Utilizing Active Machine Learning for Quality Assurance: A Case Study of Virtual Car Renderings in the Automotive Industry

Patrick Hemmer
Karlsruhe Institute of Technology
patrick.hemmer@kit.edu

Niklas Kühl
Karlsruhe Institute of Technology
niklas.kuehl@kit.edu

Jakob Schöffer
Karlsruhe Institute of Technology
jakob.schoeffer@kit.edu

Abstract

Computer-generated imagery of car models has become an indispensable part of car manufacturers' advertisement concepts. They are for instance used in car configurators to offer customers the possibility to configure their car online according to their personal preferences. However, human-led quality assurance faces the challenge to keep up with high-volume visual inspections due to the car models' increasing complexity. Even though the application of machine learning to many visual inspection tasks has demonstrated great success, its need for large labeled data sets remains a central barrier to using such systems in practice. In this paper, we propose an active machine learning-based quality assurance system that requires significantly fewer labeled instances to identify defective virtual car renderings without compromising performance. By employing our system at a German automotive manufacturer, start-up difficulties can be overcome, the inspection process efficiency can be increased, and thus economic advantages can be realized.

1. Introduction

Computer-generated imagery (CGI) has become a central element of leading car manufacturers' advertisement and sales concepts. Virtual car renderings are utilized for various marketing purposes as they enable advertisement campaigns that are efficiently customizable to different markets while not requiring costly physical car prototypes for elaborate photo shoots. Moreover, CGI introduces to customers the possibility to configure a selected car model in real-time online configurators according to their personal preferences.

Similar to the physical car assembly, high-quality standards are a central requirement for the long-term success of all advertisement offerings. Thus, it is crucial to avoid erroneous content creation. However, the continuously growing number of design options of all existing car variants including their derivatives poses increasing challenges for currently human-led quality assurance (QA), turning high-volume visual inspection processes into procedural bottlenecks. Moreover, multiple studies have indicated that the accuracy of human visual inspections may decline with endlessly repetitive routine jobs [1, 2, 3].

In this context, the continuously evolving capabilities of machine learning (ML) algorithms have led to the successful entry of this technology in many industrial sectors over the last years. Particularly, in the area of QA, artificial perception systems offer the possibility to support humans through automatic visual inspection [4]. Applications range from industrial tool wear analysis [5] over printing quality control [6] to quality estimation of porcelain [7]. However, one of the major drawbacks of supervised ML is its dependency on a vast amount of high-quality labeled training data, constituting a major impediment for the launch of such systems in practice [8]. Even though the amount of available data has been constantly increasing [9], labeling of data instances is costly because it often requires the knowledge of domain experts.

To overcome these difficulties for the deployment and operation of ML systems in practice, methods to decrease the data labeling effort, particularly when labeling resources are limited, can be applied, e.g., transfer learning [10], multi-task learning [11], few-shot learning [12], meta-learning [13] or active (machine) learning (AL) [14]. The key idea of AL is that a model can accomplish a certain performance level requiring fewer training instances if the most informative instances with regard to its learning process are selected. Initially, a model is trained on a small labeled data set. In each AL round, new data instances are chosen by an acquisition function and labeled by an expert. Subsequently, they are added to the labeled data pool. This procedure is repeated until a specific termination criterion is met [14].

In this work, we propose an *active machine*

HICSS

learning-based quality assurance (ALQA) system to assist the manual QA process for the CGI content production of virtual car renderings at Audi Business Innovation GmbH, a subsidiary of AUDI AG. Our ALQA system requires significantly less training data and therefore contributes to overcoming the initial start-up barrier of labeling a large number of data instances. Specifically, our system is able to achieve an average performance (measured in terms of the F_2 metric) of 95.81%, while reducing the labeling overhead by 35% on average. Identified defective configurations are forwarded by the system to the responsible QA managers for further root cause analysis.

With our research, we contribute to the body of knowledge with the following five core aspects: First, we propose a novel artifact—an ALQA system—to automatically identify defects in CGI of virtual car models as a support tool for human-led QA. Second, we evaluate its performance and demonstrate its ability to significantly reduce the manual labeling effort of image data required to train the system while maintaining its prediction accuracy. Third, we discuss possible economic implications of the system and show its ability to overcome start-up difficulties while increasing the QA efficiency. Fourth, by illustrating the feasibility of the artifact, we are the first to introduce deep learning into the QA of CGI within an industrial context—as similar approaches have (so far) only been utilized in the area of video game streaming [15]. Finally, we provide generalizable prescriptive design knowledge by providing precise design principles to address the design requirements of ALQA systems. These "blueprints" can serve as a foundation for the design of future artifacts within different domains.

We utilize design science research (DSR) as an overall research design and base our approach on Hevner and Chatterjee [16]. They recommend that a DSR project should consist of at least three cycles of investigation: a relevance cycle (addressing the practical problem, see Section 2), a rigor cycle (outlining the existing knowledge base, see Section 4), and one or multiple design cycles (instantiating and evaluating the research artifact, see Sections 3 and 5). We finish our work with a discussion of the potential economic impact of our artifact (Section 6) as well as a conclusion (Section 7).

2.   Relevance Cycle: Defining the Problem

Advances in CGI over the last years have paved the way for this technology's applicability to a wide range of use cases in practice. Many areas encompassing visual experiences, such as 3D animation, visual movie effects, or augmented and virtual reality, have experienced great success due to the increasingly realistic representation possibilities [17]. Today, car manufacturers leverage CGI of their automotive fleet especially in the context of advertisement purposes. Digital 3D models of all vehicle types including derivatives and vehicle configurations that customers can purchase on the market are created and regularly updated. Not only does this technology contribute to efficiently scalable advertisement options, e.g., customization to different markets worldwide without the need for real-world photo shoots, but it also offers customers the possibility to compile and experience their preferred car configuration before the purchase decision. However, representing all car types including all possible configurations as realistically as possible without any defects while continuously updating them is resource-intense and difficult to maintain by manual means.

In the use case at hand, within the virtual production process of 3D car models at the case company, specific configurations can result in parts of the vehicle not being displayed correctly, e.g., triggered by wiring errors of the individual model components. In these cases, e.g., black areas appear instead. To readjust them, occurred defects must first be identified, which is currently done in manual random checks based on tacit knowledge.

Figure 1 depicts typical errors that can occur in the model. The upper row displays correct model configurations from four different camera perspectives, while the lower row presents respective defects. In the exterior front view, the wheel arch linings are not displayed, in the exterior rear view the taillights are missing. Moreover, in the interior front view, among other parts, the upper part of the gear stick is not present, and in the interior rear view the driver's seat has no belt.

As the manual QA process on-site functions based on randomly chosen configuration and is especially driven by implicit knowledge of which configurations tend to encounter errors, this procedure carries the risk that newly occurring errors remain initially undetected. However, the goal is to identify as many defects as possible before a car model goes live to avoid sub-optimal customer experiences.

3.   Artifact Design

In line with the design science research paradigm, we start by elaborating on our design choices [18]. First, we are confronted with necessary design requirements (DRs) for the artifact.

In the work at hand, we design an ALQA system with the goal to support the identification of defects in

Figure 1. Exemplary Audi Q3 Sportback configurations (from left to right: exterior front view, exterior rear view, interior front view, interior rear view). The upper row displays correct configurations whereas the lower row depicts defective ones. We remove the copyright sign during preprocessing.

Table 1. Overview of workshop participants.

Participant	Focus	Role
Alpha	QA	Senior Manager
Beta	QA	3D Specialist
Gamma	QA/ML	Researcher
Delta	ML	ML Researcher
Epsilon	ML	ML Researcher

car renderings. We conducted an initial focus group workshop [28] with two experts (Alpha, Beta) from our case company as well as three ML researchers (Gamma, Delta, Epsilon). An overview of the roles and foci is depicted in Table 1. In the workshop, we discussed the needs and specifications for the solution and decided on the most important aspects. Based on the results of this workshop as well as related literature, we now derive a set of DRs for the proposed artifact.

We discovered that the case company needs the artifact to require significantly fewer images than conventional ML approaches like supervised ML (DR1). While they have strategically decided to introduce ML techniques to support QA in general, they are hesitant because typically large data sets are required for training [9]. The number of images rendered, the small rate of errors in the images, and the fact that ground truth images cannot automatically be generated with respective labels at the beginning of the project require that the artifact should need as little labeled input as possible. Furthermore, as the task is currently performed manually by humans with high accuracy rates, it is important that performances are comparable. While conventional approaches have shown to reach high performances in similar tasks [29], it remains unclear how other approaches perform. As a consequence, the performance of conventional approaches must be at least reached (DR2) and no or only a small fraction of images should be missed (DR3). Finally, those images that are identified as erroneous should be forwarded to a QA manager for the final decision (DR4).

In order to address the requirements, we select relevant theories from the body of knowledge, which are then incorporated into design principles (DPs). We address DR1–3 with the kernel theory of *active (machine) learning* (AL) [14]. AL is a special case of ML in which the respective algorithm can interactively query an information source (e.g., a human) to label new data points with the desired outputs. As an AL algorithm, we select DEAL [30]. We provide more details on AL, possible implementation options, and reasons for the choice of DEAL in the upcoming Section 4. By utilizing AL, we aim to select those instances within the data set which have the highest contribution to the training process (DP1). This should then result in a significant increase in the artifact's ability to learn faster (DP2). To evaluate the approach, a performance metric that minimizes false negatives is desired as our goal is to avoid defects remaining undetected. Thus, we choose the F_2 metric (see Equation (1)) as it puts a focus on recall without completely ignoring precision (DP3). As DR4 is rather a procedural requirement than a technical one, we inform our design by utilizing the theory of *human-in-the-loop* [27]. Human-in-the-loop is a theory that prescribes human interaction within decision processes that are supported by ML. By providing the deployed artifact with a forwarding mechanism, we allow QA managers direct feedback of errors so they can investigate root causes (DP4).

The resulting DPs, a summary of the DRs, as well as the related theories are shown in Table 2. These requirements are of general nature and are supported by both interviews and literature.

Table 2. Overview of design knowledge.

Design Requirement	Interview Source	Literature Source	Design Principle	Design Kernel Theory
DR1: The artifact should require significantly fewer images than a conventional machine learning approach	Alpha, Beta, Gamma, Delta, Epsilon	[19, 20]	**DP1**: Provide the artifact with active machine learning capabilities to identify instances with the highest contribution to the training period	Active machine learning [14]
DR2: The artifact's performance should be equal to a conventional machine learning approach	Alpha, Beta, Gamma, Delta, Epsilon	[21, 22]	**DP2**: Provide the artifact with active machine learning capabilities to increase the learning curve significantly	
DR3: The artifact should ensure no or only a fraction of erroneous images are missed	Alpha, Beta, Gamma, Delta, Epsilon	[23, 24]	**DP3**: Choose a performance metric with a focus on recall; F_2 metric	
DR4: Identified images that contain errors should be forwarded to QA managers	Alpha, Beta, Gamma	[25, 26]	**DP4**: Provide the artifact with a forwarding mechanism to allow QA managers direct feedback	Human-in-the-loop [27]

4. Rigor Cycle: Related Work and Research Gap

In this section, we ensure rigorous research by elaborating on related work. To that end, we specifically focus on QA in CGI as well as AL.

4.1. Quality Assurance in Computer Generated Imagery

The utilization of deep learning for computer vision tasks has shown remarkable results, especially in the area of image recognition [31]. Naturally, besides other application cases, the technology found its application soon within the area of QA tasks [32]. This field of research is often also called *image quality assessment* [33]. Computer vision in general and deep learning in specific are already utilized for a large variety of QA procedures. Examples include the control of printing quality [6], quality estimation of porcelain [7], or wear characteristics of machining tools [34].

Apart from traditional QA within physical industrial processes, quality control of CGI is an especially interesting area for our work—as we deal with digital car renderings. Interestingly, the topic of QA in CGI is little explored. While some works do generate images in a synthetic way as a basis for training models [35], the focused detection of errors in CGI is rare. The closest work relevant in this regard is the area of video game streaming [15]. As cloud gaming solutions like Google Stadia, Playstation Now or Nvidia Geforce Now [36] allow the streaming of video games via the internet, a niche group of researchers aims to detect quality issues in the generated video streams [37, 38, 39]. However, to the best of our knowledge, no work previously explored QA for CGI within an industrial context.

4.2. Active Machine Learning

Active (machine) learning (AL) has been intensively researched over the past decades. In general, AL methods can be divided into *generative* (e.g., [35, 40]) and *pool-based* (e.g., [30, 41, 42, 43, 44, 45, 46, 47]) approaches.

Generative methods employ generative adversarial networks (GANs) to create informative samples which are added to the training set [35, 40].

Pool-based approaches utilize various acquisition strategies to select the most informative data instances. The literature further distinguishes *diversity-* (e.g., *core-set* [41]) and *uncertainty-based* approaches, as well as combinations of both (e.g., *Batch BALD* [42] or *BADGE* [43]). For our work, uncertainty-based techniques are most relevant. Here, the underlying assumption is that the more uncertain a model is with respect to a prediction, the more informative the associated data point has to be. Uncertainty-based AL approaches can be further divided into *ensemble-based* (e.g., *Deep Ensemble* [46] or *MC-Dropout* [47]) or *non-ensemble-based* (e.g., *Minimal Margin* [44], *Learning Loss* [45], or DEAL [30]) methods. Generally, deriving uncertainty estimates via the softmax function carries the risk that a model may be uncertain in its predictions despite high softmax probabilities [48].

Approaches addressing this drawback, e.g., generate multiple predictions per instance, which are subsequently aggregated, by using the technique of *MC-Dropout* [47] or by employing a model ensemble (*Deep Ensemble* [46]). Another approach is to add

a separate loss prediction model to the network [45]. While the first two approaches are time-consuming regarding the acquisition of new data instances, the latter requires implementing an extra model.

In this work, we employ DEAL [30] as it allows the selection of new data through uncertainty estimates generated by only one forward pass of each instance through the network. These uncertainty estimates are derived by placing a Dirichlet distribution on the class probabilities. Its parameters are set by the output of the neural network [49].

5. Artifact Evaluation

In this section, we first describe the utilized data (Section 5.1), implement the previously described artifact design of the ALQA system (Section 5.2), and evaluate it accordingly with a common AL procedure [14] (Section 5.3). Here, the focus of the evaluation is on the AL component of the ALQA system, i.e., whether a performance comparable to a conventional ML component can be achieved with fewer data instances to-be-labeled and whether its classification performance is sufficient for practical use. The human-in-the-loop component is not evaluated because the only source of error results from the AL component.

5.1. Data Set

During the initial focus group workshop, we chose the 3D model of the Audi Q3 Sportback for the evaluation of our artifact as we were able to draw on a model status that had (at the time of the conducted project) not yet undergone the QA process on site. Thus, we can assess during the evaluation whether our artifact is capable of detecting the defects that have been identified in the manual QA process afterward. Since ML approaches applied directly to 3D data, e.g., 3D point clouds, are still in their infancy [50], we reduce the problem complexity to 2D images [51]. However, the artifact requires analyzing the largest possible area visible to customers during the QA process in order to ensure no visible key areas are missed. Thus, we select the four camera perspectives introduced in Section 2 to fulfill this requirement (see Figure 1).

As our computational resources are limited, we randomly select a subset of 4,000 uniquely identifiable vehicle configurations from a pool of customer orders. As we employ an untested car model, it inherently contains defective configurations. Furthermore, in order to represent a large diversity of theoretically manifestable defects, we randomly specify parts of the vehicle not being displayed correctly by the rendering engine in half of the subset. Subsequently, we render the entire pool of car configurations for each camera perspective. As the model inherently contains defective configurations and the procedure of sampling additional defects might affect parts invisible to the customer, we cannot automatically generate ground truth labels for the rendered data set. Consequently, each image must be subjected to manual labeling. Thus, the entire data set was labeled within the focus group over a period of several weeks. Ambiguous instances were discarded. Finally, for each camera perspective, we assign a fixed number of 2,000 images to the training set and allocate the remaining images to the test set after reserving 10% of the data per camera perspective for the validation set.

5.2. Experimental Setup

To demonstrate the effectiveness and efficiency of our approach, we evaluate it according to the following AL scheme: Initially, a model is trained on a small labeled data set. In each AL round, new data instances are chosen by an acquisition function, labeled by an expert, and added to the pool of labeled data. This procedure is repeated until a predefined termination criterion is met [14].

As a classification model, we choose a convolutional neural network (CNN) using the ResNet-18 architecture [31] with the modifications and acquisition function as described by Hemmer et al. [30]. All images are compressed to a resolution of 128×128 pixels. Initially, for each camera perspective, we provide the model with a small labeled data set consisting of 100 randomly sampled data instances. Hereupon, the model is trained from scratch for 100 epochs while monitoring the validation performance metric development. Early stopping is applied once the performance metric of the validation set does not increase within 20 epochs. We choose a training batch size of 50, a learning rate of $5 \cdot 10^{-4}$, and Adam [52] as optimizer. In each AL round, a new batch with 100 images is selected and added to the training data pool. Afterward, the model is re-trained from scratch again.

We benchmark our approach against a conventional ML component that is trained using the entire available training set and serves as an upper performance bound. Moreover, to compare the learning progress to our approach, the conventional ML component is additionally tested using the aforementioned AL evaluation. However, new batches of images are acquired uniformly at random. We terminate the AL procedure for the conventional ML component once its performance differs only marginally from its performance on the complete training data set.

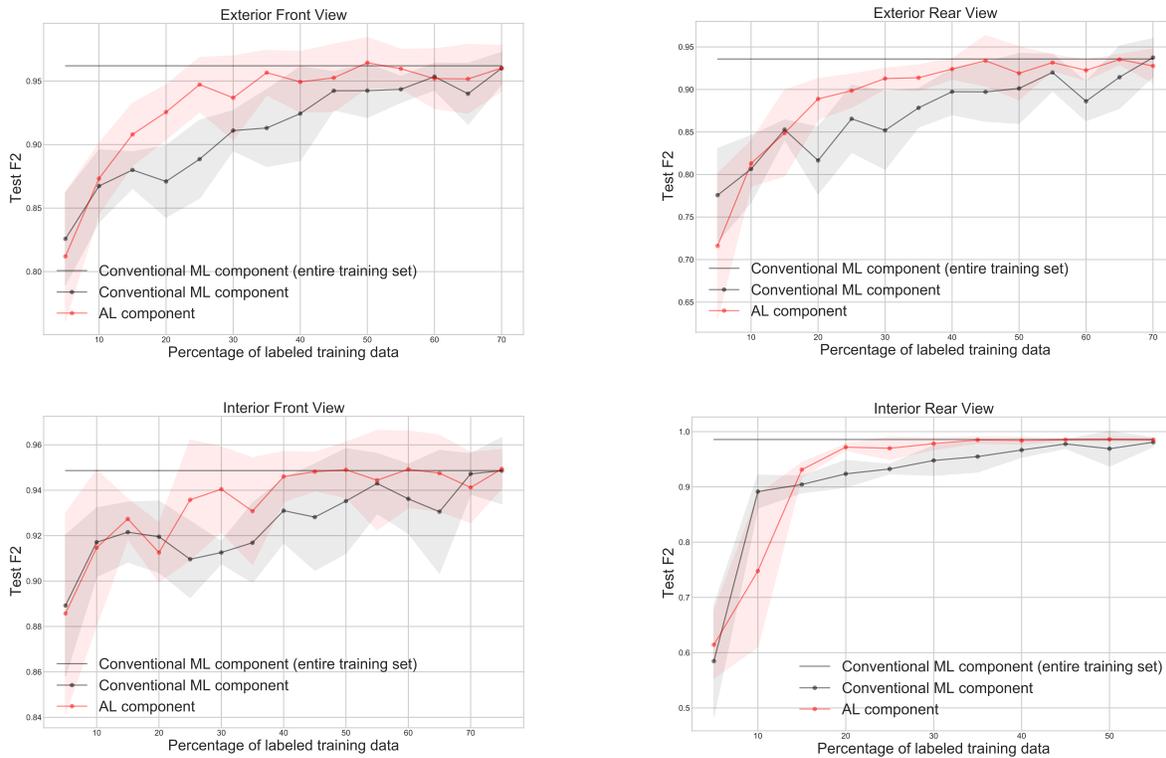For each camera perspective, we apply the resulting

Figure 2. F_2 **metric over the percentage of acquired labeled training data for exterior front view, exterior rear view, interior front view, and interior rear view. We benchmark the performance of our AL component against a conventional ML component that acquires new data uniformly at random. The solid horizontal line represents the conventional ML component trained with all labeled training data. Shaded regions display standard deviations.**

number of AL rounds equally to our AL component. This allows us to determine if the AL component is able to reach the final performance level of the ML component with fewer labeled images.

We repeat each experiment five times and report the F_2 metric (see Equation (1)) including standard deviation. As the artifact's purpose is the identification of defective configurations, minimizing false negatives becomes more important than minimizing false positives. Since the F_2 metric places more focus on recall, without neglecting precision [53], it is suitable for our case.

$$F_2 = 5 \cdot \frac{precision \cdot recall}{4 \cdot precision + recall} \qquad (1)$$

5.3. Performance Evaluation

We conduct the experimental evaluation as described in Section 5.2. The results for all four camera perspectives as a function of the number of training instances are depicted in Figure 2. The horizontal lines represent the performance of a conventional ML component trained on the entire training data set. For the exterior views, F_2 scores of 96.20% (front) and 93.58% (rear) can be achieved. The interior views result in F_2 scores of 94.86% (front) and 98.58% (rear), respectively. While repeatedly adding new data instances after each AL round to the initial training data pool, we see that the AL component identifies new data instances contributing to an overall steeper learning curve in comparison to the conventional ML component. For all camera perspectives, the AL component reaches the upper bound in terms of F_2 score using substantially fewer training data instances. Thus, DR1, DR2, and DR3 are addressed successfully.

The savings can be extracted from Figure 2 via the distance of the touchpoints of the red and black learning curves with the horizontal line. Specifically, a saving of 400 images can be achieved for the exterior front view, 500 for the exterior rear view, 600 for the interior front view, and 400 for the interior rear view.

To substantiate these findings, we apply a paired t-test assessing if the respective performance levels differ significantly between the AL and ML component.

Its assumptions are validated beforehand. The criterion of normally distributed differences of the pairs is validated based on Shapiro and Wilk [54]. We note that the null hypothesis that a sample comes from a normal distribution cannot be rejected for the exterior front view ($W_s = 0.9106, p = 0.1606$), exterior rear view ($W_s = 0.9244, p = 0.2540$), and interior front view ($W_s = 0.9258, p = 0.2361$). Performing the paired t-test results in the null hypothesis to be rejected for the exterior front view ($t = 3.6355, p = 0.0030$), exterior rear view ($t = 2.3617, p = 0.0345$), and interior front view ($t = 3.0639, p = 0.0084$). As the null hypothesis for the Shapiro-Wilk test is to be rejected for the interior rear view ($W_s = 0.6671, p = 0.0002$), following literature, we perform a Wilcoxon signed-rank test [55]. Here, the null hypothesis that two related paired samples come from the same distribution can be rejected ($W_w = 11.0, p = 0.0537$).

Finally, we briefly discuss the observed results. A possible reason explaining both the performances of the conventional ML and the AL component might be the image quality. Unlike most real-world use cases, the rendering engine provides the ability to render any vehicle configuration with the same quality and camera angle. Another aspect we want to highlight is the quality of the labeled instances. Potential errors in the labeling process by human annotators can negatively influence the overall model performance. This might explain the upper performance bounds of the models trained with the entire available training set and the observed decline in the performance of the AL component between individual AL rounds (see Figure 2).

6. Discussion

With the results at hand, we discuss a potential economic impact. We analyze possible time savings resulting from employing the ALQA system in comparison to a conventional ML-based approach and contemplate its embedding in the overall QA process. Even though more and more data is becoming available nowadays for catalyzing the usage of supervised ML algorithms [9], necessary prerequisites, e.g., available labels are often not met in practice [8]. Therefore, we intend to demonstrate with the following calculation-based example that AL-based approaches have the potential to mitigate these issues. However, our calculations are based on multiple assumptions, e.g., similar performances across all models and the same labeling quality by respective annotators for all other car types of the fleet. Thus, they should be interpreted as indicative to illustrate the potential of ALQA and similar approaches. In the case of the regarded model

(Audi Q3 Sportback), it was not possible to collect data instances including ground truth labels. Thus, manual labeling was inevitable and was conducted within the focus group over several weeks (see Section 5.1). For a randomly selected subset of images, we measured the time it took each focus group member to label each instance resulting in an average labeling time of 31 seconds (standard deviation 23 seconds) per image. By incorporating the AL component in the artifact, an average of 475 fewer images had to be labeled per camera perspective compared to a conventional ML component. The total savings of 1,900 images are equivalent to an average reduction of 35%. Taking into account the average labeling time per image and a working day of 8 hours [56], the labeling effort is reduced by approximately two full days per car model. Now, let us imagine extending the ALQA system to the entire fleet of the case company. Considering 18 vehicle types [57] would increase the potential average savings to 36 days, which emphasizes the potential to lower initial barriers for the practical use of such a system.

Finally, we want to stress that the ALQA system's objective is to complement human capabilities in the QA process, not to replace them [58]. As a large number of configurations can be assessed by the artifact efficiently, identified errors are passed to QA managers for further root cause analysis fulfilling DR4. Unblocked capacities taken up by the mere identification of defects can be used more efficiently in tasks not covered by the system, e.g., verifying that dynamic animations work properly.

7. Conclusion

Increasingly realistic representation possibilities driven by technological advances in the field of computer-generated imagery have made this technology applicable to a continuously increasing number of industrial use cases. Automotive manufacturers utilize computer-generated imagery for scalable advertisement campaigns and offer customers the possibility to configure a selected car model in online configurators according to their personal preferences. However, human-led quality assurance based on visual inspection faces the challenge to keep up with these scalable concepts in terms of efficiency and effectiveness, turning high-volume visual inspection processes into procedural bottlenecks. Even though deep learning approaches for computer vision tasks [31], and in particular for quality assurance purposes [32], have demonstrated promising results, their applicability to computer-generated imagery within an industrial context remains largely unexplored. To address this gap, this work proposes an active machine learning-based

quality assurance (ALQA) system for identifying visual defects in computer-generated car imagery. The system is capable of identifying defective configurations with an average F_2 metric of 95.81% for the Audi Q3 Sportback while requiring on average 35% fewer data instances to-be-labeled a priori.

Through the design, implementation, and evaluation of our artifact, we contribute to the body of knowledge in the following ways: The proposed artifact can be utilized as a support tool for human-led quality assurance. Additionally, we demonstrate the artifact's ability to significantly reduce the manual labeling effort while achieving a desired performance level. This has the advantage of lowering initial start-up difficulties for companies aiming to leverage the economic potentials of machine learning. Moreover, we contribute with generalizable prescriptive design knowledge by providing precise design principles to address the design requirements of ALQA systems.

However, the scope of this paper is subject to certain limitations. One shortcoming is that the evaluation of the proposed system is currently limited to one vehicle model—namely the Audi Q3 Sportback. Furthermore, the system is to be understood as a supporting tool for quality assurance to increase its efficiency and effectiveness. It does not claim to automate the entire quality assurance process, as it does not cover the verification of dynamic car model functionalities, such as the opening of doors and trunk.

Apart from these limitations, the options for future research are manifold. An obvious step may be to extend the proposed system to the entire vehicle fleet. Moreover, the system currently classifies configurations based on the features extracted from 2D images. A promising field of research to further improve its performance could be to incorporate further information provided by the rendering engine, e.g., camera distances of pixels, to better recognize whether objects that should be present are in fact present. Lastly, as new configuration options come onto the market at regular intervals, this could result in less accurate classifications over time. Thus, it could be worthwhile to explore the combination of the proposed system with approaches from the field of concept drift.

Acknowledgments

References

[1] Roland T Chin and Charles A Harlow. Automated visual inspection: A survey. *IEEE TPAMI*, 6: 557–573, 1982.

[2] James W Schoonard and John D Gould. Field of view and target uncertainty in visual search and inspection. *Human Factors*, 15(1):33–42, 1973.

[3] Judi E See, Colin G Drury, Ann Speed, Allison Williams, and Negar Khalandi. The role of visual inspection in the 21st century. In *Proceedings of HFES*, volume 61, pages 262–266, 2017.

[4] Je-Kang Park, Bae-Keun Kwon, Jun-Hyub Park, and Dong-Joong Kang. Machine learning-based imaging system for surface defect inspection. *IJPEM*, 3(3):303–310, 2016.

[5] Alexander Treiss, Jannis Walk, and Niklas Kühl. An uncertainty-based human-in-the-loop system for industrial tool wear analysis. In *Proceedings of the ECML-PKDD 2020*, Part V, pages 85–100, 2020.

[6] Javier Villalba-Diez, Daniel Schmidt, Roman Gevers, Joaquín Ordieres-Meré, Martin Buchwitz, and Wanja Wellbrock. Deep learning for industrial computer vision quality control in the printing industry 4.0. *Sensors*, 19(18):3987, 2019.

[7] Daniela Onita, Nicolae Vartan, Manuella Kadar, and Adriana Birlutiu. Quality control in porcelain industry based on computer vision techniques. In *2018 International Young Engineers Forum (YEF-ECE)*, pages 79–84. IEEE, 2018.

[8] Lucas Baier, Fabian Jöhren, and Stefan Seebacher. Challenges in the deployment and operation of machine learning in practice. *ECIS*, 2019.

[9] Xue-Wen Chen and Xiaotong Lin. Big data deep learning: challenges and perspectives. *IEEE access*, 2:514–525, 2014.

[10] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021.

[11] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.

[12] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[13] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and

Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[15] Saman Zadtootaghaj, Steven Schmidt, Saeed Shafiee Sabet, Sebastian Möller, and Carsten Griwodz. Quality estimation models for gaming video streaming services using perceptual video quality dimensions. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 213–224, 2020.

[16] Alan Hevner and Samir Chatterjee. Design science research in information systems. In *Design research in information systems*, pages 9–22. Springer, 2010.

[17] Wen-Jung Huang, Chia-Hung Yeh, Chia-Chen Kuo, Yuan-Chen Cheng, and Jia-Ying Lin. Rating realism assessment for computer generated imagery. In *IEEE ICCE-TW*, pages 327–328. IEEE, 2017.

[18] Hendrik Meth, Benjamin Mueller, and Alexander Maedche. Designing a requirement mining system. *Journal of the Association for Information Systems*, 16(9):799, 2015.

[19] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, volume 2, page 6. Citeseer, 2000.

[20] Akif Khan, Alexander Linden, and Anthony Mullen. *Boost Your Training Data for Better Machine Learning*. Gartner, 2019.

[21] Davi Pereira-Santos, Ricardo Bastos Cavalcante Prudêncio, and André CPLF de Carvalho. Empirical investigation of active learning strategies. *Neurocomputing*, 326:15–27, 2019.

[22] Yoram Baram, Ran El Yaniv, and Kobi Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5(Mar):255–291, 2004.

[23] Gordon V Cormack and Maura R Grossman. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th CIKM*, pages 1039–1048, 2016.

[24] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. A system for efficient high-recall retrieval. In *ACM SIGIR*, pages 1317–1320, 2018.

[25] Christos Emmanouilidis, Petros Pistofidis, Luka Bertoncelj, Vassilis Katsouros, Apostolos Fournaris, Christos Koulamas, and Cristobal Ruiz-Carcel. Enabling the human in the loop: Linked data and knowledge in industrial cyber-physical systems. *Annual Reviews in Control*, 47:249–265, 2019.

[26] Michael Vössing, Felix Potthoff, Niklas Kühl, and Gerhard Satzger. Designing useful transparency to improve process performance—evidence from an automated production line. *ECIS*, 2019.

[27] Fabio Massimo Zanzotto. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.

[28] David L Morgan. *Focus groups as qualitative research*, volume 16. Sage publications, 1996.

[29] Xiaoqing Zheng, Hongcheng Wang, Jie Chen, Yaguang Kong, and Song Zheng. A generic semi-supervised deep learning-based approach for automated surface inspection. *IEEE Access*, 8: 114088–114099, 2020.

[30] Patrick Hemmer, Niklas Kühl, and Jakob Schöffer. Deal: Deep evidential active learning for image classification. In *Proceedings of the 19th IEEE ICMLA*, pages 865–870, 2020.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE CVPR*, pages 770–778, 2016.

[32] Thilo Stadelmann, Mohammadreza Amirian, Ismail Arabaci, Marek Arnold, Gilbert François Duivesteijn, Ismail Elezi, Melanie Geiger, Stefan Lörwald, Benjamin Bruno Meier, Katharina Rombach, et al. Deep learning in the wild. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 17–38. Springer, 2018.

[33] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

[34] Jannis Walk, Niklas Kühl, and Jonathan Schäfer. Towards leveraging end-of-life tools as an asset: Value co-creation based on deep learning in the machining industry. In *Proceedings of the 53rd HICSS*, 2020.

[35] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *MICCAI*, pages 580–588. Springer, 2018.

[36] Andrea Di Domenico, Gianluca Perna, Martino Trevisan, Luca Vassio, and Danilo Giordano. A network analysis on cloud gaming: Stadia, geforce now and psnow. *arXiv preprint arXiv:2012.06774*, 2020.

[37] Markus Utke, Saman Zadtootaghaj, Steven Schmidt, and Sebastian Möller. Towards deep learning methods for quality assessment of computer-generated imagery. *arXiv preprint arXiv:2005.00836*, 2020.

[38] Saman Zadtootaghaj, Nabajeet Barman, Steven Schmidt, Maria G Martini, and Sebastian Möller. Nr-gvqm: A no reference gaming video quality metric. In *IEEE ISM*, pages 131–134. IEEE, 2018.

[39] Nabajeet Barman, Steven Schmidt, Saman Zadtootaghaj, Maria G Martini, and Sebastian Möller. An evaluation of video quality assessment metrics for passive gaming video streaming. In *Proceedings of the 23rd packet video workshop*, pages 7–12, 2018.

[40] Christoph Mayer and Radu Timofte. Adversarial sampling for active learning. In *IEEE WACV*, pages 3071–3079, 2020.

[41] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[42] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, pages 7024–7035, 2019.

[43] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

[44] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE TCSVT*, 27(12): 2591–2600, 2016.

[45] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE CVPR*, pages 93–102, 2019.

[46] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE CVPR*, pages 9368–9377, 2018.

[47] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th ICML*, pages 1183–1192. JMLR. org, 2017.

[48] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1:3, 2016.

[49] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *NIPS*, pages 3179–3189, 2018.

[50] Christoph Sager, Patrick Zschech, and Niklas Kühl. labelcloud: A lightweight domain-independent labeling tool for 3d object detection in point clouds. *Proceedings of CAD'21*, 2021.

[51] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE TPAMI*, 2020.

[52] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[53] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *ECIR*, pages 345–359, 2005.

[54] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[55] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[56] Federal Office of Justice. German working hours act §3, 2020.

[57] AUDI. Model overview, https://www.audi.co.uk/models.html, last accessed 06/15/2021, 2021.

[58] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. Human-ai complementarity in hybrid intelligence systems: A structured literature review. *PACIS 2021 Proceedings*, 2021.