Ensemble Methods for Visual Anomaly Detection in Manufacturing Settings

Toller Thesis Titel
Master thesis by Marc Saghir
Date of submission: April 10, 2024

1. Review: Super Supervisor
Darmstadt

Erklärung zur Abschlussarbeit gemäß § 22 Abs. 7 APB TU Darmstadt

Hiermit erkläre ich, Marc Saghir, dass ich die vorliegende Arbeit gemäß § 22 Abs. 7 APB der TU Darmstadt selbstständig, ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt habe. Ich habe mit Ausnahme der zitierten Literatur und anderer in der Arbeit genannter Quellen keine fremden Hilfsmittel benutzt. Die von mir bei der Anfertigung dieser wissenschaftlichen Arbeit wörtlich oder inhaltlich benutzte Literatur und alle anderen Quellen habe ich im Text deutlich gekennzeichnet und gesondert aufgeführt. Dies gilt auch für Quellen oder Hilfsmittel aus dem Internet.

Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mir ist bekannt, dass im Falle eines Plagiats (§ 38 Abs. 2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

Bei einer Thesis des Fachbereichs Architektur entspricht die eingereichte elektronische Fassung dem vorgestellten Modell und den vorgelegten Plänen.

Darmstadt, 10. April 2024

M. Saghir

Abstract

Abstract

Contents

1. Introduction **2**
1.1. Contributions . . . 4
1.2. Table Test Viz . . . 6

2. Background **7**
2.1. Ensembles . . . 7
2.2. Classes of Anomaly detection . . . 10
2.3. Datasets . . . 12
2.4. metrics . . . 14
2.5. Anomaly Detection Methods . . . 16

3. Related Work **17**

4. Method **18**
4.1. Discriminator . . . 18
4.2. Our own Dataset . . . 19
4.3. pipeline . . . 21
4.4. Ensemble network . . . 21
4.5. Different ensemble approaches . . . 22
4.6. Logical Anomaly Detection Using Conventional Approaches . . . 22

5. Experimental Setup **24**

6. Experimental Results **25**
6.1. SOTA Methods Performance on classical LOCO Dataset 25

7. Conclusion and Future work **27**

A. Appendix **30**

Figures and Tables

List of Figures

1.1. I am a caption 6

List of Tables

1.1. Description of metrics 6

Abbreviations, Symbols and Operators

List of Abbreviations

Notation	Description
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q Network
IAD	Image Anomaly Detection
ML	Machine Learning
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
SAC	Soft Actor Critic
TRPO	Trust Region Policy Optimization

List of Symbols

Notation	Description
A	continuous action space
S	continuous state space
$\mathcal{H}(\cdot)$	entropy
$\pi(a\|s_t)$	Policy

1. Introduction

Image Anomaly detection as a form of quality control is a widely popular practice in modern manuacturing processes. This also holds true for industrial settings. Ever since the industrial revolution, the need for manufactured metal parts has skyrocketed to a current level of roughly xyz parts of blabla being produced per year(quelle). With rising innovation and production also comes a high need for quality assurance alongside raised standards and requirements. This strict environment(synonym für rahmenbedingungen?) serves among other things to avoid product failure in situations that could cause fatal consequences. Here the quality control in form of anomaly detection often starts at the individual parts manufactured for a single purpose, which demands a large effort and lots of resources due to factors like the named increasing production rate. In earlier days this meant procedures like manual stochastic quality checks of produced parts, a practice that in its nature cannot give complete certainty and requires lots of human labour and thus time and money. Later with the rise of computers and especially sophisticated computer vision methods this process of quality control was more and more being automated using methods like IAD(abkürzung auch erklären wenn in vokabular?), to create a more efficient and easier quality control process. This alongside the striving for even higher reliability and recent developments in artificial intelligence (hier vllt eine refernz für KI geschichte?) brought forth IAD(synonym) as the popular research field that it is today. IAD in our context is a subcategory of general anomaly detection and aims at distinguishing images of a category that conform to some chosen norm from anomalous images of the same category that dont (sollte ich hier eine formel reinpacken? so a la input image i produce score etc). An example would be creating a classifier that is given the image of a screw and can detect whether or not it conforms to our expectations, which in a manufacturing setting likely means to meet the companies quality standards.

With IAD being a very recent and popular field, there are many different deep learning apporaches that have established themselves over the last couple of years. The best performing ones have generally been unsuperised learning apporaches. This stems from the fact, that in any manufacturing setting, there usually exist far less anomalous parts than regular ones, which creates a significant data imbalance. Moreover it can pose as difficult to actually obtain a large number of data points and great variance, since it is a lot of work to coordinate with adequate manufacturing facilities and also implement the necessary infrastructure to take pictures. This problem is supported by the fact that there are little well established datasets being used for modern IAD research. There are still some credible and widely used datasets, amongst them the MVTecAD [1] dataset acting as some sort of gold standard. The dataset will be discussed in greater detail in the background section. Regarding the kind of anomaly detection models, there is again a great variety of approaches that follow a somewhat different strategy of differentiating between the classes. Still most of them can be categorized with two classes: representation or reconstruction based methods. While representation approaches aim at creating a feature based representation in different forms to then compare the features of new input images, reconstruction based ones try to learn how to recreate the part shown in the image as an anomaly free object, and then comparing the constructed product to the original input. Both workflows are visualized in figure xyz, which showcases the different steps of the respetive methods as described. It is to be said that both approaches offer high quality predictions, yet feature representation methods have more frequently shown in latest research to achieve state of the art results.

The current state of IAD generally consists of very high performing classifiers. Here it is important to differentiate between different applications of those classifiers. There is anomaly detection in form of image classification, which was already mentioned. Furthermore there is anomaly locaization. This describes the process of image segmentation to point out the specific regions in which the deteted anomaly occurs. Lastly besides the applications, one can also categorize kinds of anomalies. The most researched anomaly types are so called structural anomalies, which can be described somwhat as superficial damages of the parts material or shape, i.e. a strongly bent screw or one that is broken in the middle. Yet recently there has been a new dataset from the creates of MVTecAD that covers logical anomalies, namey the MVTecAD LOCO [2] dataset. Logical anomalies denote ones that violate an abstract set of rules. More concretely this can mean instances like a metal part with an irregular number of holes, or a label missing. Whereas state of the art approaches regurlarly produce performance metrics of up to 99.6% on classification of structural anomalies, they strongly differ in anomaly localization performance. Moreso does the performance plummet when approaching to classify and localize logaical anomalies. Additionally models often show inconsitencies between differnt subtypes of structura and logical anomalies, especially during localization. These inconsitencies and perforance gabs demonstrate that IAD as such is not yet solved and still has a need for improved robustness and generalizability. This need also holds true due to logical anomalies making up an important new domain of automated quality control, as more complex parts could be tested for requirements. Moreover the showcasing of performance inconsistencies between structual and logical anomalies indicate logical anomalies of being a differnt problem domain. Achieving better translation between those domains(synonym) could serve as a basis for tackling other problems in this field that may present themselves in the future(ist der satz inhaltlich gut?)

1.1. Contributions

This research provides multiple contributions to the field of image anomaly detection, in an effort to further push the progress of robust anomaly localization in differnt domains.

1. To address the problems mentioned at the end of the last section, we attempt (vllt ohne attempt wenn der ansatz funzt) to build a heterogeneous feature level ensemble network, combining differnt state of the art IAD approaches, with the goal to improve general performance but also robustness in image localization and logical anomaly detection. This ensemble network is then tested on the MVTecAD LOCO dataset to observe its performance regarding both anomaly types.

2. Furtheremore an extensive study on the performance of a wide ranging selection of IDA methods on the MVTecAD LOCO dataset is performed. This serves to highlight the current state of anomaly detection in logical problems, and also investigate the application potential of those approaches in such domains(den satz mag ich nicht). (vllt noch einbauen dass diese experimente ggf noch nie durchgeführt wurden und auch code bereitgestellt wird)
3. Second to last we introduce a new category to the MVTecAD LOCO dataset to further increase the diversity of this dataset and strengthen the focus of this thesis on metal maufactured parts. Many datasets either use synthetic data or images in a very linical setting, therefore this attempt for variance is also a step towards IAD on more realistic datasets.
4. Finally the mentioned network and experiments are also streamlined(checken ob ich das wort richtig benutzt hab) into an easy to use pipeline to be used for future experiments in that area.

The contributions mentioned firstly benefit faster research entry and an accelerated experimentation process, with an intuitive setup, as well as potential industrial applications. Here it is to be mentioned that since the ensemble already is of heterogeneous nature, it is particularly uncomplicated to experiment using various IAD apporaches. Furthermore they give more insight into the capabilities of existing methods in an industrial setting and thus also provide a more various and practical setting than the prior categories in the MVTecAD(referenz) dataset. The same methods are also testet on their limitations regarding logical anomalies which was earlier made out to be a relevant aspect of anomaly detection in current manufacturing quality control settings. Lastly through the use of a robust ensemble approach for heterogeneous classifers, this opens up possibilities for expanding the field of application of SOTA IAD methods to other domains with robust performance and may also produce more usable results in real world IAD settings. The presented network can also be used as a foundation for future experiments in different directions. For example, the pipeline may be efficiently used to start investigations on multiperspective datasets in anomaly detection, a topic that also could further advance current IAD applciations.

1.2. Table Test Viz

This is a table:

Metric/Level	Formula	Remarks/Usage
Precision (P) ↑	$P = TP/(TP + FP)$	True Positive (TP), False Positive (FP)
Recall (R) ↑	$R = TP/(TP + FN)$	False Negative (FN), True Positive Rate (TPR)
True Positive Rate (TPR) ↑	$TPR = TP/(TP + TN)$	True Negative (TN)
False Positive Rate (FPR) ↓	$FPR = FP/(FP + TN)$	True Negative (TN)
Area Under the Receiver Operating Characteristic curve (AU-ROC) ↑	$\int_0^1 (TPR)\, d(FPR)$	Classification
Area Under Precision-Recall (AU-PR) ↑	$\int_0^1 P\, d(R)$	Localization, Segmentation
Per-Region Overlap (PRO) ↑	$PRO = \frac{1}{N} \sum_i \sum_k \frac{P_i \cap C_{i,k}}{C_{i,k}}$	Total ground-truth number (N), Predicted abnormal pixels (P), Defect ground-truth regions (C)
Saturated Per-Region Overlap (sPRO) ↑	$sPRO(P) = \frac{1}{m} \sum_{i=1}^{m} \min(\frac{A_i \cap P}{s_i}, 1)$	Total ground-truth number (m), Predicted abnormal pixels (P), Defect ground-truth regions (A), Corresponding saturation thresholds (s)
F1 Score ↑	$F1 = 2(P \cdot R)/(P + R)$	Classification
Intersection over Union (IoU) ↑	$IoU = (H \cap G)/(H \cup G)$	Prediction (H), Ground truth (G)/ Localization, Segmentation

Table 1.1.: Description of metrics

This is a citation: [3]

This is a figure:

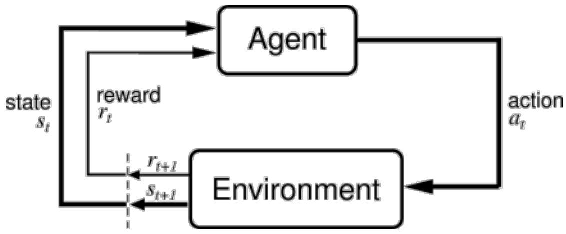

Figure 1.1.: I am a caption

2. Background

This is an algorithm

2.1. Ensembles

When it comes to ensembling classification models, there are multiple approaches to do so. Many ensembling methods are focussed on combining homogeneous models, meaning a set of related models with similar architecture but different parameters or initializations. Typical methods include (liste an methods mit referenzen, majority vote, boosting, bagging, stacking??, CAWPE, blabla), of which i.e. (ansatz für bäume ensembles) are a typical approach to boost simple classifiers like trees. Homogeneous ensembles are popular, since they tend to boost the performance and robustness of a base classifier without lots of additional work, since the ensemble is normally created by initializing the models in different ways. Heterogeneous classifier ensembles on the other hand are not necessarily combinable that easily, since they usually consist of models with different network architectures. This can lead to results, that should be interpreted as the same, differing by large margins(synonym). Yet ensembles of such variety are often desirable since they offer loads of information from different perspectives or domains when done right. Thus to bridge this gap at the output, a common approach is to first calibrate(referenz) and then ensemble each models output (beweis dafür dass das normal ist). For the last combination step, all ensemble techniques suited for homogeneous ensembling can be applied, due to the outputs being in a comparable state then. There are also approaches to collectively calibrate the hyperparameters of each heterogeneous classifier while classifying(referenz). While performance varies, combining these models in such a way is not necessarily regarded as the highest achievable robustness, especially when the classifiers work with features or some other form of inner representation. This stems from the fact that the model outputs are merely a small result of larger inner representations that may focus different aspects of information among the inputs. Therefore in turn, you cannot obtain all relevant information that can be offered by simply calibrating the model outputs. A more robust approach to address that problem, would be to ensemble the aforementioned inner representations, i.e. feature maps and in turn train another classifier for the final meaningful output. Another limitation of both kinds of ensembles, being homogeneous and heterogeneous, is that all models have to actually be trained seperately to then utilize the different classification ouputs. This leads to a highger training time and thus also higher computational cost, which is desirable to be reduced in real world manufacturing firms. The robustness and efficiency has been demonstrated in [4] (ensemble referenz). The authors utilize a feature level ensemble of multiple convolutional neural networks with different architectures and tasks to improve inference speed and accuracy in plant disease detection. (Heller referenz) show that cutting off several, potentially heterogeneous, classifiers after a couple of network layers and ensembling the resulting feature maps yields firstly a significant improvement in training time compared to classical output ensembles. This stems from the fact, that all base classifiers of the ensemble only have to be trained once for every following training approach. During this the model still stays compact(zitat aus ensemble paper markieren), giving it an advantage over most supervised approaches(satz klingt scheiße)(irgendwas mit lightweight hinschreiben? -> nochmal paper gucken). Moreover they compared the performance of different ensemble combinations with conventional output ensembles via softmax and reported in all cases no significant drop in performance. In cases where this approach allowed for different inputs via multispectral cameras(zitat markieren) there even was a similar performance of this ensemble to other state of the art ensembles visible. Keeping in mind the compactness of this new ensemble model combined with an equal performance and possible increased robustness, as argued prior, it is a promising ensemble approach for this work. To obtain ensembled feature maps the paper proposes to bring all feature maps to the same sizes using bilinear interpolation. Since it is not desirable to keep every available feature map, as this would create inputs with way too many features, the amount of feature maps is reduced using principle component analysis. This allows for the ensemble to focus only on the most important features, while maintaining an equal amount of maps as if it were composed of a single classifier. To be more specific (Heller+co) [4] introduced two different approaches to perform this ensemble. The first is a global transformation block as seen in figure xyz(figure mit global transformation block). Here the features are first all resized to the same dimensions and then connected along their channels through a concatenation layer. Afterwards PCA is applied along the channel dimension to obtain a result with N remaining feature maps, where N can be adjusted for ones needs.

- soll ich die transformation blocks erst bei den methods genauer beschreiben?

- hier vorgehensweise und findings von ensemble paper schreiben - erster ansatz für vorgehensweise: The actual combination of features from different level 1 classif

To obtain the different feature representations we would use the corresponding training methods of each IAD apporoach and then cut the model of at the respective time. Figures abc show a schematic view of each approachs respective model architecture, together with an indication of where the representations would be extracted. Proceeding in this way, we would keep all important features of each representation, resulting in a maxmium gain of information and robust predictions over all different classes. Creating such heterogeneous model ensembles on a feature map level was for instance done in (paper ref). Among other results they investigate the performance of heterogeneous models being combined and provide two main approaches to doing so: **General Transformation Block**

- talk about different ensemble approaches we discussed: ensemble model outputs and ensemble model feature maps

feature ensemble: - ground idea: have different algos extract features, and then ensemble them. Afterwards train discriminator on the ensembled features like in simplenet - reference paper that quses PCA and global block transformation - global transformation block: -> resize all feature maps to same dimensions -> append feature maps -> PCA: keep either percentage or set amount

- individual transformation block: -> first apply PCA -> sagen wann das am besten anwendbar ist, auch sagen dass für uns probably der global transformation block reicht -> dann zusammenführen mit resize und appnden

2.2. Classes of Anomaly detection

When trying to understand the choices of IAD approaches for the pipeline and ensemble, one first has to learn about a few important distinctions of models on this topic. The deep learning approaches that have established themselves as state of the art in image anomaly detection are almost exclusively unsupervised approaches. This partiall stems from the fact that naturally anomalous images occurr far less than normal images, hence the word "normal". This is especially true in industrial settings, due to the high performance of production factories nowadays. Therefore if one were to consider using a supervised learning approach to detect anomalies, either a strong class imbalance or an unrepresentative class distribution would occur. While there are some solutions for this, they often are either not goo enough for imbalances this high(synonym klänge cool) or far to extensive. Some papers like (supervised papers zitieren) utilize supervised approaches with some success, but still yield a worse performance than the popular unsupervised approaches generally used. Consequently the biggest model distinction is between unsupervised and supervised ones. Here it has to be said that there are technically also other settings of IAD one could talk about at this level of observation, but since we are also directing our focus to to RGB images, they will not be talked about. Moreover one has to make some simplifications to allow such sharp categorizations of partially interwoven approaches.

The supervised learning category could also further be split up into sub-categories at a lower level. But seeing as the performances of unsupervised approaches dominantely outweigh the performance and cost of the former, this work will solely focus on the latter kind of approaches. In the unsupervised IAD setting we then normally distinguish between reconstruction and representation based models. One of the key differences between those two is(hier dringend auch paper zitieren die das untersuchen),

...

If we now consider the classification of algorithms above, aswell as figure x, we can see that there are quite a lot of unique models and approaches to the same end. To ensure that the built pipeline is able to help experiment on images from different points of view, so to say, aswell as ensure that our ensemble approaches cover as various different aspects as possible, it is crucial to select approaches from majorly different branches. Here it may be noted that the performance of the single models is not completely disregarded, as those models may prove themselves not very useful in the ensemble setting or even as a point of view for experimentation. Therefore certain approaches from the survey papers ...., which yielded performances that were not remotely comparable with the highest performing models, were not considered, even if they might cover a previously unrepresented class of IAD setting. The main choices were: - patchcore + paper - DRAEM + paper - CSFlow + paper

With this choice we still represent reconstruction and representation based settings somewhat comparably, aswell as providing different examples for a variety of subclasses, namely distribution maps, autoencoder, memory banks, teacher-student models, diffusion models and ...

- there are different kinds of approaches to IAD - look at tree picture

- First important distinction is between supervised and unsupervised -> we focus on unsupervised -> list problems with supervised approaches and thus advantages of unsupervised ones

- briefly touch on other IAD settings like few shot, along with references

- among unsupervised approaches, there are two more fundamental distinctions -> reconstruction based vs representation/feature embedding based -> explain difference with lots of references

- for reconstruction based touch on 2-3 base categories like GANs etc and link fundamental papers for GANs etc - for representation based important to explain memory bank, teacher student, and distribution map - explain normalizing flow somehow somewhere in there

- maybe say which algos we chose and what we covered with that

2.3. Datasets

The datasets used in image anomaly detection are scarce, especially when it comes to anomaly detection in a manufacturing setting. There are many datsets and approaches that specialize on certain materials [5] [6] [7] and often only one classs. What currently stands out as a gold standard among IAD datasets is the MVTecAD [1] dataset. The authors created it as a highly representative and standardized set of anomalous images along with training images. It has 15 classes from capsules to screws. Moreover the dataset provides image labels aswell as segmentation ground truths, making it versatile and applicable for multiple algorithms. The masks come as black and white grayscale images, while the image labels are given through its folder structure. Its paradigmatic structure tree can be seen in figure xy. As shown, each class contain train images, which only consist of regular examples, and test images. The data among the testing images is categorized by a title describing the anomaly. The ground truth folder contains according ground truths on a pixel level. Example images of the dataset are to be seen in figure z. They typically are of a rectangular shape and their resolutions range from (pixel min) to (pixel max). More specifications can be found in Bergmann et al. [1] and the whole dataset is publicly available at the official website[8].
The MVTecAD(referenz) dataset is regarded highly among IAD papers, and has since its introduction been used in most relevant papers as a dataaset to benchmark the respective approaches on. This is also likely to remain the trend, since many state of the art algorithms in the recent years have primarily been benmarked on it, forcing new approaches to also be benchmarked on this dataset to be comparable to the current highest performance holding approaches. Despite this work focussing on manufacturing settings MVTecAD

is one of only two datasets relevant to this work, and serves as a comparison for the performance investigation of this papers approaches on the second dataset. This is mainly due to the dataset's importance and its relation to the second dataset.
Later in 2022 Bergman et al. has introduced another IAD dataset that is loosely related to their original MVTecAD dataset, namely the MVTecAD LOCO dataset [2]. This dataset works with the same ground ideas as their original MVTecAD set, but extends the conceptual contents of the dataset by logical anomalies(neu formulieren das klingt scheiße). It consists of five classes: breakfast box, juice bottle, pushpins, screw bag and splicing connectors. The difference to the other dataset is that the anomalous categories for each class are only seperated into good images, images with structural anomalies and images with logical anomalies. As mentioned in the introduction structural anomalies are visible damages to the objects, similar to the MVTecAD dataset. Logical anomalies denote violations against arbitrary restrictions imposed by the authors. To illustrate this by an example: The class of pushpins represents a birds view of a compartmentised box of pushpins(see figure a). A rule added was, that each compartment is only to contain one pushpin. This means that if one region were to miss their contents, or contain more than one pushpin, it would constitute a logical anomaly. If on the other hand a pushpin would have a crooked or broken tip, it would be labelled a structural anomaly. Structurally the differences of the MVTecAD and the MVTecAD LOCO dataset can be seen when comparing figures a and b, which showcases the anomaly classification, aswell as the method of storing segmentations. Here there eixists an image file for each anomalous ground truth area, which are mapped to the image by the folder name they are in. Lastly there exists a validation set in this dataset,
The addition of logical constraints opened an interesting area of research, since the high performance of current state of the art algorithms were only measured on structural anomalies so far. Yet it would be insightful to see if those models could also detect logical anomalies, since those also ocurr in real life settings, such as manufacturing settings. Another concept introduced in [2] is the saturated per-region overlap score, also sPRO. The metric is further analysed in setion (metrics section), but in short gives a measure on how well two regions overlap, while also accounting for regions overlapping in a way, that is seen as sufficient. The criterion of sufficiency is given by a file in the respective class, which

maps a saturation score to each kind of anomaly. Bergmann et al.[2] lastly also released a new IAD model together with the new dataset. The model uses autoencoders(bissi besser beschreiben hier). Since the source code has not been made public, this work refrains from using the method proposed in the paper.

2.4. metrics

Metrics are known to be an important part of developing any artificial intelligence related models. Many of them are used to infer different characteristics of model performance and should be used in different appropriate circumstances, depending on which aspect is important for the current application. Therefore, before the actual developing, one must first choose appropriate metrics to optimize and evaluate on later. IAD as a research area themselves has certain metrics that are the main performance evaluation tool across most papers. A collection of different metrics in this domain are displayed in table 1.1, which is taken from [9]. Visible are well known ones from many other machine learning models like precision, recall, TPR, FPR and the F1-Score. These are generally applicable in most cases, but are not listed in any recent important papers and thus are not important for any analyses in this work. The other metrics are more IAD specific. By a large margin, the most important scoring standard is the AUROC. This metric is usually referenced for image level binary classification and gives an indication on how good the model is able to distinguish between both classes. Its calculation can be seein in table 1.1. Moreover it can be used on a pixel level, which is also a popular approach but not utilized everywhere. Next in importance is the per-region overlap(PRO) score or also the area under the PRO score(AU-PRO). This metric denotes the per-region overlap of two areas on a pixel level and can be calculated using (PRO formel hier hin, im satz rechts dann ggf bezug auf formel zeichen nehmen). The two areas compared are generally an image mask and the according segmentation by the model. The AU-PRO is then calculated by plotting the PRO score at different threholds for the segmentations, and reporting the area under the curve. This can be used to rate the segmentation performance of different models and is also a frequently featured metric in IAD related research. Related to this score is the saturated per-region overlap (sPRO) and also the according are under the curve, the AU-sPRO. This metric was introduced in [2] and briefly mentioned in section (dataset section). The method of deriving the sPRO score is shown again in equation abc, where m denotes the amount of anomalous regions in an image, $A_i|i \in \{1,...,m\}$ an anomalous region among them and $s_i|i \in \{1,...,m\}$ a respective saturation threshold. P is considered to be the pixels classified as anomalous in the target image. The sPRO score is the calculated by averaging the intersections of all predictions and ground truths of an image, while norming the values by the saturation threshold and proividing an upper limit of 1 per region. It is to be said that this gives a similar view on the segmentation performance as the PRO score, as it is a generalized form of it and can producet the same results if the saturation threshold would be equal to the amount of anomalous pixels per region. However, due to its cap of 1, it also rates differently large segmentations equally in cases where the anomalous position posesses some uncertainty. Figure xy demonstrates this behaviour in the case of a logical anomaly of the pushpin class. As visible, the logical anomaly consists of an empty pushpin compartment. The missing pushpin could be placed in any place of this smaller box for it to be valid, therefore an amount of pixels equal to the amount a visual pushpin posesses would suffice. Yet the conventional PRO score would keep on rising as the segmented area gets larger within the anomalous region. Due to the saturation score and limit. this is prevented by the sPRO metric as figure xy shows it to be already saturated once the minimum required amount of pixels is achieved. The saturation scores for each anomaly have to be individually set for each anomaly, and are given for the five classes of MVTecAD LOCO [2].

2.5. Anomaly Detection Methods

2.5.1. PatchCore

2.5.2. SimpleNet

- highlight the use of the discriminator because its important for mine

2.5.3. AST

2.5.4. DRAEM

2.5.5. RevDist

2.5.6. CSFlow

3. Related Work

4. Method

4.1. Discriminator

Our approach to use a small, compact discriminator to differentiate between regular and anomalous image features is inspired by the approach presented in SimpleNet [10]. Since the discriminators inputs in the ensemble pipeline will be of the same nature as the inputs for SimpleNet's discriminator, it is reasonable to utilize their network architeture for this work. Looking back at section (simplenet section) and moreso figure xyz(simplenet architecture), we thus will adapt the SimpleNet pipeline after the feature adapter step. This means the discriminator, shown as the labelled circle will conceptualy be equal to ours. Instead of the merely adpated features, the ensembled features from section (ensemble feature section from methdos) will substitute. The artificial anomalous features, depicted as the red tiled pane in the figure will also be provided during training time. Here we also adapt SimpleNet's approach of gaussian noise for producing those artificial features. (satz ob wir mit simplex noise arbeiten wenn ja dann erwähnen) As also stated in SimpleNet (googlen wie man wörtliche Zitate korrekt benutzt), this discriminator "works as a normality scorer [...] estimating the normality at each location (h, w)". Moreover are positive and negative outputs expected for regular and anomalous features respectively. As to the discriminator network specifics, a regular "two-layer multi-layer perceptron"(zitat markieren) is used. As optimizer a regular adam optiizer by pytorch with a learn rate of (werte erst sauber aufschreiben bevor ich es hier hinschreibe)

- say that this is the binary discriminator for detcting the anomalies from ensembled feature maps - repeat that this is largely based of simplenets discriminator - describe model architecture as described in simplenet paper - list parameters from code like optimizer, learn rate, etc. - also describe loss

4.2. Our own Dataset

As previously mentioned in the introduction, this work will also discuss the introduction of three new dataset classes as an addition to the current ones present in the MVTecAD LOCO dataset. This was to extent the range of objects represented in datasets (referenz auf mvted und loco) and further investigate model performance on industrial manufacturing parts, as this is the mein setting for this work. Shaping the dataset in form of the MVTecAD LOCO dataset has multiple advantages. Firstly we get to make statements about the ability of SOTA algorithms detectig logical anomalies on industrial parts. Moreover we can easily infer our new datasets with all relevant IAD approaches, since they are nearly all published with MVTecAD benchmarkings, meaning they are all released with code to infer on the dataset. As discussed in section (dataset section) the only technical difference between the MVTecAD and LOCO dataset is the storage of the masks, which can be accounted for with a few minor changes in the dataset code representation. Since this work also compares AD performances of approaches between both datasets, the functionality is already implemented in the linked repository as a result. This makes for uncomplicated inference on the new dataset. Lastly these dataset classes may serve as a base for future benchmarking and research of different new IAD approaches. Therefore it is sensible to release the new dataset in the shape of if not the most referenced image anomaly detection dataset(beweis oder umformulieren). The three classes are each representing a metal part, namely a flat connector, an angle and ... . For the first to classes, each part that was acquired for the images is available in a usual hardware store. The third class was a self crafted composite part made of screws and metal sheets, which were also available to buy at similar stores as the other parts. All of the classes meet certain criteria in regards to their material nature, aswell as the possibilities of structural and logical anomalies both occuring with the same part. A solid block of steel for example would make a difficult part to represent logical anomalies. Regarding the recording of images for the dataset, we used (kamera specs) from a birds eye view (nachschauen ob das so heißt) with black cloth (maybe cloth ersetzen und spezifizieren dass es dichtes schwarzes material war) as background. The anomalies were handcrafted in the facilities of the university(suchen wie der werkzeugraum heißt und satz neu formulieren). The labels were done in the same style as the labels of the MVTecAD LOCO classes, meaning black and white segmentation images, with slightly differing pixel values to match according saturation scores.

!Subsection mit flat connector!: For the flat connector we used (maße angeben) regulatory flat connectors (wenn ich lustig bin noch DIN angeben) which are widely available (maybe link referenz). Examplary images of anomalous and good images can be seen in figure x. The structural anomalies consisted of damages to the edge of the part, cut off corners and deep scratches on the surface. Logical anomalies contained missing holes, additional holes and differently sized holes. For simulating missing holes, the holes were stuffed and then the part was spraypainted wholly. Additional holes were simply produced with a drill, likewhise the differently sized holes. The corresponding exemplary masks are also seen in fiure x, as an illustration of how the segmentation of the anomalies was held. If compared with the sample images of figure y(mvtedc loco images) the similarity is visible. The saturation scores for the anomalies, as discussed in section (dataset section loco) were put at (saturation scores) for all above listed anomalies respectively.

- repeat motivation why we added additional data in mvtec style - say that we went with loco mvtec flair(maybe give reasons) - say that we came up with a set of structural and logical anomalies for each category - list categories(flat connector, angle and special construct)

- 3 sub sections for the three categories

- flat connector - link the exact one we used(or examples of some) - give structural anomalies - give logical anomalies - for both briefly touch on how we produced them - show image examples for each

- repeat same for other categories

- also when describing angle: - touch on how there is a special case with multi perspective detection

4.3. pipeline

- explain brief structure of the pipeline - ???

4.4. Ensemble network

There are multiple approaches to ensembling models in general. When combining a heterogeneous set of classifiers, a common approach is to first calibrate(referenz) and then ensemble each models output (beweis dafür dass das normal ist). There are also approaches to collectively calibrate a heterogeneous ensemble of classifiers while classifying. While performance varies, combining the models is generally not regarded as inherently robust, especially when the classifiers work with features or some other form of representation. This stems from the fact that the model outputs do not necessarily reflect their learned representations(neu formulieren) in detail, which in turn means that you cannot obtain the optimal aspects of each part of the ensemble. A more robust approach would be to ensemble the mentioned feature maps or other representations to in turn train a discriminator for the final classification. To obtain the different feature representations we would use the corresponding training methods of each IAD apporoach and then cut the model of at the respective time. Figures abc show a schematic view of each approachs respective model architecture, together with an indication of where the representations would be extracted. Proceeding in this way, we would keep all important features of each representation, resulting in a maxmium gain of information and robust predictions over all different classes. Creating such heterogeneous model ensembles on a feature map level was for instance done in (paper ref). Among other results they investigate the performance of heterogeneous models being combined and provide two main approaches to doing so: **General Transformation Block**

- talk about different ensemble approaches we discussed: ensemble model outputs and ensemble model feature maps

feature ensemble: - ground idea: have different algos extract features, and then ensemble them. Afterwards train discriminator on the ensembled features like in simplenet - reference paper that quses PCA and global block transformation - global transformation block: -> resize all feature maps to same dimensions -> append feature maps -> PCA: keep either percentage or set amount

- individual transformation block: -> first apply PCA -> sagen wann das am besten anwendbar ist, auch sagen dass für uns probably der global transformation block reicht -> dann zusammenführen mit resize und appnden

4.5. Different ensemble approaches

- weighted, random forest etc - specifics

4.6. Logical Anomaly Detection Using Conventional Approaches

As discussed in the related works section, logical anomalies represent a signifacant part of image anomaly detection in modern manufacturing settings. The experiments also serve as an extensive comparison of SOTA methods for IAD versus recent approches that where introduced with special mind to logical anomalies, like GCAD [2](GCAD reference von Paul Bergmann). Moreover, for a qualitative evaluation of the performance change when using feature level ensembles, one first needs to evaluate the base performance of each relevant classifier of the set. Hence this work features experiments to evaluate IAD approaches mainly evaluated on the classical MVTecAD dataset. To do so, the original code from each paper was taken and not modified in regards to any reportet parameters and/or arguments. This was to prevent possible unwanted deviations in original performance by changing up synergies. This paper recognizes the possibility of improved performances on the logical anomalies dataset with different combinations of model parameters. Yet this work focusses on the performance(synonym) of current unmodified apporoaches and more importantly the increased robustness through the use of ensembles. Therefore research regarding this hypothetical improvement would have to be done in another work. Metrics that are specifically looked at in this context are the AUROC, pixel AUROC(weitere maybe einfpgen) and the sPRO. If the functionality to evaluate these metrics was already given, the results of inference were (übernommen), else the according functionality was implemented in this work and used to produce the according metrics. Papers whose approaches were evaluated using the MVTecAD LOCO dataset were: SimpleNet [10], PatchCore [3] (list of paper references with names). These papers were discussed in depth in the backgrounds section and any specifics like hyperparameters can be viewed in the corresponding paper. Furthermore all named classifiers were including, among other variable measures, a preprocessing step to resize the input image. This makes for a variable model input and also the ability to process rectangular images, which is important due to MVTecAD LOCO images being rectangular unlike the squared input from the standard MVTecAD dataset. The only necessary modification to the whole process of anomaly detection was the generation of image masks. The MVTecAD LOCO dataset stores its masks in multiple seperate black and white images, one for each individual anomaly. To fix errors stemming from this fact, additional code was added that pastes all masks belonging to one image into a single mask before iterating through the data.

Überschrift reformulieren!

What i wanna say in this section: - what we did to do the survey on LOCO IAD detection - what we did to the methods(nothing) - aspekte anhand welcher wir die experimente analyiert haben

- wenn ich es actually auch mache dann ablation experiments nennen in welchen ich die images square

5. Experimental Setup

- get information from cluster what it is running on etc. - look in other papers for how they did it -> ensemble paper did this section but also look in IAD papers

6. Experimental Results

- analysis on how methods worked on own dataset individually -> if poor performance error analysis and also address different subclasses

- analysis of how ensemble model worked and if it improved performance

6.1. SOTA Methods Performance on classical LOCO Dataset

In this section we review the performance of prior introduced anomaly detection methods. All experiments were performed with the same experimental setup as explained in section (referenz of experimental setup section), the conditions explained in section (referenz von methods section über loco) and on the mvtec LOCO dataset [2]. The results of inference on the test set can be seen in table x (tabelle mit ergebnissen). As it can be seen, all models scored a significantly lower result on the MVTecAD LOCO dataset than on the normal MVTecAD one(exemplary scores seen in table xy(table mit normalen mvtec scores)). A lower performance is generally to be expected, since firstly logial anomalies are regarded as a more difficult problem than structual ones and secondly the average SOTA performances as seen in table x(tabelle mit ergebnissen) is already closing in on an AUROC of of 1. (den satz rechts von hier müsste man maybe rausmachen oder umschreiben)Therefore there is not much room for further improvement in similar settings, and a worse performance still aknowledgeable as very good. Yet there is an drop in cross-model average AUROC of approcimately (durchschnitts drop ausrechnen), which is a remarkable(synonym) difference. Most other metrics, namely (metrics names), also declined with an respective average of (respective averages). As explained in section (referenz zu metrics section von background), the sPRO (or rather AU-sPRO) was a score introduced in [2] to gain an advanced insight on the quality of segmentations. This means that all approaches who either were published before or did not include this paper in their research likely did not include this metric, which holds true for the approahces used for this experiment. Therefore no comparison in sPRO/AU-sPRO can be shown(vllt einfach spro auch für allte ansätze implementieren?? dann kann ich den satz ändern). Comparing the sPRO scores of the SOTA methods in this experiment with the ones from compared to GCAD [2] shows asignificantly (abchecken ob wirklich) worse performance. Among the different models, the highest scoring one was PatchCore [3]. It scored an average (metrics einfügen) feature embedding based approaches like achieved the highest scoring

Interpretation of results hier, weiß nicht in welche section das eigentlich muss:

7. Conclusion and Future work

Bibliography

[1] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection," *International Journal of Computer Vision*, vol. 129, p. 1038–1059, Jan. 2021.

[2] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *International Journal of Computer Vision*, vol. 130, p. 947–969, Feb. 2022.

[3] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 14318–14328, Jun 2022.

[4] G. Heller, E. Perrin, V. Vrabie, C. Dusart, M.-L. Panon, M. Loyaux, and S. Le Roux, "Multisource neural network feature map fusion: An efficient strategy to detect plant diseases," *Intelligent Systems with Applications*, vol. 19, p. 200264, Sept. 2023.

[5] C. S. Tsang, H. Y. Ngan, and G. K. Pang, "Fabric inspection based on the elo rating method," *Pattern Recognition*, vol. 51, p. 378–394, Mar. 2016.

[6] D. Yang, Y. Cui, Z. Yu, and H. Yuan, "Deep learning based steel pipe weld defect detection," *Applied Artificial Intelligence*, vol. 35, p. 1237–1249, Sept. 2021.

[7] Y. Huang, C. Qiu, Y. Guo, X. Wang, and K. Yuan, "Surface defect saliency of magnetic tile," in *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, IEEE, Aug. 2018.

[8] "Download page of the mvtec ad dataset."

[9] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, vol. 21, no. 1, pp. 104–135, 2024.

[10] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023.

A. Appendix

Appendix here