Remote Sensing Image Semantic Segmentation via Text-Guided Visual Foundation Models. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 63–66, Atlanta, GA, USA, October 2024. ACM. doi:10.1145/3687123.3698287. URL https://dl.acm.org/doi/10.1145/3687123.3698287.

[13] Yunkang Cao, Xiaohao Xu, Yuqi Cheng, Chen Sun, Zongwei Du, Liang Gao, and Weiming Shen. Segment Any Anomaly without Training via Hybrid Prompt Regularization. Technical report, arXiv:2305.10724, 2023. URL https://arxiv.org/abs/2305.10724.

[14] Xiaohao Xu, Yunkang Cao, Huaxin Zhang, Nong Sang, and Xiaonan Huang. Customizing visual-language foundation models for multi-modal anomaly detection and reasoning. Technical report, arXiv:2403.11083, 2025. URL https://arxiv.org/abs/2403.11083.

[15] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012. (Images under license CC BY-SA 4.0, see https://creativecommons.org/licenses/by-sa/4.0/).

[16] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD - A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, Long Beach, CA, USA, June 2019. doi:10.1109/CVPR.2019.00982. (Images under license CC BY-NC-SA 4.0, see https://creativecommons.org/licenses/by-nc-sa/4.0/).

[17] Kilian Batzner, Lars Heckler, and Rebecca König. EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies. In *2024*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 127–137, Waikoloa, HI, USA, January 2024. IEEE. doi:10.1109/WACV57701.2024.00020. URL https://ieeexplore.ieee.org/document/10484326/.

[18] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, volume 12664, pages 475–489. Springer International Publishing, Cham, Switzerland, 2021. doi:10.1007/978-3-030-68799-1_35. Series Title: Lecture Notes in Computer Science.

[19] Fan Liu, Lei Su, Mengying Fan, Jian Yin, Zhenzhi He, and Xiangning Lu. Using scanning acoustic microscopy and LM-BP algorithm for defect inspection of micro solder bumps. *Microelectronics Reliability*, 79:166–174, December 2017. ISSN 00262714. doi:10.1016/j.microrel.2017.10.029. URL https://linkinghub.elsevier.com/retrieve/pii/S002627141730505X.

[20] Hyunung Yu. Scanning acoustic microscopy for material evaluation. *Applied Microscopy*, 50(1):25, December 2020. ISSN 2287-4445. doi:10.1186/s42649-020-00045-4. URL https://appmicro.springeropen.com/articles/10.1186/s42649-020-00045-4.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. Technical report, arXiv:1512.03385, 2015. URL https://arxiv.org/abs/1512.03385.

[22] xmed-lab. GitHub - xmed-lab - CLIPSurgery. Web-page, 2025. URL https://github.com/xmed-lab/CLIP_Surgery.

[23] OpenAI. GitHub - openai - CLIP. Web-page, 2025. URL https://github.com/openai/CLIP.

[24] IDEA-Research. GitHub - idea-research - GroundingDINO. Web-page, 2025. URL https://github.com/IDEA-Research/GroundingDINO.

[25] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava and Jo Scott. Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2):151–155, 2015.

[26] Seyed Raein Hashemi, Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, Sanjay P. Prabhu, Simon K. Warfield, and Ali Gholipour. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: application to multiple sclerosis lesion detection. *IEEE Access*, 7:1721–1735, 2019.

[27] Pratik Khanna. Foundation models for semantic segmentation of industrial defect patterns. M.Sc. thesis, Karlsruher Institut für Technologie (KIT), Karlsruhe, Germany (2025).