Fachhochschule Aachen
Campus Jülich

Fachbereich 9: Medizintechnik und Technomathematik
Studiengang: Angewandte Mathematik und Informatik

Data Augmentation Tuning for Visual Deep Learning Utilizing Expertise-Guided Bayesian Optimization

Bachelorarbeit

von

Karl Johannes

Erstprüfer:	Prof. Dr. rer. nat. Stephan Bialonski
Zweitprüfer:	M. Sc. Maximilian Motz
Matrikelnummer:	3527999

Aachen, den 21. Juli 2024

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Bachelorarbeit mit dem Thema

> *Data Augmentation Tuning for Visual Deep Learning Utilizing Expertise-Guided Bayesian Optimization*

selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, wobei ich alle wörtlichen und sinngemäßen Zitate als solche gekennzeichnet habe. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Aachen, den 21. Juli 2024

KJohannes

Karl Johannes

Abstract

Deep learning (DL) models require a substantial amount of data in order to achieve a high accuracy and be robust to new, unseen scenarios. This makes it particularly difficult for domains such as visual quality inspection, which often lack sufficient data. Since defects occur infrequently, visual quality inspection suffers from severe data imbalance and data scarcity. This results in less accurate and less robust DL models.

A possible solution to this problem is to augment the data, generating more data points and increasing the variance of the data to make the model more accurate and robust. This can be achieved through automatic data augmentation (AutoDA). Bayesian optimization (BO) can be used as an AutoDA framework to identify the best data augmentation policies. However, since this takes a lot of time and resources, this work proposes to integrate expert knowledge into the Bayesian optimization of data augmentation policies, to accelerate the optimization. This involves identifying the state of the art in how expert knowledge is currently utilized within BO. Then derive an own methodology to extract and integrate visual quality inspection expert knowledge into BO. Eventually, testing the proposed methodology and analyzing its limitations.

Contents

1 Introduction

Recent breakthroughs in Deep Learning (DL) have made it possible to address various tasks to be solved, which previously were considered intractable. The visual quality inspection domain is just one example of where DL enabled to have much more accurate quality inspections and enabled new inspection types to be possible [1–5]. However, the drawbacks as outline by [2] is that these DL models require huge amounts of data to achieve high accuracy scores and become robust enough. The problem is that data is not infinite and occasionally scarce. Therefore, this work will build upon current state-of-the-art approaches to increase the volume and quality of data within this domain more efficiently. Specifically, the approach of Bayesian optimization (BO) (2.3) is utilized to identify optimal data augmentation policies, which is referred to as AutoDA (2.2). Such image data augmentation policies serve as instructions on increasing the amount and quality of data. Identifying the best augmentation policy is challenging due to the high dimensionality and the expensive nature of evaluations. BO is an appropriate choice, as it is intended to estimate functions, such as the described policy goodness function within a limited number of iterations. Functions suitable for BO may be black-box functions, may be costly to evaluate with some degree of noise, or may lack an easily estimatable gradient [6, p.1]. Even with the help of BO finding these augmentation polices can take a long time and be computationally very expensive. Hence, accelerating this search process is a very important topic. In fact, a survey by Bouthillier and Varoquaux at NeurIPS2019 and ICLR2020 found that the vast majority of interviewed researchers still prefer to tune their hyperparameters manually or only with traditional random and grid search [7]. This proofs that hyperparameter optimization methods including BO still have drawbacks to improve on. One reason for these results could be that many believe that their intuition and experience with tuning hyperparameters has more value than statistically evaluating the points with the most potential for information gain. Not making use of the human insights can lead to having multiple costly evaluations wasted on obviously bad parameter combinations. In case the researchers would be able to integrate all or at least a sufficient amount of their knowledge into BO, more researchers and practitioners might trust doing hyperparameter optimization with more advanced techniques.

Visual Quality Inspection is a domain where data samples are often very limited, posing a significant challenge for DL models. This is a domain in which images with defects are even more scarce. In case, specific defect types should be classified it becomes even more difficult. Consequently, having an instruction to enrich the dataset synthetically has enormous potential on DL models.

This work will focus on the visual quality inspection domain. Therefore, it is mainly about how the experience and intuition of visual quality inspection process experts can be integrated into the Bayesian optimization process. With the aim of accelerating the identification of optimal image data augmentation policies to augment image data sets from the visual quality inspection domain. However, there is not a lot of research about integrating specifically expert knowledge into BO, instead other sources of integrating knowledge is much more common. Therefore, the theoretical investigation of the current state of the art will focus on integrating any kind of knowledge into BO. Subsequently, the found research gaps will be elaborated (3.2), followed by defining this works research questions (3.3).

2 Fundamentals

This chapter introduces the core methods used within this work. The aim of this work is to integrate expert knowledge into the Bayesian Optimization process (2.3) to search for optimal image data augmentation policies (2.2.1). The research field of automatically searching for optimal data augmentation policies is often referred to as Automatic Data Augmentation (2.2).

2.1 Data Augmentation

The process of Data Augmentation enriches a dataset by creating new data points by altering existing data to show the machine learning model more perspectives of the data. Within our use case of image classification, this could involve augmentations like rotating, cropping or recoloring the image. For humans some of these augmentations might not seem too informative. Nonetheless, since these machine learning models learn through features like structures and patterns, the named slight alterations can give the model completely new perspectives.

2.2 AutoDA

Automatic Data Augmentation (AutoDA) tries to find the optimal data augmentations to be used to maximize the quantity and quality of the data. These augmentation instructions are called augmentation policies (2.2.1). The most common Augmentation Policy structure was introduced by Cubuk et al. [8]. Each augmentation policy consists of five sub-policies, which include two subsequently applied data augmentations (e.g. rotation, recoloring, etc.). Both augmentations have their own two parameters for magnitude and probability of application. Augmentations that have no use of the magnitude parameter will also not be used.

2.2.1 Data Augmentation Policy

An example image data augmentation policy from the AutoAugment paper [8] can be seen in Figure 2.1. On the very left column the original image is displayed. The other columns show the application of one policy, consisting of five sub-policies. Consequently each image from the second column was first equalized with a probability of 40% with magnitude of 4. Subsequently, each image was rotated with a probability of 0.8 and a magnitude of 8. As illustrated, due to the probabilities, each application of the same sub-policy can result in a different augmented image.

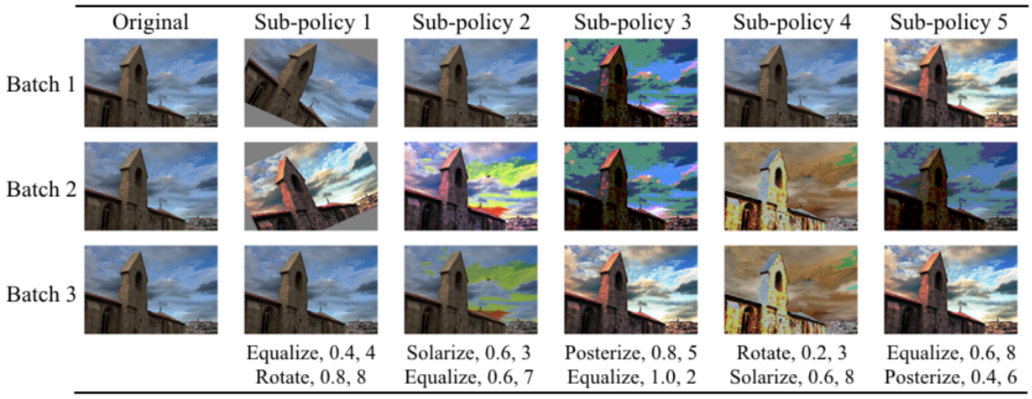


Figure 2.1: Example policy from the AutoAugment paper [8] trained on ImageNet.

2.2.2 FastAA

Fast AutoAugment (FastAA) as proposed by Lim et al. [9] and highly inspired by AutoAugment from Cubuk et al. [8]. The main contribution of FastAA is the introduction of Density Matching as their evaluation function. This concept in AutoDA is also known as evaluating a proxy task instead of the target task, as the proxy task has similar results but is much cheaper to evaluate. In particular, the original target task of augmenting data with a policy and then training an entire neural network on that augmented data to figure out an accuracy score or loss, is not needed anymore. Instead Density Matching directly compares the augmented data to the original validation data. The assumption being that the closer their distributions are, the more realistic the augmented images should be. This approach achieves similar results with the original AutoAugment paper [8] that used the target task for evaluation, while being orders of magnitude faster.

2.3 Bayesian Optimization

Bayesian optimization is a standard methodology for estimating and optimizing expensive objective functions. It is especially prominent for optimizing the parameters of a black-box function. These functions have no useful expression or non-practical gradient estimation mechanisms [6, p.1]. The general BO framework can be divided into four key components. These being the surrogate (2.3.1) model that estimates the objective function and serves as en estimated replacement, as the objective function is unknown. Additionally, the acquisition function (2.3.2) which calculates the likelihood of x-values yielding high improvements or knowledge. The evaluation function (2.3.3) is also a very important component of BO, as well as the search space (2.3.4). The initialization is here not considered as a component, as in most general BO implementations random sampling is employed.

2.3.1 Surrogate Model

The surrogate model in BO is an estimation of the objective function. It it used to provide predictions about not yet evaluated areas and uncertainty estimates. It is also referred to as the prior and then the posterior, after it got updated by the Bayes' Theorem with new evaluations on the objective function. Most use cases utilize the Gaussian Process (GP) [**Rasmussen.2005classification**], whereas within this work Tree-structured Parzen Estimator (TPE) [10] is selected, as it is intended to work also with nominal categorical variables. In particular, it can better optimize the selection of augmentation operations. In state of the art TPE works, random sampling is utilized by default for the initial parameters [11] just like in BO with GPs.

2.3.2 Acquisition Function

The acquisition function in BO balances exploration and exploitation to explore unknown areas of the function while finding the single best x-value within good areas. Some common acquisition functions are Expected Improvement (EI) [12] and Upper Confidence Bound (UCB) [13].

2.3.3 Evaluation Function

The evaluation function in BO is the actual objective function that is tried to be estimated and optimized. However, there are multiple works that focus on evaluating a proxy task, that is much cheaper to evaluate but has similar results like the target task or objective function. One such work is Fast AutoAugment (FastAA) by Lim et al. [9] as described in subsection 2.2.2.

2.3.4 Search Space

The search space encompasses all possible parameter combinations that can be evaluated on the objective function. The starting points for BO are most often randomly sampled within this space. However during this work Latin Hypercube Sampling (LHS) will be utilized. LHS is a statistical method used to generate samples which are evenly spread over the range of each variable. It is commonly used to sample efficiently from multidimensional distributions. This work employs this sampling in addition to the random sampling in hope to reduce the variance of the experiments.

2.3.5 Algorithm

The algorithm for standard BO is given in algorithm 1. Initially, Bayesian optimization randomly samples some x-values and evaluates them to have a starting data set D_0 to work with. Then the BO iterations loop starts for T iterations. First, the surrogate model (2.3.1) is fitted on all evaluated data points, creating data set D_t. In the second step, the acquisition function (2.3.2), which balances exploration vs exploitation, is maximized, given the current data set D_t. As the acquisition function is easily derivable, the x-value that maximizes the acquisition function is taken as the next point to be evaluated x_{t+1}. In the third step this x-value is evaluated on the objective function $f(x_{t+1})$. Lastly, the new information gained about the objective function $(x_{t+1}, f(x_{t+1}))$ is added to the data set of all evaluated datapoints D_{t+1}. After the budget of BO Iterations T was fully used, the best found $\hat{x}$ value which minimizes the objective function, is returned.

In the context of optimizing image data augmentation policies, this would mean that the policy, that resulted in the lowest density loss, is returned. Hence, the policy that minimized the difference and thus its augmented images are the most similar to the original images data distribution.

Algorithm 1 Bayesian Optimization

```
Input: Initial data points $D_0$, acquisition function $a$, maximum iterations $T$
Output: Minimizer of $f(x)$ for $t = 0$ to $T - 1$ do
1. Train surrogate model on $D_t$
2. $x_{t+1} = argmax_{x \in \mathcal{X}} a(x|D_t)$
3. $f_{t+1} = f(x_{t+1})$
4. $D_{t+1} = D_t \cup \{(x_{t+1}, f_{t+1})\}$
end for Return: $\hat{x} = argmin_{x \in D_T} f(x)$
```

3 State-of-the-Art

Numerous approaches have been developed with the objective of integrating prior knowledge into Bayesian optimization. This chapter introduces the four key components into which knowledge can be effectively integrated. Furthermore, this chapter will examine the variations in integrating knowledge into these key components. Emphasis will be placed on the scarcity of ways to specifically integrate expert knowledge, as most approaches rely on other sources of knowledge. In particular, many methods require unified databases, sometimes containing dozens of logged BO campaigns. However, such knowledge sources are not realistic in current industry standards.

3.1 Key Components to Integrate Expert Knowledge

This work began with a literature review, which served to construct a categorization (see Figure 3.1) and to provide an overview of the state of the art in expert knowledge integration. Fortunately, Bai et al. [14] recently conducted a thorough literature survey on transfer learning approaches for BO. They divided their categorization into four transfer learning BO components that are similar to the components that BO consists of. Another categorization of recent advances in BO was proposed by Wang et al. [15], which can also be used as a reference for comparison with the categorization in Figure 3.1. This work agrees with [14], that knowledge integration is achieved through the design of the surrogate model, the acquisition function, the initialization, and the search space. Each of the transfer learning BO components has its own advantages and disadvantages. The same is true for the individual approaches within each category. Therefore, criteria that is particularly important in the field of visual quality inspection is used to evaluate a few representative approaches. The approaches selected for evaluation were chosen either because of their influential impact (citations) or their distinctness to the other approaches within one transfer learning BO component. The criteria consists of the following:

Search Space Mapping

Search space mapping describes the ability of a method to optimize mixed search spaces, i.e. to optimize a combination of discrete and continuous parameters. This is important as the optimization of augmentation policies requires both the selection of augmentation operations (discrete) as well as the parameterization of magnitude and application probability (continuous).

- **Completely Fulfilled:** Mixed-value (continuous and discrete) optimization is supported
- **Partially Fulfilled:** Any essential BO component is mainly intended for continuous value optimization (e.g. GP), but a mechanism for handling discrete values is proposed and tested
- **Not Fulfilled:** Only continuous value optimization is supported

Practicality

The practicality of a method is determined by whether the requirements of the method are realistic and practical for current industry standards. Given that most research is focused on the invention of novel methods, many studies do not consider the practical issues that come with unrealistic requirements. In particular, the form and structure of knowledge is often expected to be stored in an easily accessible and structured format. However, this is not the case in most industrial settings.

- **Completely Fulfilled:** Required form and structure of knowledge is available in most industrial setting
- **Partially Fulfilled:** Required form and structure of knowledge is sometimes available in industry
- **Not Fulfilled:** Requirements are not realistic for most industrial settings

Reproducibility

In contrast, reproducibility considers only if and how the method is reproducible. How much effort is required to actually use the method, regardless of how realistic the must-have requirements are. Many industrial use cases for visual quality inspection would rather use an older method that has been tested and is publicly available than put a lot of effort into reproducing a technique. Therefore, reproducibility is a very important factor for many industrial domains.

- **Completely Fulfilled:** Method is publicly available
- **Partially Fulfilled:** Algorithm is clearly explained and method is reproducible with some development efforts
- **Not Fulfilled:** Method is not reproducible or only with extensive research and development

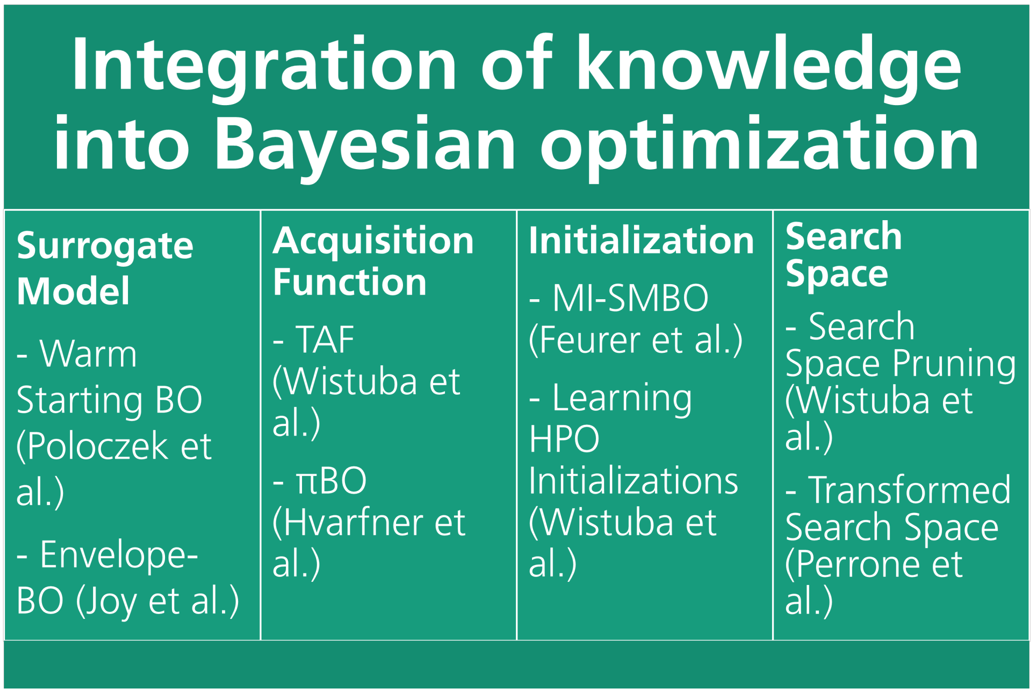
Usability
This category applies only to methods that support the integration of human expertise. Consequently, some human-machine interaction is required. The assessment evaluates how easy and intuitive the interaction is for those without a background in mathematics or coding. Once a method benefits from human input, the application should encourage them to enter new data, not discourage them, which would reduce the method's available data and thus its performance.

- **Completely Fulfilled:** Method is intuitive and easy for anyone to use
- **Partially Fulfilled:** Method can be used with a substantial background in mathematics or coding
- **Not Fulfilled:** Method is difficult to use even with a background in mathematics or coding (e.g. defining probability distributions without any user interface)

In order to gain an initial understanding of how knowledge can be infused into the BO process to accelerate convergence, some examples from Figure 3.1 will be elaborated. In addition, each method explanation is followed by an evaluation based on the defined visual quality inspection criteria.

3.1.1 Surrogate model

Starting with the design of the surrogate model, some examples include [16] and [17], both of which infuse prior knowledge from a database into the surrogate model. They take the data from previous experiments and utilize it to make



Integration of knowledge into Bayesian optimization

Surrogate Model	Acquisition Function	Initialization	Search Space
- Warm Starting BO (Poloczek et al.) - Envelope-BO (Joy et al.)	- TAF (Wistuba et al.) - πBO (Hvarfner et al.)	- MI-SMBO (Feurer et al.) - Learning HPO Initializations (Wistuba et al.)	- Search Space Pruning (Wistuba et al.) - Transformed Search Space (Perrone et al.)

Figure 3.1: Categories for knowledge integration in BO with some corresponding methods - inspired by [14]

the surrogate model smarter and more robust.

In particular, [16] creates a Bayesian prior probability distribution, or more specifically a Gaussian process (GP), to model the relationships of past objective functions to the current objective. This modeling is not done independently; rather, the GP jointly models the objective across tasks to leverage shared information, enabling the learning past tasks. The initial prior, namely the GP, is updated to the first posterior only with samples from previous tasks, thereby avoiding the need for new evaluations for randomly sampled initial points, as typically employed.

In particular, [16] introduces a Gaussian process (GP) as a Bayesian prior to model the relationships between past and current objective functions. This approach does not treat each task in isolation; instead, the GP jointly models the objectives across various tasks, leveraging shared information to enhance learning from previous experiences. By using samples from past tasks to update the initial GP prior to the first posterior, this method bypasses the necessity for new evaluations at randomly sampled initial points, as it is typically employed by default in BO. This method belongs to the "surrogate model" category, since the data and knowledge are integrated into the GP instead of the initial points. The fact that the initial points are also sampled more intelligently rather than randomly is merely a side effect, as some methods in one category interfere with other categories. Once the initial posterior has been computed from the past data points, the usual Bayesian optimization loop begins with the goal of maximizing the specific acquisition function. This is followed by evaluating the x-value on the target objective function and updating the prior to the posterior. The loop continues to make use of all samples from the current and previous tasks.

The method proposed by [16] does not score well on the proposed criteria specific to visual quality inspection. First, their approach is based on a GP and its properties without explicitly mentioning or testing a mechanism to optimize discrete values. Thus, the criterion of search space mapping is not fulfilled. However, they do explain their algorithm clearly. Although they have not published source code, it could be reproduced with some development effort, which partially fulfills the reproducibility criterion. Their approach assumes that users already have a clean, consistent database with multiple BO campaigns from similar use cases. Since this is not the case in most industry environments, the practicality of their approach is not fulfilled. They do not require any human-machine interaction, hence there is no need for an interface. Consequently, there is also no need to evaluate usability.

In contrast, the approach of [17] employs a distinct methodology to integrate historical knowledge into the surrogate model. Their approach takes observations from past experiments as noisy observations for the target task. The flexible noise level is estimated by an inverse gamma distribution. Whenever a point is evaluated on the target objective function, the inverse gamma distribution is updated based on the difference between its estimated value and the observed target value. As their calculations are based on observational data, the method is superior to selected previous approaches, including Yogatama & Mann [18] and Bardenet et al. [19] in unrelated task scenarios. Furthermore, the method can even completely ignore data points if the corresponding task is too unrelated to the target task.

The evaluation of [17] shows that it improves only slightly on our criteria, but is still far from being a good solution for the visual quality inspection domain. In the experimental section, the method was explored in much more detail. In addition to the clearly explained algorithm, the source code was made publicly available. This completely fulfills the reproducibility criterion. Once again, the method requires a unified database of past experiments, which is impractical for most in the industry, so the practicality criterion is not fulfilled. Consequently, the usability also does not need to be evaluated. The search space mapping criterion is not met because their implementation and approach is based on a GP without mentioning how to handle discrete optimization problems.

3.1.2 Acquisition Function

Some methods that integrate knowledge into the surrogate model encounter difficulties when scaling the model. Furthermore, many methods fail to acknowledge that new data is often more valuable than data from previous experiments [14]. In contrast, integration of knowledge into the acquisition function can eliminate these problems.

This exact argumentation also led Wistuba et al. [20] to infuse the knowledge solely within the acquisition function, as opposed to the surrogate model, as Witsuba et al. [21] did previously. The novel approach, as presented in [20] is called transfer acquisition function (TAF). Their acquisition function is consists of two main components, the predicted improvement on all previous experiments, and on the other hand the expected improvement (EI) on the new data. These two components complement each other well and result in a weighted average of the EI, with the former becoming less important as its knowledge is consumed and more new data becomes available. Regarding the process of obtaining the predicted improvement over all past experiments, they train a GP individually for each past experiment. Concerning their surrogate model component, they also chose a GP that was soley trained on the new data without any knowledge from past experiments. As discussed above, this acquisition function devalues data from past experiments over time. Hence, countering the problem of decaying importance of past data and the scalability problem.

The TAF method can be used with any surrogate model, since the knowledge is integrated into the acquisition function and does not depend on the properties of a specific surrogate model. However [20] only examines the case where a GP is integrated. Other surrogate models that support discrete optimization are not tested, nor is a mechanism for working with discrete data tested. Therefore, the search space mapping criterion is only partially met. Again, the method does not support a common source of knowledge in the industry. Since most companies do not have a large BO database, this criterion is not met. The method is partially reproducible, since it takes a considerable amount of development effort to code the algorithm from scratch using its algorithm description. Since the source code is not available, manual coding would be required to integrate the knowledge. However, the mere fact that there is not even an application makes the method fail the usability criterion.

Next, Hvarfner et al. [22] introduces a method that can be applied to many different acquisition functions. The Method is called Meta-learning-based Initialization Sequential Model-based Bayesian Optimization (MI-SMBO). In the paper, the authors concentrate on the application of their method to the EI acquisition function. They propose a generalization of the acquisition function that allows for the integration of knowledge. In this case, the knowledge is represented by a probability distribution, or prior, indicating the likelihood of the optimum being located at a given point. In essence, a weight is multiplied by the result of the acquisition function. This weight is based on how likely the parameter settings x are to be the optimum according to the given prior. In particular, if the user has defined the parameter combination as unlikely to be the optimum, the final value of the weighted acquisition function will be lower. Conversely, for parameter settings that are more likely to be the optimum, the value of the acquisition function is multiplied by a value greater than one. Furthermore, a decaying weight for the prior is introduced to provide the convergence guarantees of vanilla BO. Its mathematical background was inspired by works such as Souza et al. [23], which decays the impact of the prior converging to zero over time. Consequently, the theoretical convergence guarantees of [22] are also independent of the prior provided by the process expert. The experimental section in [22] is a good example of how to benchmark a new method. They compare their method to several state-of-the-art prior-guided approaches, including BOPrO by Souza et al. [23] and BOWS by Ramachandran et al. [24]. Unfortunately, not many researchers have the resources to do such a thorough comparison. A brief statement on this issue is presented in section 3.2.

The method proposed by Hvarfner et al. [22] allows the acquisition function to be augmented with knowledge in the form of a prior. Since it does not require a specific acquisition function or surrogate model, it can be used with many different frameworks. The paper even explores and tests the results using a random forest as a surrogate model, thus fully satisfying the search space mapping criterion by supporting mixed-value optimization. Practicality is also completely fulfilled because they don't require a single database with a lot of data. Instead, they use the insights and experience of process experts to warm start the BO. Reproducibility even has its own section in their paper, where they provide two links to their own implementations. In addition, they have implemented their algorithm in two popular BO frameworks (SMAC and HyperMapper), which greatly improves reproducibility and thus completely fulfills the the reproducibility criterion. The only criticism is that the actual application still requires the user to have a mathematical background. As the user is supposed to provide a handmade prior probability distribution for the location of the optimum. Since there is also no supporting interface to help the user create this essential parameter as defined in the criteria section (3.1), the method does not fulfill the usability criterion. This is still an important factor for practitioners in industry, as process experts are asked for something they cannot easily provide, which reduces the actual usage of the method.

3.1.3 Initialization

The convergence speed of Bayesian optimization depends heavily on the initial evaluations that form the initial posterior. This initial posterior has the potential to direct the optimization process towards promising regions, thereby facilitating faster convergence. The default initialization for Bayesian optimization is to randomly sample starting points. However, the following methods allow for the sampling of parameter combinations that are likely to perform well based on knowledge gained from past experiments. This process is also often referred to as warm starting.

One of the earliest and most successful papers in the field, focusing on the initialization component, was published by Feurer et al. [25]. Their approach generates meta-features for datasets, which can then be utilized to assess their similarity. Subsequent to the computation of these meta-features, they rank the source datasets based on the freely choosable distance metric, or in other words, they are ranked on the similarity to the target task. Ultimately, they take the top k closest datasets and their corresponding k-most promising parameter combinations as initial points for the target task. The experimental section of the paper presents substantial evidence that the proposed method for initializing BO is more effective than the previously state-of-the-art BO algorithms, namely Spearmint and SMAC. The experimental setup comprised 57 classification datasets, each with 46 computed meta-features. This is another good example of benchmarking a method, although as previously noted, it is not a viable approach for the majority of researchers. It should also be noted that at the end of the same year, Feurer et al. [26] published another article applying this methodology in practice. In particular, they applied the methodology with some modifications and another key ensemble technique to the popular sklearns AutoML system, successfully demonstrating its enhanced efficiency.

The methodology proposed by Feurer et al. [25] can be applied to any BO setup, as the calculated meta-features and the initialization points are not interfering with any other components of the BO process. Consequently, the methodology is fully compatible with mixed-value optimization, thereby completely fulfilling the search space mapping criterion. On the other hand, the number of datasets used to achieve these results is not even replicable for many researchers. Moreover, the number of companies that possess such a multitude of use cases in a unified database database is nearly nonexistent. Therefore, the practicality criterion is not fulfilled. In regard to the reproducibility criterion, the algorithm is clearly explained, as well as the parameter setup for the experimental section. It should be noted that, due to the unavailability of the source code, a significant amount of development would be required to effectively recreate the setup. Furthermore, it is challenging to obtain the necessary computational resources and data to reproduce similar outcomes. Consequently, the method only partially fulfills the reproducibility criterion. As the method does not interact with human process experts but extracts the knowledge from a database, the usability does not have to be evaluated.

Wistuba et al. [27] proposed a method that is based on gradient-based learning. In essence, the approximation of the meta loss is minimized through an iterative process. In this context, the optimizing meta loss refers to an optimization based on evaluations of parameter combinations across multiple datasets, with the utilization of predictions from plug-in models that have been trained on past campaigns. Each plug-in model is learned on an individual source task, thereby enabling the prediction of outcomes for never-evaluated parameter combinations. Additionally, their method does not require meta-features, a common practice in alternative approaches, which started with the just aforementioned paper [27]. These two fundamental concepts make their approach highly distinctive back then. The primary challenge of making the meta-loss differentiable was addressed by approximating it via a differentiable softmin function, thereby enabling their gradient descent-based approach to function. Finally, they proposed the addition of a weighting term based on the similarity of the data sets, with the objective of weighting predictions from similar campaigns more.

The method proposed by Wistuba et al. [27] is also a plug-in solution that does not interfere with other BO components, in accordance with the majority of works focusing on the initialization phase. Therefore, mixed-value optimization is supported, fulfilling the search space mapping criterion. However, this method also requires a uniform database of logged past experiments, which precludes the integration of alternative, more commonly available sources of knowledge. As a result, the method does not fulfill the practicality criterion. The authors provide a detailed explanation of their algorithm, yet the replication would demand considerable time and resources, thus only partially fulfilling the criterion of reproducibility. As the method relies on a database, its usage is not realistic for many industrial settings. Hence, the practicality criterion is not fulfilled. In the absence of human-machine interaction, usability does not need to be evaluated.

3.1.4 Search Space

Another potentially highly orthogonal component is the search space. Since the search space can be reshaped or pruned, without interfering with other BO components, it is an ideal candidate for combination with other knowledge-enhanced components.

Search space shaping has the advantage that it is often a modular component. It is possible to modify the search space without affecting the functionality of the surrogate model or an acquisition function. One such approach was proposed by Peronne et al. [28]. They eliminate the sets of arbitrary search ranges and introduce automatically learning search space geometries on source tasks. The search space geometries are learned via a defined constraint optimization problem. Their paper proposes two such space geometries, a bounding box (or hyperrectangle) and a hyperellipsoid. The hypothesis that hyperellipsoids are more appropriate when the parameter combinations do not cluster in the corners of the bounding box, was empirically supported in their experimental section.

It is explicitly stated that any search space can be transformed by their approach. Furthermore, the experimental section included an investigation of mixed-value optimization. As mixed-value optimization is fully supported, the search space mapping criterion is completely fulfilled. Next, the practicality is not fulfilled, as the approach once again requires the practitioner to have a uniform database of logged experiments, which is not an industry standard. The method is described as straightforward to implement, and the algorithm is clearly explained. However, there is no official implementation nor any other third-party source code available. Implying that the recreation and eventual utilisation of the method would require a certain effort in terms of development. This results in the method's reproducibility being only partially fulfilled. A usability evaluation is not necessary, as there is no option of extracting knowledge from process experts.

Wistuba et al. [27] was the first to propose adding this fourth component of search space shaping to hyperparameter optimization. As the technique is designed to prune any search space, it is also applicable to search spaces within Bayesian optimization. The fundamental concept is to prune regions that are unlikely to contain good parameter combinations based on knowledge learned from source tasks as well as evaluations already performed on the target task. The method assumes that source tasks with similar data sets also have similar or even the same regions with bad hyperparameter combinations. The identification of these bad regions is accomplished through training plug-in GPs on all normalized source tasks. The distance between each source task and the target task is calculated, resulting in the preservation of the k-closest datasets. The trained plug-in estimators are employed to estimate the potential of parameter combinations. The parameter combinations with minimal potential define the regions with radius δ which are to be pruned and are referred to as δ-regions. The technique was evaluated using two newly created metadata sets, one of which consisted of 59 different data sets with 19 different classifiers, for a total of 1.3 million experiments. The results demonstrated that in all cases, the technique either resulted in a significant improvement in performance or had no detrimental effect.

This technique operates in a manner that is orthogonal to other BO components, thereby ensuring that it does not interfere with them. In addition to facilitating both continuous and discrete search space optimization, the technique also explicitly proposes a distance metric for categorical values, thereby completely fulfilling the search space mapping criterion. However, the technique also requires a uniform database comprising a sufficient number of datasets to identify the k most similar ones. As the majority of companies lack even a single clean logged data set, this approach does not fulfill the practicality criterion. The algorithm for recreating the technique is provided and clearly explained. However, it would require significant development efforts to utilize it in practice. Consequently, it only partially fulfills the criterion of reproducibility. The usability of the technique does not require any evaluation.

3.1.5 Summarizing the Evaluation

The principal conclusion to be drawn from the assessment of criteria tailored to the domain of visual quality inspection is that there are some discrepancies between the findings of research and the practices of industry. A significant issue is that research frequently fails to consider which industrial standards are in place and instead prioritizes the development of novel techniques that push the boundaries of the state of the art. The most troubling finding of the evaluation is that there is a lack of research aimed at leveraging the process experts' intuition, despite its pervasive presence in the industry. Consequently, the proposal of novel methods for the integration of expert knowledge could prove highly beneficial for practitioners in industrial settings. One such example is the work of Hvarfner et al. [22], which made a significant contribution towards the integration of expert knowledge into BO. However, the proposed acquisition function requires the user to manually define probability distributions, and no human-machine interface is provided to assist the process expert in designing such priors. Therefore, another research gap exists, namely the development of more intuitive human-machine interactive interfaces that can be used to extract intuition and knowledge from ordinary process experts who lack a background in mathematics or coding.

In essence, the objective is to expand the range of usable knowledge sources to align with industry standards. Enabling process experts to input their experience and intuition into the BO process in an accessible manner has the potential to significantly enhance the visual quality inspection domain.

3.2 Identified Research Gaps

The literature review has revealed several research gaps within the field of visual quality inspection. This section will explain and summarize these gaps, highlighting the areas where further research and development are needed to enhance the effectiveness and accuracy of visual quality inspection technologies.

3.2.1 Uncomparable Results

This first aspect is not only specific to the visual quality inspection domain, rather every single domain. The problem of having no common benchmark to quantify results is a real problem. Many other research areas like traditional image classification, object detection, segmentation and many more have public datasets to benchmark their methods [29] [30]. These data sets are then used to have a fair benchmark for their method against others. Instead, the newly developed methods are most often compared against weak baselines, consisting of much older approaches or even basic BO without any knowledge integration (e.g. [17, 24, 31]).

3.2.2 Selection of Component to Integrate Knowledge

Another problem is that not only the methods themselves have uncomparable performances, but consequently the components have the exact same problem. If methods would have transparent results, it could be derived, into which component practitioners should infuse their knowledge. Not only that, but researchers might also want to focus on the components that bring more benefits, or on the other hand components that lack research. It would overall make the entire process much more transparent.

3.2.3 Unrealistic Method Requirements

Currently, in research within the integration of knowledge into Bayesian optimization it is the norm to ignore industry standards. Almost all methods focus on using knowledge from uniform databases, without acknowledging the fact that most practitioners are not be able to use their method as they can not fulfill these unrealistic requirements.

3.2.4 Unreproducible Results

The literature review shows that most approaches do not have public source code from either the authors or any other third-parties. This makes it extremely difficult for any practitioner to reproduce the results or even apply any of the novel methods to a new use case.

3.2.5 Methods Are Not Used

All these problems lead to making most of the not being used after they have been developed. This is not only true for most people in industry but apparently even the most knowledgeable researchers tend to use traditional random and grid search over novel hyper parameter optimization methods. The question arises, why is so much effort put into creating novel methods and not making them viable, if in the end even the best researchers themselves do not use these methods [7]. Identifying causes within the visual quality inspection domain was done in this literature review. Possible solutions will be discussed in the discussion section (6).
Nevertheless, one particular issue clearly stood out specifically important in the visual quality inspection domain. That being the limited range of knowledge sources. As analyzed above most methods do not incorporate any kind of expert knowledge, instead they lean on unrealistic large unified data bases. This shows that there is a significant research gap in regard to extracting knowledge from process experts and eventually integrating that extracted knowledge into the Bayesian optimization process. These research gaps will guide this works research question (3.3).

3.3 Research Questions

The analysis of current state-of-the-art in integrating knowledge into the Bayesian optimization process opens up many different research gaps. This work is supposed to contribute towards filling two of them specifically important to the visual quality inspection domain. Namely, how can expert knowledge be extracted from process experts, as this is not even handled within the single approach analyzed here that did incorporate knowledge from process experts [22]. The other research gap being how this extracted knowledge then can be incorporated into the Bayesian optimization process, to make more sources of knowledge transferable. The deriving primary research question can be broken down into four sub-research questions, as follows:

Primary Research Question: How can expert knowledge be utilized for faster convergence in the Bayesian optimization of augmentation policies in visual quality inspection?

Sub-Research Questions 1: How to extract expert knowledge from a process expert within the visual quality inspection domain?

This refers to designing and developing an interface that can extract an experts experience and intuition. The focus should lie in making the human-machine interaction as easily accessible for the process expert, to not exclude the expertise of people without a mathematical or coding background. As every domain might require a different interface for a seamless interaction, this interface should aim at highlighting how visual quality inspection experts can contribute their knowledge.

Sub-Research Questions 2: How to integrate the extracted expert knowledge into the Bayesian optimization process?

Development of two methods to integrate expert knowledge into the Bayesian optimization process. This means identifying which components might be suitable to integrate expert knowledge and then developing a method for these components.

Sub-Research Questions 3: Which benefits does the integration of expert knowledge result in?

The two developed methods should be tested in experiments to figure out the advantages of integrating expert knowledge either way. Resulting in a statement on which component using the developed way of extracting and integration knowledge was more successful.

Sub-Research Questions 4: Does the combination of integrating expert knowledge into two components give any additional benefits?

After testing the methods individually, both components should be combined and tested in combination. This idea was proposed as having huge potential by [14, p.29].

4 Methodology

4.1 Concept

The theoretical investigation of the current state of the art revealed that there are very few methods that actually integrate expert knowledge. The only method within the analysis that did integrate expert knowledge, namely Hvarfner et al. [22], did not assist the human in creating the required probability distribution. The same usability problem is prevalent in similar approaches, such as an earlier approach by Souza et al. [23]. Thus, it is not easily usable even by people with some mathematical or programming background. This section attempts to answer the first two derived research questions (3.3). Namely, by presenting a method for extracting the intuition and experience of visual quality inspection process experts to answer the first question. The process of extracting expert knowledge is illustrated as a flow chart in Figure 4.1. The second research question is answered by introducing two methods that integrate the extracted knowledge into two different transfer learning components.

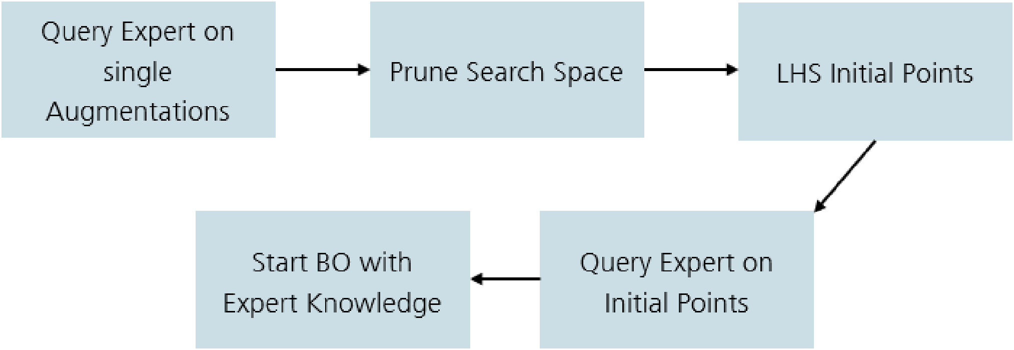

Query Expert on single Augmentations
Prune Search Space
LHS Initial Points
Query Expert on Initial Points
Start BO with Expert Knowledge

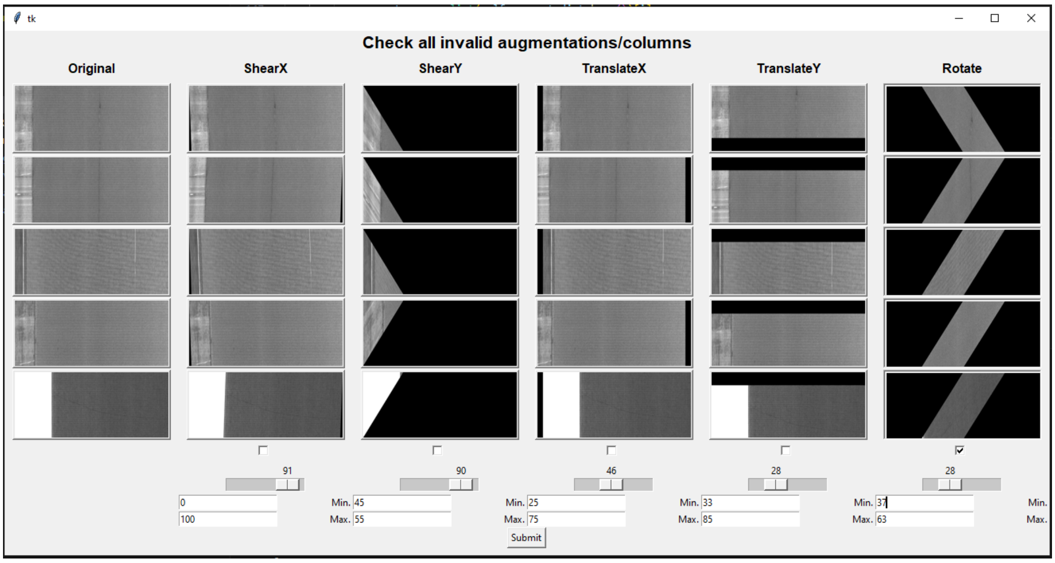
Figure 4.1: Flow Chart of the developed method

4.1.1 Extracting Expert Knowledge

Optimizing data augmentation policies in the context of visual quality inspection has the advantage that the effect of data augmentation can be easily presented to the process expert. In particular, if a data augmentation policy is to be evaluated based on the knowledge of the process expert, a particular sub-policy (2.2) can be applied to just one image and the process expert can already provide his insights to the machine. This makes it a highly intuitive approach for knowledge extraction, placing great emphasis on what the process expert is used to. Given that they have extensive experience dealing with visual defects on a daily basis, for several years, they possess a lot of experience and intuition about which augmented images might be realistic and useful, or unrealistic and probably even confuse the model.

Naturally, the proposed method takes full advantage of this fact by initially displaying all available augmentations to the process expert. However, an initial challenge is encountered when considering the variability in the effectiveness of augmentations depending on the specific image to which they are applied. In other words, one augmentation might make some defects undetectable, while other defects might still be visible. To address this issue, each augmentation is applied to five distinct images. Consequently, each column in Figure 4.2 consists of five augmented images, with the magnitude (if applicable) being adjustable via a slider. The first column on the leftmost side always shows the original images as a comparison. The process expert can then determine which individual augmentations make defects undetectable and hence should be pruned from the search space.

Additionally, a "magnitude tuning" approach is introduced, which enables the process expert to prune not only entire augmentation but also specific magnitudes. This allows experts to define a valid range within which the augmentation remains effective and identify the magnitudes at which the augmentation begins to negatively impact the data. Moreover, they can use the slider located below the images column to test how changing the magnitude alters the augmented images. The min-max input fields, located below the slider, allow the user to specify the magnitude range that remains valid. This functionality is exemplified in Figure 4.2 In summary, only the image columns checked as invalid, representing individual augmentations, and the magnitude ranges that fall outside the as valid defined min-max ranges for the remaining valid augmentations are being pruned. This method is related to the search space pruning proposed by Wistuba et al. [27], which was analyzed in the theoretical investigation. However, the method [27] is distinct in not only the fact that it does not support expert knowledge integration, but also employs a very different approach to pruning the search space.



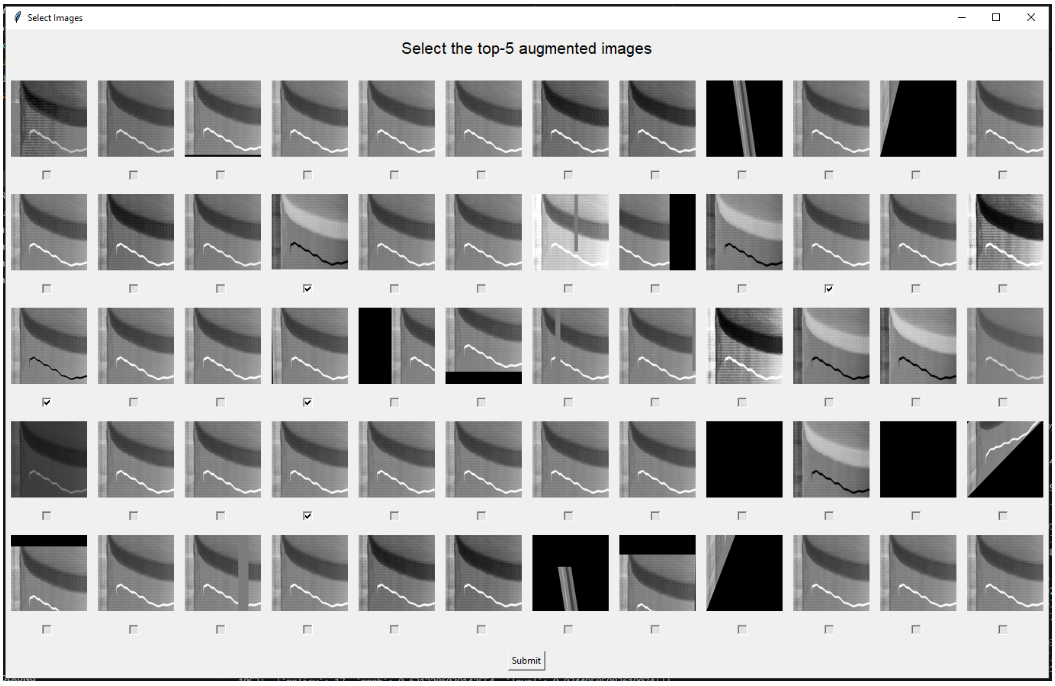
Check all invalid augmentations/columns

Original	ShearX	ShearY	TranslateX	TranslateY	Rotate
	□	□	□	□	☑
	91	90	46	28	28
	0	Min. 45	Min. 25	Min. 33	Min. 37
	100	Max. 55	Max. 75	Max. 85	Max. 63

Min.
Max.

Figure 4.2: Search Space Shaping - removing augmentations and tuning magnitude ranges

Subsequently, after pruning the search space, Latin Hypercube Sampling (2.3.4) is utilized to create sub-policies that are evenly sampled from the pruned search space. It is cruicial to note that the sub-policies consist of two subsequent augmentations, in contrast to the initial step, which involved the pruning of individual augmentations. The process expert is presented a collection of the same original image, each of which has been augmented by an LHS-sampled sub-policy. The expert is then tasked with selecting the k-most promising augmented images, as seen in Figure 4.3. These selected augmented images, or more precisely the underlying sub-policies used to augment them, are then passed into the fifth and final step, as illustrated in Figure 4.1.


Select Images

Select the top-5 augmented images

Figure 4.3: Initial Points Selection

4.1.2 Integrating Expert Knowledge

The integration of expert knowledge into the selected transfer learning components is a relatively straightforward process. Especially after understanding what knowledge is extracted from the process experts. Given that the previous extraction returns a search space that has been pruned by the process expert and that the sub-policies within that search space have also been selected by the process expert, both objects can be employed directly. Firstly, it is not necessary to search through the entire search space, instead, only the pruned search space, which the process expert believes does not contain any confusing augmentation operations. Thus, our first method is categorized into the search space transfer learning component for BO. Next, the Bayesian optimization process can be warm started with the selected most promising sub-policies. Motivated by the potential that these promising points can guide the BO into the right direction, potentially accelerating the identification of optimal image augmentation policies. Therefore, the second proposed method can be considered to belong to the initialization transfer learning component.

An important assumption that is being made here is that the initialization method is not conditionally related to the search space component, given a sufficiently large number of sampled sub-policies using LHS. This is because the process experts only prune those augmentations which they deem unfavorable. Consequently, with a sufficiently large collection of LHS sampled sub-policies, it is reasonable to conclude they would not select other sub-policies that include previously pruned augmentations. In conclusion, it is assumed that the initialization and search space methods are unrelated, as long as the collection of sampled sub-policies is sufficiently large.

4.2 Implementation

The implementation is again divided into the implementation of the expert knowledge extraction method and the development of the two expert knowledge integration methods.

4.2.1 Extraction Implementation

In order to ensure that the developed method is accessible to any process expert, regardless of their educational background, a significant emphasis was placed on the design of an intuitive interface. This interface is intended to facilitate the interaction with the process expert during the two query phases, as described above and illustrated in Figure 4.1. The user interface, which enables human-machine interaction, is a crucial element influencing the extent to which a method is utilized after its development. This extraction method was implemented in Python using a library called Tkinter, which was introduced by Lundh, Frederik [32]. The application has been developed in Python 3 [33] and consists of four Python scripts. The main script initiates the search space pruning script. The search space pruning script then generates the columns of images with the corresponding interactable objects. These include a checkbox to prune the entire augmentation, a slider to adjust the applied magnitude, and the min-max input fields to only allow a specific magnitude range. The augmentations are encoded within the augmentations script, which was taken from the official source code of FastAA by Lim et al. [9]. Once the search space has been pruned, the main script calls the selection of initial points script. In this script, the LHS method is used to equally sample from the remaining pruned sub-policy search space. The LHS implementation was taken from Scipy [34] a very popular and powerful Python library for data science. The user-selected augmented image indexes with the corresponding sub-policies are returned to the main script. Ultimately, the definition of the pruned search space and the sub-policies selected as initial points for the Bayesian optimization process are printed into the console, without direct integration with the other two knowledge integration methods.

4.2.2 Integration into Search Space Implementation

The integration of expert knowledge was conducted within a preexisting framework developed by the Fraunhofer IPT for the tuning of hyperparameters with Bayesian optimization. This work focused exclusively on the extraction of expert knowledge and integration of expert knowledge into Bayesian optimization for the purpose of identifying optimal image augmentation policies. It is noteworthy that no effort was made to construct a suitable convolutional neural network (CNN), tune the CNN, build the entire pipeline of Bayesian optimization, or undertake any related tasks.
Instead, a considerable amount of time was spent familiarizing oneself with the provided framework and understanding how the pipelines work and what exactly needs to be changed to integrate the extracted knowledge into the Bayesian optimization process within this custom framework. The provided framework is based on the HyperOpt framework [35] with their Tree-structured Parzen Estimator (TPE) [36], which serves as the surrogate model (2.3.1). The search space was implemented in precisely the same manner as in the official source code from FastAA [9]. Consequently, the contribution of this work involved the development of a function that maps the BO chosen *magnitude* to the valid ranges defined by the process expert. This was done using the following mathematical operation:

$$\text{mapped\_magnitude} = \text{min\_val} + \text{magnitude} \times (\text{max\_val} - \text{min\_val}) \tag{4.1}$$

With *mapped_magnitude* being the final magnitude used for evaluation, and [*min_val*, *max_val*] being the valid magnitude range for one specific augmentation. Both, *min_val* and *max_val* are the user defined valid ranges within [0,1]. Consequently, the user-inputted values from the application that fell within the range of [0,100] were divided by 100. It is noteworthy that the BO process is unaware of the resized parameters. BO suggests a new sub-policy with two defined augmentations and their respective magnitude and probability parameters. Ultimately, it only receives a score, in our case the density matching score (2.3.3) of the augmented image and the original image. The parameter range mapping is called within the augmentations script. In this script, each augmentation uses the parameter range mapping function with its corresponding *min_val* and *max_val* parameters, as defined by the process expert.

The pruning of entire augmentations is very straightforward. As the search space is defined as a list, the augmentations selected as invalid are simply removed from the list.

4.2.3 Integration into Initialization Implementation

The objective of this subsection is to integrate the sub-policies selected by the process expert as the initial points for evaluation. The most challenging aspect of initiating Bayesian optimization with pre-defined initial points is that each search space has its own representation of how the data should be structured. In our case the list of dictionaries was encoded as follows: each key contained the sub-policy id, in addition to either the name of an operation, probability, or a magnitude. The value of this long key was either a string of an augmentation, or a decimal value for the parameters. Each initial point must include 5 sub-policies, resulting in a total of 10 augmentation operations, each with its own probability and magnitude parameters. This design aligns with the original policy structure used in FastAA [9]. This list of dictionaries can then be parsed into the AugmentHyperOptSearch object as a parameter called *points_to_evaluate*. As soon as the structure of the list of dictionaries is figured out the HyperOpt searcher uses these points as the initial evaluation points, without any further issues.

4.3 Experimentation

The technical details of the implementation are described in section 4.2. However, in this section, the general setup including which data set is used and which parameters are used within the provided BO framework is elaborated (4.3.1). Additionally, within this section the experiment plan with its design choices is presented (4.3.3). The results of this experimentation will then be presented in the results section (5.2) and discussed in the discussion section ().

4.3.1 Setup

Starting with the setup. In particular the dataset chosen to be a good representation for the visual quality inspection domain is called AITEX Fabric Image Database [37]. The dataset is taken from Kaggle, a popular data science platform for competitions and discussions. The data set consists of 245 images of 7 different fabrics with 12 defect classes, ranging from point defects to large scratch defects, as shown in Figure 4.4. The data set contains 140 defect-free images and 20 defect images for each of the 12 defect classes. The images have a size of 4096x256 pixels.

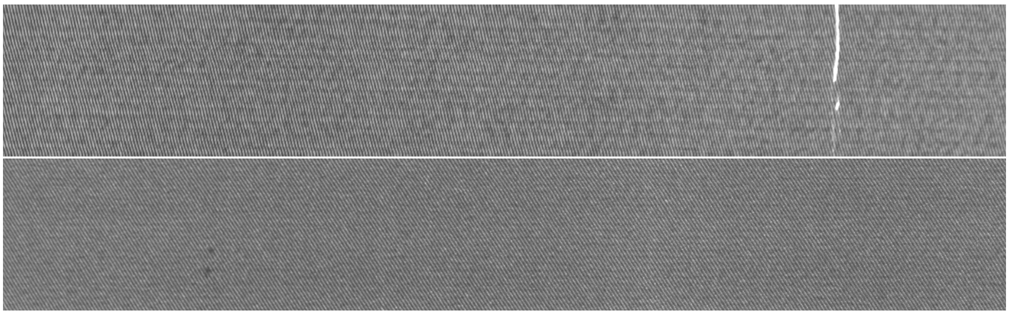


Figure 4.4: Aitex data set - zoomed in images with scratch defect (top-right) and dot defects (bottom-left)

In the beginning the data set is splitted into a train and test set (see Table 4.1). The test split is kept as a holdout set.

Table 4.1: Custom Split of the Aitex Data Set

	No Defect	Defect
Train	75	75
Test	66	31

4.3.2 Training and Optimization Pipeline

The explanation of the pipeline is kept to a minimum, as this work did not contribute to the development of the general pipeline. In the first step of the provided pipeline (see Table 2), a CNN is trained with k-fold Cross Validation [38, p.241] on the non-augmented train split, with $k = 5$. The CNN is a binary image classification model that classifies images into 'defect' or 'no defect', and is always trained for 100 epochs for the following experiments. Subsequently, the Bayesian optimization to find an optimal augmentation policy is started. Optionally within this second step (algorithm 2), a pruned search space or/and a warm started initialization can be integrated, as described in section 4.2. The components used for this Bayesian optimization are as described in subsection 4.2.2, TPE as a surrogate model to effectively search through the mixed-value search space. Consisting of the continuous magnitude and probability parameters, as well as the nominal augmentation operations. The evaluation function employed within the BO step is the mentioned density matching (2.3.3) to make optimization feasible. After the number of BO iterations ran out, the best-policies are used to augment the original Aitex train split, in step 3 of algorithm 2. Finally, another CNN is trained on the augmented data set, also with k-fold cross validation and $k = 5$. The last step involves evaluating the ten k-fold CNN models, of which 5 were trained on a non-augmented k-fold split and the other 5 on the augmented k-fold split. An important not mentioned parameter is the number of BO iterations, which will be a part of the experimental plan (4.3.3).

4.3.3 Design choices

Starting with the design choice of how many LHS-sampled sub-policies should be computed, the investigated rule of thumb $n = 10d$ suggested by Loeppky

Algorithm 2 Training and Optimization Pipeline

```
Train a CNN with k-fold on the original train data split
Run BO to find optimal image augmentation policies using density matching
Apply best policies to original data set
Retrain the CNN with k-fold on the augmented dataset
Evaluate the k-CNN models on the holdout data =0
```

et al. [39] will be followed. With n being the amount of drawn samples and d being the number of search space dimensions. As each sub-policy contains two augmentation operations, two magnitude parameters and two probability parameters, the search space dimension is six. Resulting in a total of $n = 10 \times 6 = 60$ LHS-sampled sub-policies. This number is also expected to fulfill our requirement that it is large enough to conclude that the initialization method and the search space pruning method are unrelated (4.1.2).
In regard to selecting the amount of initial points for Bayesian optimization, there is only an uninvestigated rule of thumb of selecting between $10\% - 20\%$ of BO iterations as initial random evaluation. But as we do not utilize the default random sampling and as we have much more BO iterations than in most use cases, due to the efficient evaluations with density matching, this rule of thumb will not be used. Instead, priority is given to the process expert, and it is unlikely that they are interested in selecting dozens of images out of an even larger collection. Therefore, this approach will take the $n = 60$ LHS-sampled sub-policies and let the process expert select five final sub-policies out of them, that they think are realistic and beneficial.

The experimentation plan with the as above elaborated setup, is presented in the following.

Setup:

- LHS-samples: 60 ($n = 10 \times 6 = 60$ [39])
- BO initial points: 5
- BO iterations: 3000

With BO iterations being the budget for how many iterations the optimization can search through the provided sub-policies search space. The value was determined by singular runs on different budgets, as presented in algorithm 5.2. LHS-samples, referring to the number of sub-policies sampled with LHS and then presented to the process expert. BO initial points being the budget of how many evaluations BO can use before stopping the optimization process.

Experiments:

- Baseline: default BO
- Warm Started Initialization: Initial Points are selected by expert from LHS samples
- Pruned Search Space: Search Space is pruned via augmentations removal and magnitude tuning
- Full Expert Knowledge: Warm start and pruned search space combined

Baseline

The first experiment, simply termed Baseline, utilizes the default Bayesian optimization implementation. Hence, no expert knowledge is integrated. The default initial sampling with a TPE surrogate model is random sampling [11], analogous to that employed in BO. Consequently, this baseline also samples the initial points randomly.

Warm Started Initialization

The Warm Start experiment aims to investigate the benefits of incorporating expert knowledge-infused initial points. The extraction of the initial points was already described above (4.2.3). It is important to note that this experiment still uses the entire search space, hence the improvements can be entirely attributed to the initialization method.

Pruned Search Space

As its name indicates, the Pruned Search Space experiment is designed to identify improvements resulting from the reduction of the search space. This is achieved as described in subsection 4.2.2, allowing the user to eliminate invalid augmentations and adjust the magnitude range of the valid ones. Once again, this experiment does also not contain any initial points that have been infused with expert knowledge. Instead, random sampling is used, as this is the default setting.

Full Expert Knowledge

Lastly, the Full Expert Knowledge experiment encompasses both expert-selected initial points and a pruned search space. This approach is inspired by Bai et al. [14, p.29], trying to assess the mentioned potential of combining orthogonal methods from different BO transfer learning components. This method, might be best evaluated against the Robust Baseline, as LHS is once again used before querying the process experts.

5 Results

This chapter will summarize the key findings from the theoretical investigation of how expert knowledge is integrated into the BO process (5.1) to collect the content needed to answer this works research questions within the discussion chapter (6.1). Furthermore, the results from the practical investigation as described in the experimental section (4.3) will be described and visualized (5.2). As they are also a key contribution of this work to answer the research questions.

5.1 Theoretical Investigation

5.1.1 How to Integrate Expert Knowledge into BO

Firstly, it was found that expert knowledge is very rarely integrated into BO. Instead, most methods do their transfer learning approach by utilizing a database which contains data from past experiments [16, 17, 20, 25, 27, 28]. However, these methods can not be applied in most industrial settings, as most practitioners in industry do not have a structured database. Thus, leading to a new open research gap, in particular that current transfer learning methods for BO have unrealistic requirements (3.2.3). Consequently, making this work and other works contributing towards transferring more accessible sources of knowledge into BO even more important. One such approach that transfers expert knolwedge instead of knowledge from a large database was proposed by Hvarfner et al. [22]. In essence, it allows to augment the aquisition function based on a user defined probability distribution of how likely specific parameters are to perform well. As this work was inspired by an earlier approach from Souza et al. [23], there are a very few works that focus on making more sources of knowledge transferable. However, they do not assist the experts by giving them an interface to define such priors. Therefore, there is still some research to be done.

5.1.2 More Research Gaps Discovered

More research gaps not directly related to this works research questions were also found and pointed out. Including that methods are not easily comparable within this research field (3.2.1, 3.2.2) and the methods are not easily reproducible (3.2.4) making it much harder for practitioners to actually use or select one of these novel methods (3.2.5).

5.2 Practical Investigation

This section will present the results from the practical investigation, in particular from the experiments defined in 4.3.

The amount of BO iterations was determined to be 3000, which is very large compared to most BO settings. However, due to the very cheap density matching evaluation function it is a viable amount. Nonetheless, each experiment, including the training of the first k-fold CNN with $k = 5$ and the policy optimization, lasted on average 4.75 hours. Additionally, the final retraining on the augmented data set lasted another 1.5 hours.

5.2.1 Experimentation Results

The results of all experiments are presented in Figure 5.1. The plot on the left depicts the final k-folds CNN models trained on the augmented data set. Each boxplot contains five accuracy scores from the final k-CNN models, evaluated on the never-seen holdout set. The plot on the right contains all CNN models trained on the original data set and evaluated on the holdout set. It is noteworthy that the models trained on the augmented data for the selected parameter settings perform worse on the holdout set than the models trained on the original data. This effect is made even more obvious in table 5.1.

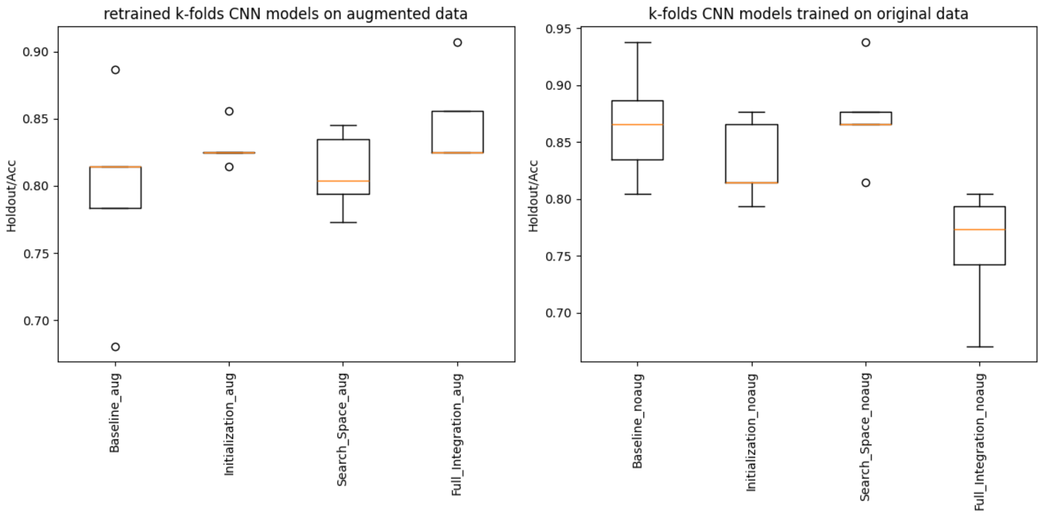


Figure 5.1: Boxplots of all experiments as defined in section 4.3

Experiments	**Mean**	**Median**	**Variance**	**Std Dev**
CNN models trained on the augmented data set				
Baseline_aug	0.795876	0.814433	0.004481	0.066939
Initialization_aug	0.828866	0.824742	0.000196	0.013984
Search_Space_aug	0.810309	0.804124	0.000706	0.026565
Full_Integration_aug	0.847423	0.824742	0.001037	0.032207
CNN models trained on the original data set				
Baseline_noaug	0.865979	0.865979	0.002083	0.045641
Initialization_noaug	0.832990	0.814433	0.001037	0.032207
Search_Space_noaug	0.872165	0.865979	0.001556	0.039446
Full_Integration_noaug	0.756701	0.773196	0.002321	0.048179

Table 5.1: Accuracy Statistics Summary on Holdout Split

6 Discussion

This chapter addresses the research questions of this work (3.3), which will be answered in section 6.1. Most contributions are already outlined in the results chapter (5). However, further discussions and interpretations are required, particularly in regard to the practical investigation.

6.1 Research Questions

This section initially addresses the sub-research questions, which are a segmented version of the primary research question. Subsequently, a concise response is provided to the primary research question.

Sub-Research Questions:

S-RQ1: How to extract expert knowledge from a process expert within the visual quality inspection domain?

The comprehensive literature review (3) revealed that the majority of existing methods do not integrate expert knowledge, but rather focus on other sources of knowledge (3.2.3). Moreover, the few methods that do integrate expert knowledge lack an effective approach for extracting that knowledge. This work proposes a method for extracting expert knowledge that enables process experts to input their knowledge in an intuitive and highly accessible manner (4.1.1). In particular, the method allows process experts to prune specific augmentation operations or specific magnitude ranges that result in overly aggressive augmentations (4.2.1).

S-RQ2: How to integrate the extracted expert knowledge into the Bayesian optimization process?

for integrating any form of knowledge into BO. Ultimately, it was decided to integrate the extracted expert knowledge into the initialization transfer learning BO component and the search space transfer learning BO component (3.1), as described in subsection 4.1.2.

S-RQ3: Which benefits does the integration of expert knowledge result in? The results of the experiments in section 5.2 can not demonstrate significant advantages of integrating expert knowledge into BO. As this experimental section has a narrow scope, the variance is too high to conclude any empirical proofs. The ride plot in figure 5.1, containing the models trained on the augmented data set, seems to indicate that the knowledge infused methods perform slightly better on the holdout accuracy score. However when the variance described in table 5.1 is taken into account, or when considering the spread within the boxplots, it becomes evident that this is not indicative for of superior performance.

S-RQ4: Does the combination of integrating expert knowledge into two components give any additional benefits?
Last but not least, the combination of orthogonal transfer learning BO components was tested, as proposed by Bai et al. [14, p.29]. However, the experiments as displayed in the results section 5.2.1 do contain very high variance. It could indicate at performing better, but it is definitely not possible to draw a concrete conclusion.

Primary Research Question

RQ: How can expert knowledge be utilized for faster convergence in the Bayesian optimization of augmentation policies in visual quality inspection?
Ultimately, to utilize expert knowledge in BO, a method was developed to extract expert knowledge (4.2.1) and integrate it into two distinct transfer learning BO components (4.2.3, 4.2.2). Moreover, both methods were evaluated within a limited experimental scope. However, the selected experimental parameters were insufficient to demonstrate any notable enhancements in accuracy or accelerated convergence.

6.2 Limitations

The theoretical investigation successfully devised an effective method for extracting and integrating expert knowledge into BO. However, the practical investigation lacks a comprehensive benchmarking, as the variance is too high. Another limitation is that the warm started initialization method may prove to be a more effective approach in a different optimization scenario that does not involve density matching. The inclusion of density matching allows for an abnormally high number of BO iterations, which makes the initial sampling process almost irrelevant.

7 Conclusion & Outlook

This bachelor's thesis investigated the potential for leveraging expert knowledge to accelerate BO. In particular, BO optimizes image data augmentation policies with the aim of increasing the data set volume and variety within the visual quality inspection domain, which suffers from data imbalance and data scarcity. In order to utilize expert knowledge within BO, a thorough literature review about the state of the art of integrating any kind of knowledge into BO was conducted (3.1). Moreover, further research gaps were identified and highlighted (3.2). Subsequently, a method for querying and extracting the knowledge of process experts within the visual quality inspection domain was developed. This extracted knowledge was then integrated via two methods that work orthogonal to each other. The effectiveness of these two expert knowledge integration methods was then tested in a narrow experiment. However, due to an underestimation of the experimental scope, no significant improvements can be concluded. Further details and other limitations are outlined in subsection 6.2.

Further research could entail a more comprehensive practical investigation of the advantages of integrating expert knowledge. Furthermore, examining retrained models across distinct BO iterations may prove beneficial in gaining insights into the methods' performance. A comparison of expert knowledge integration to transfer learning approaches that utilize databases would also be of interest. However, it is important to remember the findings from the theoretical investigation, that the majority of methods only allow the integration of structured knowledge from a database. This is a significant limitation, as it does not align with the standard of most industrial settings. Consequently, any research aimed at enriching the range of knowledge sources that are transferable, such as expert knowledge, is important and can have significant benefits for practitioners.

Bibliography

[1] X. Dong, C. J. Taylor and T. F. Cootes. 'Automatic aerospace weld inspection using unsupervised local deep feature learning'. In: *Knowledge-Based Systems* 221 (2021), p. 106892. ISSN: 09507051. DOI: 10.1016/j.knosys.2021.106892.

[2] Y. Gao, X. Li, X. V. Wang, L. Wang and L. Gao. 'A Review on Recent Advances in Vision-based Defect Recognition towards Industrial Intelligence'. In: *Journal of Manufacturing Systems* 62 (2022), pp. 753–766. ISSN: 02786125. DOI: 10.1016/j.jmsy.2021.05.008.

[3] J. Mirapeix, P. B. García-Allende, A. Cobo, O. M. Conde and J. M. López-Higuera. 'Real-time arc-welding defect detection and classification with principal component analysis and artificial neural networks'. In: *NDT & E International* 40 (2007), pp. 315–323. ISSN: 09638695. DOI: 10.1016/j.ndteint.2006.12.001.

[4] Y. F. Shu, B. Li, X. Li, C. Xiong, S. Cao and X. Y. Wen. 'Deep learning-based fast recognition of commutator surface defects'. In: *Measurement* 178 (2021), p. 109324. ISSN: 02632241. DOI: 10.1016/j.measurement.2021.109324.

[5] J. Park, H. Riaz, H. Kim and J. Kim. 'Advanced cover glass defect detection and classification based on multi-DNN model'. In: *Manufacturing Letters* 23 (2020), pp. 53–61. ISSN: 22138463. DOI: 10.1016/j.mfglet.2019.12.006.

[6] R. Garnett. *Bayesian Optimization.* Cambridge University Press, 2023. DOI: 10.1017/9781108348973.

[7] Bouthillier, X., Varoquaux, G. *Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020.* 2020. URL: https://hal.archives-ouvertes.fr/hal-02447823.

[8] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan and Q. V. Le. 'AutoAugment: Learning Augmentation Strategies From Data'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2019.

[9] S. Lim, I. Kim, T. Kim, C. Kim and S. Kim. 'Fast AutoAugment'. In: *Advances in Neural Information Processing Systems.* Vol. 32. Curran Associates, Inc, 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/6add07cf50424b14fdf649da87843d01-Paper.pdf.

[10] J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl. 'Algorithms for Hyper-Parameter Optimization'. In: *Advances in Neural Information Processing Systems.* Vol. 24. Curran Associates, Inc, 2011. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

[11] S. Watanabe. *Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance.* 2023. DOI: 10.48550/arXiv.2304.11127.

[12] D. R. Jones, M. Schonlau and W. J. Welch. 'Efficient Global Optimization of Expensive Black-Box Functions'. In: *Journal of Global Optimization* 13 (1998), pp. 455–492. ISSN: 09255001. DOI: 10.1023/A:1008306431147.

[13] N. Srinivas, A. Krause, S. Kakade and M. Seeger. 'Gaussian process optimization in the bandit setting: no regret and experimental design'. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning.* ICML'10. Madison, WI, USA: Omnipress, 2010, pp. 1015–1022.

[14] T. Bai, Y. Li, Y. Shen, X. Zhang, W. Zhang and B. Cui. *Transfer Learning for Bayesian Optimization: A Survey.* 2023. DOI: 10.48550/arXiv.2302.05927.

[15] X. Wang, Y. Jin, S. Schmitt and M. Olhofer. 'Recent Advances in Bayesian Optimization'. In: *ACM Computing Surveys* 55 (2023), pp. 1–36. ISSN: 0360-0300. DOI: 10.1145/3582078.

[16] M. Poloczek, J. Wang and P. I. Frazier. *Warm Starting Bayesian Optimization.* 2016. DOI: 10.48550/arXiv.1608.03585.

[17] T. Theckel Joy, S. Rana, S. Gupta and S. Venkatesh. 'A flexible transfer learning framework for Bayesian optimization with convergence guarantee'. In: *Expert Systems with Applications* 115 (2019), pp. 656–672. ISSN: 09574174. DOI: 10.1016/j.eswa.2018.08.023.

[18] D. Yogatama and G. Mann. 'Efficient Transfer Learning Method for Automatic Hyperparameter Tuning'. In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics.* Vol. 33. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR, 2014, pp. 1077–1085. URL: https://proceedings.mlr.press/v33/yogatama14.html.

[19] R. Bardenet, M. Brendel, B. Kégl and M. Sebag. 'Collaborative hyperparameter tuning'. In: *Proceedings of the 30th International Conference on Machine Learning.* Vol. 28. Proceedings of Machine Learning Research. Atlanta, Georgia, USA: PMLR, 2013, pp. 199–207. URL: https://proceedings.mlr.press/v28/bardenet13.html.

[20] M. Wistuba, N. Schilling and L. Schmidt-Thieme. 'Scalable Gaussian process-based transfer surrogates for hyperparameter optimization'. In: *Machine Learning* 107 (2018), pp. 43–78. ISSN: 0885-6125. DOI: 10.1007/s10994-017-5684-y.

[21] M. Wistuba, N. Schilling and L. Schmidt-Thieme. 'Two-Stage Transfer Surrogate Model for Automatic Hyperparameter Optimization'. In: *Machine Learning and Knowledge Discovery in Databases.* Vol. 9851. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 199–214. DOI: 10.1007/978-3-319-46128-1{\textunderscore}13.

[22] C. Hvarfner, D. Stoll, A. Souza, M. Lindauer, F. Hutter and L. Nardi. *BO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization.* 2022. DOI: `10.48550/arXiv.2204.11051`.

[23] A. Souza, L. Nardi, L. B. Oliveira, K. Olukotun, M. Lindauer and F. Hutter. *Bayesian Optimization with a Prior for the Optimum.* 2020. DOI: `10.48550/arXiv.2006.14608`.

[24] A. Ramachandran, S. Gupta, S. Rana, C. Li and S. Venkatesh. 'Incorporating expert prior in Bayesian optimisation via space warping'. In: *Knowledge-Based Systems* 195 (2020), p. 105663. ISSN: 09507051. DOI: `10.1016/j.knosys.2020.105663`.

[25] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter. 'Efficient and Robust Automated Machine Learning'. In: *Advances in Neural Information Processing Systems.* Vol. 28. Curran Associates, Inc, 2015. URL: `https://proceedings.neurips.cc/paper_files/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf`.

[26] M. Feurer, J. Springenberg and F. Hutter. 'Initializing Bayesian Hyperparameter Optimization via Meta-Learning'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29 (2015). ISSN: 2159-5399. DOI: `10.1609/aaai.v29i1.9354`.

[27] M. Wistuba, N. Schilling and L. Schmidt-Thieme. 'Hyperparameter Search Space Pruning – A New Component for Sequential Model-Based Hyperparameter Optimization'. In: *Machine Learning and Knowledge Discovery in Databases.* Vol. 9285. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 104–119. DOI: `10.1007/978-3-319-23525-7{\textunderscore}7`.

[28] V. Perrone, H. Shen, M. W. Seeger, C. Archambeau and R. Jenatton. 'Learning search spaces for Bayesian optimization: Another view of hyperparameter transfer learning'. In: *Advances in Neural Information Processing Systems.* Vol. 32. Curran Associates, Inc, 2019. URL: `https://proceedings.neurips.cc/paper_files/paper/2019/file/6ea3f1874b188558fafbab78e8c3a968-Paper.pdf`.

[29] A. Krizhevsky. 'Learning Multiple Layers of Features from Tiny Images'. In: (2009), pp. 32–33. URL: `https://www.cs.toronto.edu/%C2%A0kriz/learning-features-2009-TR.pdf`.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. 'Imagenet: A large-scale hierarchical image database'. In: *2009 IEEE conference on computer vision and pattern recognition.* 2009, pp. 248–255.

[31] D. Khatamsaz, R. Neuberger, A. M. Roy, S. H. Zadeh, R. Otis and R. Arróyave. 'A physics informed bayesian optimization approach for material design: application to NiTi shape memory alloys'. In: *npj Computational Materials* 9 (2023). DOI: `10.1038/s41524-023-01173-7`.

[32] F. Lundh. 'An introduction to tkinter'. In: *URL: www. pythonware. com/library/tkinter/introduction/index. htm* (1999).

[33] G. van Rossum and F. L. Drake. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace, 2009.

[34] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Polat, VanderPlas, Jake, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors. 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python'. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: `10.1038/s41592-019-0686-2`.

[35] J. Bergstra, D. Yamins and D. Cox. 'Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures'. In: *Proceedings of the 30th International Conference on Machine Learning.* Vol. 28. Proceedings of Machine Learning Research. Atlanta, Georgia, USA: PMLR, 2013, pp. 115–123. URL: `https://proceedings.mlr.press/v28/bergstra13.html`.

[36] J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl. 'Algorithms for Hyper-Parameter Optimization'. In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc, 2011. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

[37] J. Silvestre-Blanes, T. Albero-Albero, I. Miralles, R. Pérez-Llorens and J. Moreno. 'A Public Fabric Database for Defect Detection Methods and Results'. In: *Autex Research Journal* 19 (2019), pp. 363–374. DOI: 10.2478/aut-2019-0035.

[38] T. Hastie, R. Tibshirani and J. H. Friedman. *The elements of statistical learning: Data mining, inference, and prediction.* 2. ed., corr. at 4. print. Springer series in statistics. New York, NY: Springer, 2009.

[39] J. L. Loeppky, J. Sacks and W. J. Welch. 'Choosing the Sample Size of a Computer Experiment: A Practical Guide'. In: *Technometrics* 51 (2009), pp. 366–376. ISSN: 0040-1706. DOI: 10.1198/TECH.2009.08040.