Ensemble Methods for Visual Anomaly Detection in Manufacturing Settings

Toller Thesis Titel
Master thesis by Marc Saghir
Date of submission: April 7, 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, 7. April 2024

M. Saghir

Abstract

Abstract

Contents

Figures and Tables

List of Figures

List of Tables

Abbreviations, Symbols and Operators

List of Abbreviations

Notation	Description
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q Network
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

This is a citation: [1]

This is a figure:

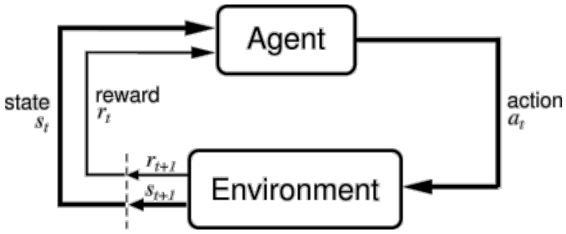

Figure 1.1.: I am a caption

- It is important to note somewhere in the paper that we are dealing with very high variance in our ensemble since we only have 5 models ish

Concept for introduction: - General approach -> Start very wide and narrow it down

- we are in manuacturing setting -> quality control of metal parts -> write one sentence about industrial revolution and address the rising needs for metal parts lately(find credible numbers of metal parts production) - say that most things we engineer nowadays has strict requirements and needs to be functioning properly to avoid dangerous situations/fatalities etc - one of many problems of this requirement is that parts may be insufficiently produced which can lead to unecpected breakage/ausfällen

- to combat this, people have startet inspecting produced parts with different approaches. Yet all approaches required extensive human labor - in recent decades with the rise of computers alongside computer vision advancements, this quality controll process has often been automated whereever possible - Mention that the striving for increasingly accurate controls brought forth the field of anomaly detection which has been a very popular and ectensively researched topic in the last couple of years - The subfield of specifically image anomaly detection aims at successfullly distinguishing between images of good /normal and anomalous parts. - nochmal bezug auf manufacturing

- erklären dass es verschiedene arten von IAD gibt -> bspw. image detection and anomaly localization - auch logical vs structural anomalies anreißen

- sagen dass IAD best performance eig unsupervised sind weil Gründe(= große data imbalance, wenig data overall, etc...(maybe noch einer aber 3 gründe reichen)) - irgendwo erwähnen dass IAD aktuell sehr gute ergebnisse erzielt

- Eventuell kurz auch arten von IAD (embedding vs reconstruction) anschneiden? aber nicht zu detailiert weil der rest kommt in background

- gegen ende aber probleme aufzählen -> gute ergebnisse sind meist in sehr klinischen settings nur beobachtbar -> localization bislang schwieriger als image detection -> ansätze welche gut sind performen doch deutlich schlehter bei logical anomalies, ergebnisse oft inkonsistent -> logical anomalies sind aber wichtig weil sie neue felder der quality control ermöglichen

- Damit einleiten in die contributions

1.1. Begin Intro

In recent years, image anomaly detection has become significantly more important among many scientific communities, especially in industrial applications. This is no surprise, considering the amount of mechanically manufactured parts in factories all over the world.

Since in most parts of the world, manufactured items undergo rather strict regulations and are expected to work in real case scenarios, there is a need for sufficient quality control, that is rising with the amount of produced components. A long time ago, it has come to a point where human based quality checks are not adequate anymore for the production volume, which has led to computer solutions for the problem. Generally speaking, anomaly detection has first been proposed in 1986 for intrusion detection systems(referenz). While the methods and modalities may change, the high level idea stays the same: Detecting data that deviates from a set standard to a degree that is becoming problematatic regarding the own requirements(letzer part maybe neu formulieren). Besides many approaches that were used over the years, deep learning approaches for image anomaly detection have become very popular lately. A likely reason for this are impressively high performance scores with state of the art models achieving an area under the receiver operateor curve of around 0.96 and sometimes even more. It is difficult to say what the first deep learning approaches to this topic were(fact checking), but a notable milestone is definetly Bergmann 2021(Referenz). Among blabla, they introduced the MVTecAD dataset which is used widely and serves as a dataset to benchmark on for nearly every IAD paper released afterwards. - Übergang benötigt

1.2. Table Test Viz

This is a table:

1.3. 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.

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

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. This ensemble network is tested on its performance regarding structural but also logical anomalies from differnt parts. 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. Lastly 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. 4. 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. 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 ground work(fundament oder synonym oder so) 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.

– in my work i contribute the following things: - pipeline to infer new images on different algorithms and compare them -> pipeline is industry focussed for benefits of the guys where i write my thesis

- research on multi perspective detection

- research of ensemble output learning to enhance individual network performance -> simple network over 5-6 outputs

- introduction of very new dataset categories in style of mvtec LOCO dataset

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 [2] (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) [2] 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. The Datasets

The datasets used in image anomaly detection are scarce, especially when it comes to anomaly detection in a manufacturing setting. There are a few that specialize on certain textures(references) and some that can be used for wide ranging categories. What currently stands out as a gold standard among IAD datasets is the MVTecAD(referenz) dataset. It was designed by Bergman et al.(referenz) as a highly representative and standardized set of anomalous images along with training images. It has 15 classes from (some examples) to (...). It 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 iamge labels are given through its folder structure. Its paradigmatic structure tree can be seen in figure xy.(hier ein satz der die ordner struktur beschreibt) Example images of the dataset are to be seen in figure z. They typically are of a rectangular shape and their resolutions range from blabla to blabla. More specifications can be found in (mvtec reference) and the whole dataset is publicly available at (dataset link).
The MVTecAD(referenz) dataset is regarded as the go to dataset(wissenschaftlich formulieren) among IAD papers, and has since its introduction been used in nearly every paper as a dataaset to benchmark ones approaches on. This is also likely to remain the trend, since many important 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 SOTA approaches. Due to its importance MVTecAD is one of only two datasets relevant to this work, and serves as a comparison to investigate SOTA algoithm performances of the second dataset, which will be our main focus.

Later in (loco year) Bergman et al(referenz) has introduced another IAD dataset that is loosely related to their original MVTecAD dataset, namely the MVTecAD LOCO dataset(reference). 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 class:(class names). 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. Structural anomalies being visible damages to the objects, similar to the MVTecAD dataset. Logical anomalies denote violations against arbitrary restrictions imposed by Bergmann et al.(referenz). 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 two pushpins, it would constitute a logical anomaly. If on the other hand a pushpin would have a crooked or broken tip, it would be a structural anomaly. The addition of logical constraints opened an interesting area of research, since the high performance of current SOTA 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. (Noch ansprechen dass LOCO eine neue metric -> sPRO ermöglicht und die saturation configs ansprechen) Bergmann et al.(referenz) also released a new IAD model together with the new dataset. The model uses autoencoders(bissi besser beschreiben hier). Unfortunately the code has not been made public. Aside from approaches tailored specifically towards the detection of logical anomalies, it would be interesting to see how SOTA methods of structural anomaly detection perform on the LOCO dataset. The performance of previous methods on the LOCO dataset is already partially evaluated in some papers like(referenzen von benchmark papers), but will comprehensively be investigated later in this work. Moreover the novel dataset categories introduced later are composed of structural aswell as logical anomalies and formatted in the MVTecAD LOCO dataset style. Aside from the conceptual differences in the two datasets, there are slight changes to the structure tree aswell. The anomaly classes are only changed by name, since it is irrellevant for the models whether the anomaly name is "" anmerkungen für text oben: - beschreiben was mvtec neu bringt: zb dass es näher an real world ist - saturation thesholds ansprechen

2.4. metrics

Metrics are (bekannterweise (auf englisch)) an important part of developing any artificial intelligence related models. Many of them (sagen aus) different aspects of model performance and should be used in different appropriate circumstances, depending on which aspect of the results 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 community(synonym!) also has certain metrics that are the main performance evaluation tool across papers. A (zusammentragung (auf englisch)) of different metrics in this domain are displayed in table 1.1, which is taken from (paper aus dem ich das table hab). 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. The undisputed(synonym)/most popular scoring method is the AUROC. This metric is normally(synonym) used for image level binary classification and gives an indication on how good the model an distinguish between both classes. Its calculation can be seein in table 1.1 and it can also be used on a pixel level, which is done in some papers like (paper refernzen) but not everywhere. Next in importance is the 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 levels for the segmentations, and reporting the area under this curve. This can be used to rate the segmentation performance of different models and is also a metric featured frequently 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 (bergmanns paper) [3]

- show metrics from survey papers - some metrics are well known from other ML applications - metrics that are important for our work/in most recent published papers are: -> auroc: image/instance and pixel level -> Area under PRO -> explain formula

-> sPRO for LOCO and also Area under sPRO –> Extra section where i explain sPRO on basis of dents and scratches paper –> very detailed with saturation threshold and also include figure of comparison for PRO

2.5. description of patchcore algo

2.6. description of simplenet

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

2.7. description of AST

2.8. description of DRAEM

2.9. description of another reconstruction based algo

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 [4]. 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 [3] (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 [4], PatchCore [1] (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 [3]. 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 [3] 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 [3] shows asignificantly (abchecken ob wirklich) worse performance. Among the different models, the highest scoring one was PatchCore [1]. 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] 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.

[2] 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.

[3] 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.

[4] 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