Visual car brand classification by implementing a synthetic image dataset creation pipeline

Jan Lippemeier[1], Stefanie Hittmeyer[2], Oliver Niehörster[3], and Markus Lange-Hegermann[4]

[1] TH OWL University of Applied Sciences and Arts,
Campusallee 12, 32657 Lemgo
[2] Fraunhofer IOSB-INA,
Campusallee 1, 32657 Lemgo
[3] iplus1 GmbH,
Vogelsang 12, 33104 Paderborn
[4] Institute for Industrial Information Technology (inIT),
Campusallee 6, 32657 Lemgo

Abstract Recent advancements in machine learning, particularly in deep learning and object detection, have significantly improved performance in various tasks, including image classification and synthesis. However, challenges persist, particularly in acquiring labeled data that accurately represents specific use cases. In this work, we propose an automatic pipeline for generating synthetic image datasets using Stable Diffusion, an image synthesis model capable of producing highly realistic images. We leverage YOLOv8 for automatic bounding box detection and quality assessment of synthesized images. Our contributions include demonstrating the feasibility of training image classifiers solely on synthetic data, automating the image generation pipeline, and describing the computational requirements for our approach. We evaluate the usability of different modes of Stable Diffusion and achieve a classification accuracy of 75%.

Keywords Image synthesis, image classification, computer vision car brand classification, traffic monitoring, synthetic training data

DOI: 10.58895/ksp/1000174496-16 erschienen in:
Forum Bildverarbeitung 2024
DOI: 10.58895/ksp/1000174496 | https://www.ksp.kit.edu/site/books/m/10.58895/ksp/1000174496/

1 Introduction

In the last twelve years advancements in machine learning, deep learning and object detection achieved remarkable results in performance. The deep learning revolution in image classification started with the publication of AlexNet [1] in 2012. Further performance enhancements have been achieved in the following years with models such as ResNet [2]. Object detection has achieved significant success with models like YOLO (You Only Look Once) [3], which have produced high-accuracy results.

Transfer learning which uses pre-trained existing models is a common approach for solving image classification tasks [4]. Although this approach needs less data than training a model from ground up, it still requires large amounts of labeled data. Existing publicly available datasets can only be successfully used for training if they actually resemble the use case. Even if large amounts of data from the actual use case can be acquired, labeling this data remains time consuming and therefore expensive, as this is often a manual task. Biases within the data present a challenge as there is a compromise to be made, either in the form of potentially keeping the bias, reducing the dataset size to balance classes or oversampling underrepresented classes. In most cases this leads to small available datasets and consequently overfitting. Another common challenge is a low variance within the available data.

Solving computer vision tasks when only limited data is available, is a common major challenge in practice. Limited datasets often lead to overfitted or poorly performing models, endangering the success of a project. We face this challenge by illustrating an adaptable approach. For this approach we synthesize images on demand that are tailored to the respective use case. This leads us to posing our abstract main research question: Is it possible to use image synthesis in an automated manner to create suitable datasets for computer vision tasks with otherwise limited existing data? We evaluate this general approach on a specific real-world application.

Our real-world image classification task is to visually predict the brand of a car as an unequivocal visually determinable feature (see Figure 1). We recorded and labeled data that represents the German automotive traffic; this data was recorded by traffic cameras in Lemgo

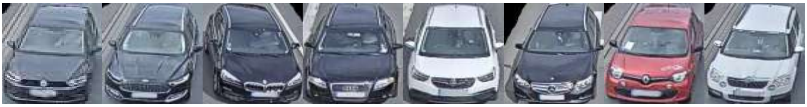


Figure 1: The selected brands we aim to classify. These eight brands occur the most in our recorded footage.

- a medium sized town in Germany. However, we are limited by the traffic volume and the capacities for human labeling. Even if unlimited gathering of real labeled data from the German traffic was possible, the data would still include the biases of the real world. Filtering these out and labeling the images would still remain a time consuming task. Existing related datasets such as the Stanford Car Dataset [5] tend to resemble the North American market. Some car brands common in Germany such as Skoda are normally not even present within existing datasets. With limited data from the actual application and no usable existing dataset this problem is a prime example for our main research question.

With the emergence of image synthesis models that create highly realistic images with correct proportions and details we propose the usage of synthetic images as training data for image classification tasks. In theory image synthesis models are a prompt-guided way to synthesize an image of a desired object. Stable Diffusion is an open source model that allows for programmatic image synthesis [6]. The creation of images therefore becomes a question of time and computing power.

We create an automatic pipeline for image dataset creation using Stable Diffusion as a tool to synthesize images. We are able to control the distribution of the generated images by controlling the distribution of the used prompts. By using YOLO we automatically determine the bounding boxes of a car inside a generated image. YOLO also allows us to estimate whether the synthetic image is suitable by giving a confidence score, the bounding box and the class of the detected object.

Although we can avoid class distribution bias by balancing the number of generated images per manufacturer, image synthesis models may still introduce inherent biases. If there are biases within the training data of the image synthesis model it might pass this bias on to the generated images. In regards to cars the data used in the training of Stable Diffusion could be unbalanced, for example, by favoring new over old, famous over unfamous, and popular over unpopular cars. Further it is not automatically confirmable whether a synthetic image actually matches the desired output encoded by the prompt. Also the perspective, illumination, contrast and other photographic properties might differ from the actual task.

Our main contributions are: We automate an image generation pipeline that also includes labels, bounding boxes and a quality assessment for the synthetic images. We show that training on purely synthetic data from our image generation pipeline is sufficient to train an image classifier that can visually predict the car brand on a real photograph. We describe the required amount of and necessary computation time for synthetic images in our use case. We include and compare different modes of Stable Diffusion for synthesizing images in our pipeline.

2 Related Work

Large-scale text-to-image diffusion models can be fine-tuned to augment the ImageNet training set [7] leading to significant improvements in ImageNet classification accuracy [8]. Moreover, the authors of [9] investigate using synthetic images produced with Stable Diffusion [6] when training models for ImageNet classification. Whether and how synthetic images generated from text-to-image generation models can be used for image classification in data-scarce settings and in large-scale model pre-training for transfer learning is considered in [10] using the GLIDE diffusion model [11].

Synthetic data has been successfully used to improve identification and classification tasks in other applications such as lung edema identification in chest X-ray images [12] and the diagnosis of skin diseases [13]. In the latter two references Stable Diffusion was used to generate the corresponding synthetic image datasets. Introducing synthetic test data has been proposed as a means to improve model evaluation on diverse and underrepresented population subgroups [14].

In the field of vehicle type classification, or, more specifically, car brand and model identification, mainly models trained on real-world data have been investigated so far. Examples are the extension of models trained on limited-size datasets to handle extreme lighting conditions [15], balanced sampling to address the challenge of classifying imbalanced data from visual traffic surveillance sensors [16], improving accuracy of car type classification through the adaptation of specific CNN architecture models [17], as well as adapting deep learning techniques for vehicle color classification [18] and vehicle logo recognition [19]. The detection, recognition, and counting of vehicles based on their car types using a combination of YOLOv5 and ResNet has been investigated in [20].

3 Method

The goal of this work is to develop a pipeline (illustrated in Figure 2), which can generate a balanced dataset for a computer vision task with otherwise limited available labeled data.

I Dataset Input For this work we used the official car registrations [21] from the Federal Motor Transport Authority of Germany (Kraftfahrt-Bundesamt). They provide data for registered car models in Germany. The features are the vehicle class, the brand, the model name, the build years and the registered number of cars in each category. An example for a car model is the Skoda Karoq, a SUV of the brand Skoda produced in the years 2017, 2018 and 2020. However we ruled out production years earlier than 1990 as they rarely occur in everyday traffic.

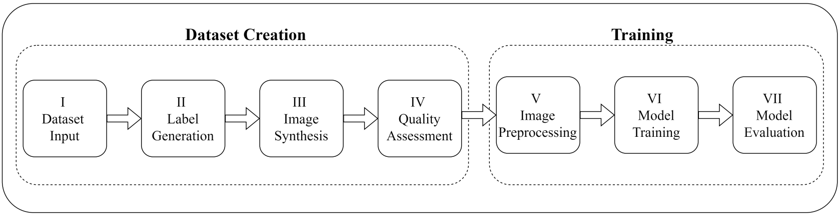


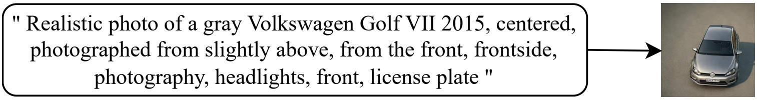
Figure 2: The scheme of the developed pipeline consisting of dataset creation and training. The illustrated pipeline produces a dataset of synthetic images with corresponding labels and bounding boxes. The dataset can be used to automatically train an image classification model.

While it might be possible to generate suitable data even for rare brands we limit our approach to the brands shown in Figure 1. By focusing on these brands we aim to include the often occurring brands such as Volkswagen and Ford but also more rarely occurring brands such as Skoda and Renault.

II Label Generation In order to balance the dataset we use a multi-step hierarchical uniform probability distribution. In the first hierarchy step each brand has the same probability. For each brand the probability of the respective car models is also uniformly distributed. The same principle applies to the construction years for each model as well as the most common colors.

III Image Synthesis In this pipeline step we sample from the labels created in the previous step and create a prompt for each sample as shown in Figure 3. We create two datasets, one for Text-to-Image and one for Image-to-Image using Stable Diffusion XL Turbo. If not otherwise noted we use the standard parameters set by the Python Diffusers library [22]. For Text-to-Image we use four inference steps with one image per prompt and a guidance scale of zero. This guidance scale is recommended in the documentation for the usage of this model [23]. For Image-to-Image we use ten inference step with a guidance scale of 0.4 and a strength of 0.6. These parameters are manually tuned to subjectively fit the desired output.

When using Stable Diffusion in Image-to-Image mode we also have to provide a base image in conjunction with a prompt as the input for the model. To create these base images we use real photographs of cars at different positions on the road cropped to the car with padding. With these base images we intend to implicitly give the desired perspective so that the generated images strongly resemble the real images. The input base image for this mode is scaled up to 720x720 pixels



" Realistic photo of a gray Volkswagen Golf VII 2015, centered, photographed from slightly above, from the front, frontside, photography, headlights, front, license plate "

Figure 3: The prompt used to generate the images with Stable Diffusion alongside a generated image using Text-to-Image. The substring *"gray Volkswagen Golf VII 2015"* is changed accordingly for different car models.

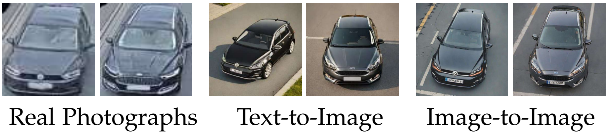


Figure 4: Illustration of differences between real images and modes of image generation. Text-to-Image tends to encompass more perspectives contrary to the narrow range of perspectives with Image-to-Image.

as this is the minimal size for Stable Diffusion XL. Figure 4 illustrates real photographs compared to images generated by Text-to-Image and Image-to-Image.

IV Quality Assessment The output of image synthesis models such as Stable Diffusion normally matches the expected output. In most of the cases there is exactly one car as the main subject of the image. The location and the size of the image's subject differs. Therefore an object detection model such as YOLOv8x has to be used to automatically determine the bounding boxes (see Figure 5). However, we observe that in rare cases Stable Diffusion produces something other than the desired output of a singular car (see Figure 5). Therefore, we use YOLOv8x in object detection mode on each image. This allows us to automatically confirm that there is exactly one car in the image. It further assesses the quality of the generated image as YOLO provides a score for the certainty of a detected bounding box. We then crop the image to the bounding boxes of the detected car.

V Image Preprocessing Stable Diffusion models allow to specify the dimensions of the resulting image. However, the subject of the images vary in size so that the images cropped to their subjects' bounding boxes differ in aspect ratios and sizes. We transformed the images to

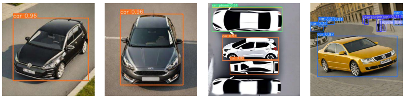


Figure 5: Bounding Boxes detected with YOLOv8x. This allows to crop the image and provides a confidence score for the presence of a car. It also allows to automatically sort out the two undesired images on the right where more than one car is detected.

64x64 pixels. The small resolution is chosen as the traffic cameras in our use case record in HD and cars at different positions in the image may therefore have a similar resolution. Random Rotation as a classical data augmentation method is also applied to the dataset.

VI Model Training We then use the generated and pre-processed image dataset to train image classifiers. In using the model Resnet-18 that is pre-trained on ImageNet [7], we apply the principle of transfer learning [4]. For adapting this model we replace the last fully connected layer by a new fully connected layer with the same input size and an output size of eight which encodes the classes we want to classify. We do not lock any pretrained layer.

VII Evaluation The performance of the model is validated against real-world data recorded by traffic cameras mounted in Lemgo, Germany. These images were manually labeled. Images of the same car at different positions can exist in the datasets as the cars are moving forward. The split between validation and test dataset is therefore performed based on location of the respective camera and the time of recording. Due to the biases in the real world and the described split the classes are unevenly distributed.

4 Experiments and Results

The time for generating images with Stable Diffusion depends on output size and the number of inference steps. In the following we consider the performance for the parameters described in Section 3 (Method) when run on a Nvidia RTX 3060. When using Text-to-Image (four inference steps) it took 0.85 seconds, for Image-to-Image (ten inference steps) 2.33 seconds. These durations account for image synthesis, bounding box detection, automatic quality assessment and storing of the image.

We retrain Resnet-18 on varying datasets. The model is trained on the images in random order with an exponentially decaying learning rate starting at 0.01 and the stochastic gradient descent optimizer. An epoch for training the Resnet-18 on 100.000 images takes approximately 18 seconds on a Nvidia RTX 3060.

We evaluate the performance of this model regarding the different modes of image synthesis and the required amount of data. As we

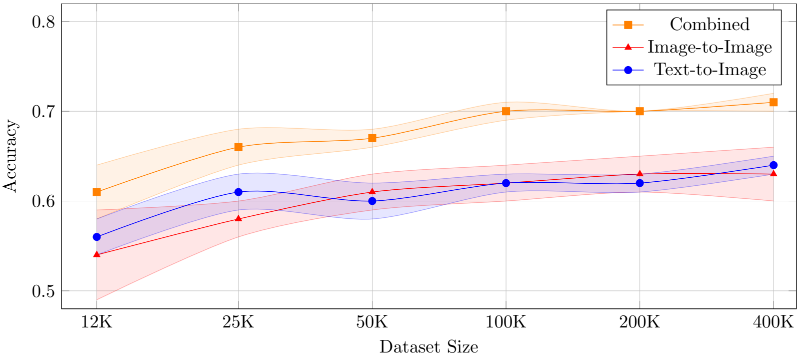


Figure 6: Training Results for Resnet-18 trained on the different dataset sizes. Model performance and dataset size correlate. The experiments were performed five times per dataset size and Stable Diffusion mode. The graph shows a confidence interval of one standard deviation.

use real world photographs of cars in traffic in Lemgo, Germany, we have an unbalanced dataset. To illustrate the unbalanced distribution: VW occurs the most with 976 images in contrast to Renault with only 165 images. The split, described earlier, results in a validation dataset consisting of 1503 images and a test dataset consisting of 1317 images.

Figure 6 shows the training results for different dataset sizes and datasets generated by Image-to-Image mode, Text-to-Image mode and a combination of these modes. These results show that we are able to train a model on the generated images that can exceed the primitive baseline by far, even reaching up to 75% in accuracy on the given eight classes.

We observe that the performance of the retrained Resnet-18 model has a reliable and stable performance on both modes of image generation and both modes combined. The performance of this model in regards to the mode of image generation does not differ significantly. The Resnet-18 model clearly benefits from using images from both modes combined leading to a score of 75% at maximum with a dataset size of 400.000 images. We can see that the model performance correlates with the dataset size and the variety in images, as the performance when trained on images from both modes combined surpasses the performance when trained only on images from one mode.

The retrained models are typically biased towards predicting Volkswagen which is the most common brand in the photographs of the real traffic. We also observe a difference in accuracy per brand. Volkswagen, Ford, BMW, Audi and Mercedes achieve a performance of at least 70% while Opel, Renault and Skoda are only correct in about half of the cases.

5 Conclusion

We are able to train image classifiers for real world data solely on synthetic images that require no human labeling. The images we evaluate the classifiers on are taken in real moving traffic. Therefore, we face challenges such as a large variety of objects, image artifacts, different lighting, low resolution and motion blur. The generated data is sufficient to exceed the primitive baseline by far. These results are achieved whilst needing human work only for engineering the pipeline, tweaking hyperparameters and labeling the validation and test images. Thus the engineered pipeline may illustrate a potential approach to overcome challenges associated with traditional data acquisition methods.

On average the Resnet-18 performs better when retrained on the combined images instead of the same amount from just one mode of image generation. This may result from the fact that both modes combined cover a broader variety regarding the characteristics of the images. We assume that using varying prompts and varying parameters for the Stable Diffusion model could increase the variation in images and could therefore be beneficial.

To illustrate one possibility of this pipeline: With our pipeline we are able to create a perfectly balanced dataset of 100.000 images by using both modes combined. We can directly train a Resnet-18 model on this generated dataset. The time to perform this consecutively on a singular Nvidia RTX 3060 without further optimizations sums up to about two days. This provides a solid baseline model on short term with very little human work required.

There are significant differences in performance of the retrained Resnet-18 per class. These differences may be explained by biases inside of the Stable Diffusion models as they are more likely trained on more images of Volkswagen than images of Skoda. Another may be that, on one hand, brands like Volkswagen, Mercedes, BMW and Audi have very prominent visual features that are easy recognizable for humans. On the other hand, earlier models of Renault have very small logos and Skoda has a dark radiator grill with only a small logo as an identifier. This also can attribute to a lower performance for these brands.

The pipeline introduced in this work is possible as we can automatically assess and crop the output of the Stable Diffusion model with YOLO. For other computer vision tasks with classes that are a subset of the classes YOLO can predict, we can adapt the pipeline easily. However, for completely other classes one would have to engineer another way to implement the fourth step of the pipeline. As this presents a challenge, the possible use cases of the introduced pipeline are limited by the capabilities of object detection models like YOLO. Another limitation lies within the capability of Stable Diffusion as it is unlikely that these models can generate usable images for every situation.

Acknowledgments

Markus Lange-Hegermann acknowledges support from the German Federal Ministry of Education and Research (BMBF) for the project SyDaPro with grant number 01IS21066A.

References

1. A. Krizhevsky *et al.*, “ImageNet classification with deep convolutional neural networks,” in *NeurIPS*, vol. 25, 2012.
2. K. He *et al.*, “Deep residual learning for image recognition,” 2015.
3. J. Redmon *et al.*, “You only look once: Unified, real-time object detection,” 2016.
4. F. Zhuang *et al.*, “A comprehensive survey on transfer learning,” *Proceedings of the IEEE*, vol. 109, no. 1, 2021.
5. J. Krause *et al.*, “3d object representations for fine-grained categorization,” in *IEEE ICCVW*, 2013.
6. R. Rombach *et al.*, “High-resolution image synthesis with latent diffusion models,” 2022.

7. J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009.
8. S. Azizi *et al.*, "Synthetic data from diffusion models improves imagenet classification," 2023.
9. M. B. Sariyildiz *et al.*, "Fake it till you make it: Learning transferable representations from synthetic imagenet clones," in *CVPR*, 2023.
10. R. He *et al.*, "Is synthetic data from generative models ready for image recognition?" *ICLR, spotlight*, 2023.
11. A. Nichol *et al.*, "GLIDE: towards photorealistic image generation and editing with text-guided diffusion models," *CoRR*, vol. abs/2112.10741, 2021.
12. Z. Liang *et al.*, "Covid-19 pneumonia chest x-ray pattern synthesis by Stable Diffusion," in *IEEE SSIAI*, 2024.
13. P. Patcharapimpisut and P. Khanarsa, "Generating synthetic images using Stable Diffusion model for skin lesion classification," in *16th KST*, 2024.
14. B. van Breugel *et al.*, "Can you rely on your model evaluation? improving model evaluation with synthetic test data," in *NeurIPS*, vol. 36, 2023.
15. Y. Zhou *et al.*, "Image-based vehicle analysis using deep neural network: A systematic study," in *IEEE ICDSP*, 2016.
16. W. Liu *et al.*, "An ensemble deep learning method for vehicle type classification on visual traffic surveillance sensors," *IEEE Access*, vol. 5, 2017.
17. M. Taqiyuddin *et al.*, "Accuracy improvement of cnn mobilenet-v1 and residual network 50 layers models using adam setting for car type classification," in *ISESD*, 2022.
18. J. Kim, "A study on the trend of vehicle types and color classification technology for intelligent transportation systems," in *IEEE ICCE-Asia*, 2021.
19. W. Lu *et al.*, "Category-consistent deep network learning for accurate vehicle logo recognition," *Neurocomputing*, vol. 463, 2021.
20. A. S. Rao *et al.*, "Identification of car make and model using deep learning and computer vision techniques," in *AIDE*, 2022.
21. Kraftfahrt-Bundesamt, "Official registrations of cars by segments and models (fz12, fz2, sv 4.2 - 2023)," 2023.
22. P. von Platen *et al.*, "Diffusers: State-of-the-art diffusion models," https://github.com/huggingface/diffusers, 2022.
23. Huggingface, "Using Diffusers Stable Diffusion XL Turbo," https://huggingface.co/docs/diffusers/using-diffusers/sdxl_turbo, 2023.