2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Synset Boulevard: A Synthetic Image Dataset for VMMR*

Anne Sielemann[1], Stefan Wolf[2,1], Masoud Roschani[1], Jens Ziehn[1] and Juergen Beyerer[1,2]

***Abstract*— We present and discuss the Synset Boulevard dataset, designed for the task of surveillance-nature vehicle make and model recognition (VMMR)—to the best of our knowledge the first entirely synthetically generated large-scale VMMR image dataset. Through the simulation of image data rather than the manual annotation of real data, we intend to mitigate common challenges in state-of-the-art VMMR datasets, namely bias, human error, privacy, and the challenge of providing systematic updates. On the other hand, the provision and use of synthetic data introduce individual challenges, such as potential domain gaps and a less pronounced intra-class variance. Our approach to address these challenges, using path tracing and physically-based, data-driven models, is evaluated on an existing large real-world dataset. Overall, our synthetic dataset contains 32 400 independent images (each with different imaging simulations and with/without masked license plates, leading to a total of 259 200 images) from 162 different vehicle models of 43 makes depicted in front view. It is split into 8 sub-datasets to investigate the influence of optical/imaging effects on the classification ability.**

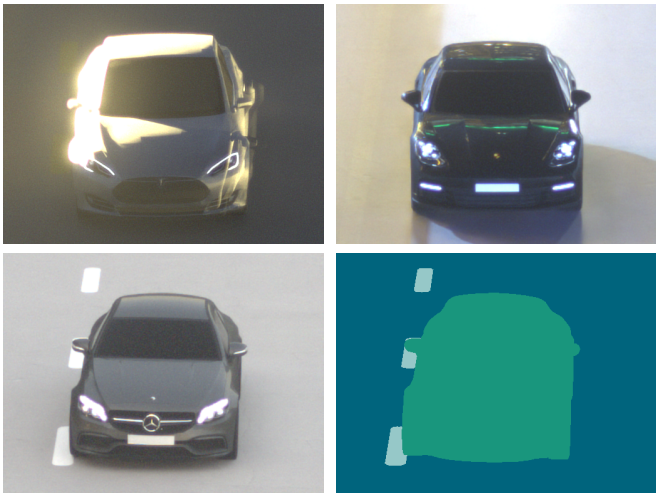


Fig. 1: Example images and one label image from the Synset Blvd. dataset.

I. INTRODUCTION AND MOTIVATION

THE task of "vehicle make and model recognition" (VMMR) has a wide range of applications, from traffic analysis to police surveillance. It involves classifying a given image of a vehicle into a fine-grained class hierarchy including the make (e.g., "Audi"), the model (e.g., "A8"), and possibly the model year (e.g., "2016"). Implementations may further include extracting features such as vehicle color.

Use cases include analyzing the traffic share of vehicles on given roads e.g., for intelligent transportation systems (ITS), traffic management systems (TMS), or the identification of vehicles within automated vehicle surveillance (AVS), or electronic toll collection (ETC). The latter applications usually primarily rely on license plate recognition; however, this recognition can be inhibited either by unintentional effects (e.g., stains on the plate) or, commonly in organized crime, by deliberately manipulating or hiding the license plates. Here, VMMR can considerably support the identification of vehicles as described by Pan, Zhou, Zhou, *et al.* [1].

Besides conventional image processing challenges, such as variations in light and weather, occlusions, shadows, and reflections, VMMR involves the particular challenge arising from a large class set of vehicle types, with low inter-class and high intra-class variance. In addition, new vehicle models are constantly being released and old models are modernized through facelifts, causing VMMR datasets to go out of date within few years if they are not maintained at great expense.

Creating training data for VMMR is a time-consuming, laborious, and expensive process: Manual labeling of image data requires expert knowledge and is still prone to errors and inaccuracies, affecting the quality of results [2]. When acquiring the data, it is difficult to ensure balanced classes and representative variance. This can challenge the certification of AI-based systems and services, e.g., in the European Union, where the proposed "AI Act" [3] introduces high requirements for high-risk AI systems (including the domains of law enforcement and operation of road traffic, cf. [3, annex III 2a, 6f]), such as the requirement that "Training, validation and testing datasets should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system". Furthermore, the generation of datasets is subject to privacy requirements, such as the European General Data Protection Regulation (GDPR) [4] limiting not only the processing and redistribution but also already the recording and storage of any data containing privacy-critical information.

The synthetic generation of data can, in principle, mitigate these challenges, by reducing acquisition costs and manual effort for the provision of ground truth, by enabling parametric, controlled variation and thus a means of systematic balancing, and by avoiding privacy issues. It can also directly

Download Synset Blvd.: synset.de/datasets/synset-blvd/

[1]Fraunhofer IOSB, 76131 Karlsruhe, Germany, {anne.sielemann, stefan.wolf, masoud.roschani, jens.ziehn}@iosb.fraunhofer.de
[2]Karlsruhe Institute of Technology (KIT), Vision and Fusion Laboratory (IES), 76131 Karlsruhe, Germany

*This work was supported by the Fraunhofer Internal Programs under Grant No. PREPARE 40-02702 within the "ML4Safety" project, and by the Ministry of Economic Affairs, Labour and Housing of the state of Baden-Wuerttemberg, Germany, as part of the "FeinSyn" research project.

TABLE I
MOST RELEVANT PUBLICLY AVAILABLE WEB-NATURE VMMR DATASETS.

Dataset	# Samples	# Classes	Perspective
Cars-196 [11]	16 185	196	mixed
CompCars [12]	136 726	1,716	mixed
VMMRdb [13]	291 752	9 170	mixed
Frontal-103 [14]	65 433	1 759	front
DeepCar 5.0 [15]	40 185	480	front & 3/4

TABLE II
MOST RELEVANT PUBLICLY AVAILABLE SURVEILLANCE-NATURE VMMR DATASETS.

Dataset	# Samples	# Classes	Persp.
CompCars [12]	50 000	281	front
BoxCars21k [16]	63 750	148	mixed
BoxCars116k [17]	116 286	693	mixed
Synset Blvd.	32 400	162	front

provide a methodology for expanding and updating the dataset, by generating new data once new vehicle models and corresponding 3D models are available.

The precondition for successfully leveraging these advantages is resolving a key challenge in simulation: The "domain gap", describing a systematic and significant deviation between the distribution of two data sources, here the simulation vs. the real world. This gap, also called "sim-to-real gap", is known to impact the results of machine learning and specifically transfer learning [5] significantly. In general, its resolution is considered challenging and is the subject of active research. Generic approaches to bridge this gap are domain randomization (DR) [6] and structured DR (SDR) [7], used for the creation of several synthetic vehicle detection datasets [8]–[10]. The presented dataset is intended to contribute to understanding the effect of the "sim-to-real" domain gap and evaluating potentials for limiting its effects.

II. STATE OF THE ART

A. VMMR Datasets

In the field of VMMR, numerous datasets exist, most of which, however, are of relatively small scale. This section focuses only on publicly available VMMR datasets of a scale that is comparable with the presented dataset, as these predominantly serve as benchmarks for upcoming approaches.

Existing datasets can be categorized into *web-nature* (Tab. I) and *surveillance-nature* (Tab. II). Web-nature datasets include web images as published by manufacturers and vendors complete with make and model information. In general, this category comprises a greater number of large-scale datasets, presumably because they are easier to label and acquire. Surveillance-nature images are captured in public traffic, usually by traffic cameras. Here, only two common publicly available large-scale datasets are known.

For either category, existing datasets are difficult to compare, interchange, or combine, due to the different recording perspectives, styles, and a limited overlap in classes. Sánchez, Parra, *et al.* [2] analyzed the Cars-196 [11], CompCars [12], BoxCars21k [16], VMMRdb [13], and Frontal-103 [14] datasets with regard to strengths but also possible limitations and biases. These considerations were taken as a significant basis for the design of the presented dataset.

For all analyzed datasets, a geographical bias was determined. For instance, the CompCars [12] and Frontal-103 [14] datasets captured in China include a large percentage of Asian vehicle models. The entire BoxCars21k [16] dataset was recorded in one city in the Czech Republic. Other problems stated by Sánchez, Parra, *et al.* [2] are mislabeling of vehicles, the combination of variants and facelifts into a single class, or the class imbalance problem, leading to a poor classification performance of underrepresented classes. Web-nature datasets suffer less from under-representation, but instead from the domain gap between professional images and actual public traffic scenes. These effects were found to limit the performance in practical applications.

With their experiments on a cross dataset combining samples from the CompCars [12], VMMRdb [13], and Frontal-103 [14] datasets, Sánchez, Parra, *et al.* [2] showed that the classification performance degrades considerably (for CompCars up to 59 %) when evaluating on data of differing domains, for instance, conditioned by more challenging data or data of another dataset. They conclude that although state-of-the-art results achieve over 95 % accuracy on available VMMR datasets, there is still a lot of work to do for unbiased VMMR in realistic traffic and driving scenarios. This is where we see great potential for improvement by using synthetic datasets.

B. Synthetic Vehicle Datasets

To the best of our knowledge, there exists no entirely synthetic VMMR dataset so far—in contrast to the tasks of vehicle detection or semantic segmentation, where the resulting performance of deep learning (DL) approaches was already successfully increased by (additionally) using synthetic datasets for training [10], [18]–[21]. Tab. III provides an overview of some popular synthetic vehicle datasets ordered by year of publication. All listed datasets were generated by using real-time (i.e., rasterization-based) rendering approaches. One possibility is to make use of a computer game [18], [19]. Modern games provide large, open worlds with a high level of detail. However, label generation is challenging, since labels are not natively provided, but have to be reconstructed from the communication between the computer game and graphics hardware. Another disadvantage is the extensibility of game datasets, which are typically limited to objects present in the original game. Therefore, other approaches directly adopt the underlying game *engines*, such as Unity [22] or Unreal Engine [10], [20], [21], [23]–[25], and thereby gain flexibility in exchange for increased effort in building complex worlds.

III. DATASET GENERATION

The presented dataset is generated by the Fraunhofer simulation platform OCTANE [26], which provides a mod-

TABLE III
POPULAR SYNTHETIC VEHICLE DATASETS.

Dataset	Tool	Main Purpose	# Samples
[18]	GTA5	Sem. Segmentation	24 966
[19]	GTA5	Object Detection	200 000
SYNTHIA [20]	Unity	Sem. Segmentation	200 000
vKITTI [24]	Unity	Object Detection	21 260
VehicleX [21]	Unity	Vehicle Re-ID	—
vKITTI2 [25]	Unity	Object Detection	42 520
SAVED [10]	Unreal	Vehicle Part Recog.	586 340
Synset Blvd.	OCTANE	VMMR	32 400

TABLE IV
RESULTS OF THE CONDUCTED ABLATION STUDY.

Type of Variation	F1	Diff.
Base Dataset	74.9	—
Reduced Number of Samples (50)	64.49	−10.41
Reduced Number of Samples (20)	51.46	−23.44
No Environment Variation	59.48	−15.42
No Denoising	70.63	−4.27
DR 1 (Background & Distractors)	70.03	−4.87
DR 2 (Background, Effects & Distractors)	75.51	+0.61
SDR (Background & Effects)	76.62	+1.72

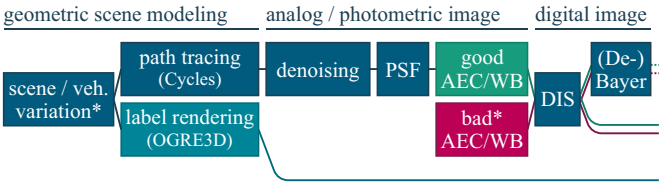


Fig. 2: Overview of the synthetization steps discussed in Sec. III, in particular Sec. III-D, ordered by the domain of operations. Steps marked with an asterisk (*) contain significant modeled variations.

ular, plugin-based architecture written in C++, allowing to extend functionality at runtime. Two largely interchangeable rendering plugins are available, one based on the real-time rasterization graphics engine OGRE3D [27], the other based on the physically-based path tracer Cycles [28]. The Cycles render engine was developed by the Blender project and implements an unidirectional path tracing algorithm with multiple importance sampling. The image rendering was performed exclusively using the Cycles plugin i.e., path tracing, while the OGRE3D plugin to OCTANE was used to generate texture-level label images including road markings.

Modeling in OCTANE enables a direct stochastic approach of scenario generation; hence, the variations described in Sec. III-B et seqq. provide a full overview of the main steps in defining the parameters to generate the dataset.

A. Preliminary Examinations

To investigate the influence of synthetization properties such as render settings, environment, and vehicle variation on the classification ability, we conducted an ablation study on small sub-datasets consisting of 49 vehicle models with 100 images per class (200 in case of DR and SDR) before generating the Synset Blvd. dataset. For the experiments, we utilized the ResNet50 [29] backbone of the OpenMMLab Classification Toolbox [30], pretrained on the ImageNet1k [31] dataset. The resulting network was evaluated on six overlapping classes[1] of the CompCars dataset [12]. Note that only daytime images of the CompCars dataset were used since solely daytime images are included in the Synset Blvd. dataset. The results are given in Tab. IV. Therein, the specified differences refer to the F1 score of the base dataset.

We drew the following conclusions from this study:

- The Synset Blvd. dataset should be rendered with 100 samples per image (like the base dataset) since a reduction to 20 or 50 samples leads to performance loss.
- A lack of environment variation significantly reduces the resulting F1 score. Therefore, a realistic environment modulation is of great importance.
- Lack of denoising had a negative effect (albeit small). Although some authors [32] describe a positive impact (i.e., a higher robustness) of using noisy images for training, we decided to apply denoising on the raw Synset Blvd. dataset renders, and add model-based synthetic camera noise subsequently.
- DR [6] respectively SDR [7] holds potential for future improvements; particularly, a realistic scene variation typically outperforms purely random augmentations.

[1]Mercedes S-Class, Mercedes C-Class, Porsche Panamera, Skoda Rapid, VW CC, and Porsche Cayenne

B. Environment Variation

In the actual Synset Blvd. dataset, the visible scene contains a straight road segment (uniformly distributed either with or without a concrete barrier) with procedural texture variations equivalent to 2.4 km of road length using textures of texturelib.com [33] as a basis. Road markings were added randomly and include single solid and dashed, white and yellow lines of random proportions, along with ground truth labels (cf. Fig. 1), such that a total of 150 different road surfaces occur uniformly across the dataset, both as wet and dry surfaces. Environment lighting uses image-based lighting (IBL) based on 183 environment maps, collected from Polyhaven [34], which are sampled uniformly, and whose azimuth is varied uniformly.

C. Vehicle Variation

1) Model: The main part of the scene is the particular vehicle which is depicted in the center of the images. For the generation of this dataset, we have purchased a total of 162 vehicle models, predominantly from DOSCH DESIGN [35]; one model stems from cgtrader [36]. We attached importance to two conditions for the selection of vehicle models: On one hand, we paid attention that our selection includes vehicles from diverse continents to counteract the geographical bias. Therefore, we selected sets of 3D vehicle models common in

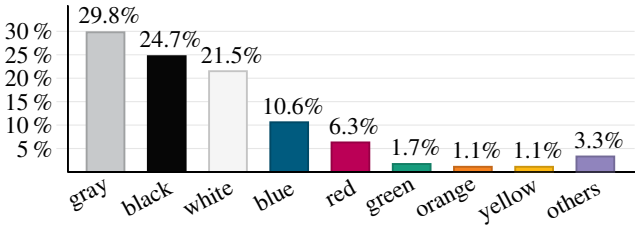


Fig. 3: A statistic from the German Federal Motor Transport Authority (KBA) [37] showing new vehicle registrations by color in 2021. It is based on 2 622 132 newly registered cars.

Europe, USA, and Asia. On the other hand, we aimed for a notable overlap of classes with the CompCars dataset [12]—the dataset we chose for our evaluation—to get meaningful evaluation results. In contrast to the available real-world datasets, our Synset Blvd. dataset offers for each class the same number of images, 200 per vehicle, i.e., the dataset is balanced to equal class representation.

2) Car Paint: We distributed the car paint color according to a statistic of the German Federal Motor Transport Authority (Kraftfahrt-Bundesamt, KBA) [37] that quantitatively describes the colors of newly registered vehicles based on 2 622 132 registrations in 2021 as visualized in Fig. 3. Within a selected color we nuance the shade to cover more variation.

3) Pose: Vehicle positions vary across the entire width of the road (with the camera always centered on the vehicle), while its orientation (roll, pitch yaw) is normally distributed by ($\sigma_{\text{roll}} = 0.7°$, $\sigma_{\text{pitch}} = 0.3°$, $\sigma_{\text{yaw}} = 1.3°$) respectively. Final images are cropped to a randomly padded bounding box around the vehicle to match the established training dataset format as used in CompCars for example.

4) Lights: The vehicle lights are equally likely either completely on or off[2], but no distinction is made concerning specific light functions due to a lack of available data.

D. Optical and Imaging Effects

The path tracing simulates ideal light transport per camera pixel (containing only noise from the render sampling, which is denoised directly through the Nvidia AI Denoiser [38]). Actual camera images contain a variety of effects from light transport that are impractical to consider during path tracing, as well as effects from the digital imaging process. We distinguish the quality levels into *good* and *bad*, each with and without simulated Bayer demosaicing, leading to four levels of quality (cf. Fig. 4c) per path tracing result.

Prominently, a *point spread function* (PSF) describes a convolutional effect per object ray caused by focusing, lens optics, diffraction, and may also include scattering in the atmosphere and on the lens. The channel-dependent parameters are approximated as a mixture-of-Gaussian model, based on a Tamron M112FM35 35 mm lens (but taken for the simulated focal length of 50 mm) used for the real-world reference dataset acquisition, and is identical across all levels of quality in the dataset (cf. Fig. 4b). The *noise* levels depend on the overall scene intensity (with lower light leading to stronger noise), with low and constant overall levels for the *good* quality, and random, high levels for the *bad* quality. Additionally, the *bad* quality introduces lens flare effects assuming an approximately centered optical axis. No image distortions were simulated due to the narrow field of view.

Automatic exposure control (AEC) and *white balancing* (WB) affect the digital image brightness and tint. For the *good* quality, AEC and white balance settings are optimal (white balance according to the gray world assumption) w.r.t. the visible frame. For the *bad* quality, both AEC and white balance deviate randomly.

After digitalization, *digital image sharpening* (DIS) is applied as a 3×3 highpass kernel uniformly across all qualities. For the *Bayer* quality variants, a final step introduces artifacts from simple Bayer BGGR bilinear demosaicing.

E. Structure and Annotations

The dataset is subdivided into the four different simulated imaging qualities regular/Bayer good/bad (cf. Fig. 4), each with and without masked license plates (MLP), within which 162 vehicle classes, each with 200 samples, are placed.

Each vehicle sample contains one simulated camera image cropped to the approximate bounding box, one corresponding semantic segmentation image (cf. Fig. 1), along with car paint color (as category and RGB), car paint metalness, approximate time of day, and the road condition (wet, dry).

Per vehicle class, we specify the make, model, model year(s) as far as known today, and the number of doors, predominantly based on data from the ADAC database [39].

For experiments with masked license plates, we trained a YOLOX [40] model on the Car Plate Detection [41] dataset for 60 epochs. Afterward, we evaluated it on the images of Synset Blvd. and filled the area of all bounding boxes with a confidence greater than 0.2 with the normalization mean used for training our classification networks, i.e., a gray tone.

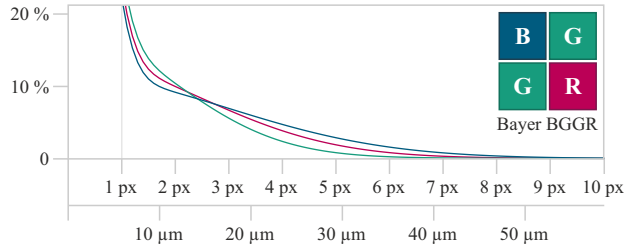
IV. EVALUATION

A. Training Setup

For the evaluations, we employ a ConvNeXt-Small [42] classification network of the OpenMMLab Classification Toolbox [30]. The model is trained on the full train set of the Synset Blvd. dataset for 100 epochs to classify all 162 classes. We apply an AdamW optimizer, an initial learning rate of 10^{-2} with a cosine decay towards 10^{-6}, a weight decay of 10^{-2}, and a batch size of 64. We load backbone weights trained on ImageNet1k [31] as pre-training. To reduce overfitting, we freeze the first two stages of the ConvNeXt backbone and apply label smoothing with a value of 0.1 and exponential moving average with a momentum of 10^{-4}. During training, a randomly resized crop, a random horizontal flip, RandAugment [43], and random erasing are applied as data augmentation. During inference, the image is rescaled to 256 pixels on the shorter side preserving the image ratio and a center crop to 224×224 pixels is performed. We train each configuration with three different seeds and report the mean and the standard deviation of the three runs.

[2]Following a manual annotation of the CCSV Audi Q5 with 69 samples with low beam, and 110 with daytime running lights, out of 216 total.

(a) Sample images from real dataset acquired with a Basler ace acA1920-40gc camera (Sony IMX249) and a Tamron M112FM35 35 mm lens.



(b) Used point spread function approximated based on the Tamron M112FM35 (while dataset images, as below in (c), use a 50 mm lens to achieve a resolution comparable to CompCars) and Bayer pattern on the Sony IMX249. The functions observe energy conservation; the amplitude is indicated w.r.t. the green peak. Values in µm refer to the Sony IMX249.

(c) Three of the four qualities (good, Bayer not shown) computed per geometric path tracing result. The "good" quality set has optimal balance and low noise levels. Bad introduces AEC and white balance deviations, lens flares and higher noise levels. Bayer introduces demosaicing effects from a BGGR Bayer pattern (visible, e.g., in the headlight fringes).

Fig. 4: Overview of optical and imaging effects applied after path tracing and subsequent denoising of render sample noise, cf. Sec. III-D.

B. Testing on CompCars

We evaluate the trained network on the daytime images of the test set of the CompCars Surveillance [12] (CCSV) dataset. To investigate the impact of the vehicle model differences, we select subsets of the classes for the evaluation that match the classes of Synset Blvd. by different degrees of granularity. First, we pick all classes of CCSV with matching make and model. Afterward, we increase the class specificity step-by-step by considering additional attributes like the model year, the facelift version, and whether the images contain country-specific versions. For the model year, we consider the initial year of generation without considering facelifts. We aggregate the predictions of models for which multiple versions are present in Synset Blvd., e.g., with different model years, since CCSV also refrains from this distinction. The results are shown in Tab. V.

TABLE V

EVALUATION OF CONVNEXT-SMALL TRAINED ON SYNSET BLVD. AND TESTED ON CCSV. WE ADD ADDITIONAL CRITERIA FOR THE MATCHING OF MODELS LEADING TO CAR MODELS BEING MORE SIMILAR BETWEEN BOTH DATASETS FOR A SINGLE CLASS, BUT REDUCING THE TOTAL NUMBER OF OVERLAPPING CAR MODELS. ABBREVIATIONS: Y – YEAR, F – FACELIFT, C – COUNTRY, MLP – MASKED LICENSE PLATE.

Y	F	C	F1	F1 MLP	F1 CCSV	# Models
			$50.3_{\pm 0.6}$	$50.1_{\pm 0.4}$	$97.4_{\pm 0.0}$	51
✓			$93.5_{\pm 0.9}$	$94.1_{\pm 0.7}$	$100.0_{\pm 0.0}$	21
✓	✓		$100.0_{\pm 0.0}$	$99.8_{\pm 0.3}$	$100.0_{\pm 0.0}$	12
✓	✓	✓	$100.0_{\pm 0.0}$	$99.8_{\pm 0.3}$	$100.0_{\pm 0.0}$	10

Overall, they indicate a very good generalization ability and usefulness of synthetic data for training VMMR models, with a perfect classification if the vehicle models in the training set match the models in the test set in terms of model year and facelift variant. Additionally, the results show the importance of annotating the model years for a dataset, since not considering this leads to a drop of 43.2 points in terms of F1 score in our experiments. This insight is particularly important considering that CCSV has not annotated any years. For the overlapping models, we have manually annotated model years, facelift variant, and country variant to check for a fine-grained matching with the Synset Blvd. models. As expected, including the facelift constraint increases the accuracy significantly since the training examples now better match the test examples. It indicates that considering the facelift can be equivalently important as considering the model generation of a car. Adding the facelift constraint is increasing the F1 score to 100% with no improvement to be gained anymore by additionally considering the country variant of the vehicle models. So, the differences between car models of different countries, such as different positioning of the front light elements seem to have a negligible impact. We additionally evaluated a training with masked license plates to suppress overfitting on the limited variation of license plates as described in Tab. VII. However, for the evaluated scenario the differences to the regular dataset in terms of F1 score are lower than the standard deviation of the training runs. This indicates that the design of the license plate has only a minor impact on the classification. For reference, we include results of a model trained on the CCSV train set which achieves an F1 score close to 100 for all evaluated scenarios but the set of vehicles which do not match the model year with the equivalent SynSet Blvd. vehicle models. Thus, this specific set of vehicle models is likely harder to distinguish than the other evaluated models. While the higher F1 scores of the CCSV training might seem to indicate a domain gap between the synthetic dataset and the real-world dataset, most of the accuracy gap can be attributed to the older generations of vehicle models present in CCSV, biased sampling strategy of datasets like CompCars, and general domain gaps between datasets. In this regard, our results are in line with other authors evaluating in cross-dataset

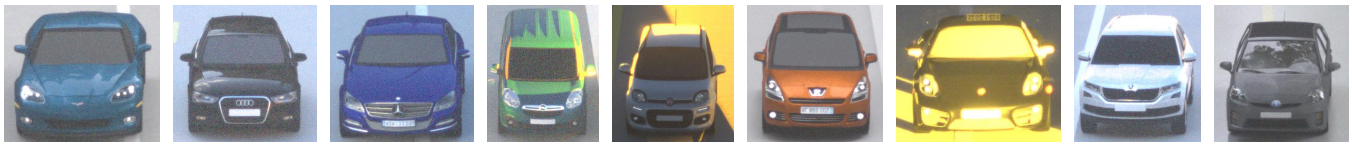


Fig. 5: Samples from the dataset of the "Bayer, bad" configuration, vehicle images cropped to their (slightly and randomly padded) bounding box.

TABLE VI
EVALUATION OF CONVNEXT-SMALL TRAINED ON DIFFERENT CONFIGURATIONS OF SYNSET BLVD. AND TESTED ON CCSV. ABBREVIATIONS: CONF. – CONFIGURATION, MLP – MASKED LICENSE PLATE, BAY. – BAYER.

Conf.	F1	F1 MLP	Conf.	F1	F1 MLP
Bay., good	$99.1_{\pm 0.8}$	$99.7_{\pm 0.2}$	good	$100_{\pm 0.0}$	$99.8_{\pm 0.3}$
Bay., bad	$99.4_{\pm 0.4}$	$99.6_{\pm 0.1}$	bad	$99.3_{\pm 0.5}$	$99.5_{\pm 0.1}$

scenarios, as e.g., Sánchez et al. [2].

In Tab. VI, we analyze the impact of different post-processing configurations of the Synset Blvd. dataset. We use these car models for evaluation for which model year and facelift variant match the models in CCSV. While Bayer good, Bayer bad, and (regular) bad show similar accuracies, (regular) good is slightly in advantage by less than 1 percentage point with an F1 score of 100%. Masking the license plates improves the accuracy slightly for all configurations but (regular) good. Nonetheless, the differences between the configurations are below the standard deviation of the diffrent training seeds for most cases.

V. CONCLUSION AND OUTLOOK

We have presented the Synset Blvd. synthetic dataset for the task of surveillance-nature vehicle make and model recognition (VMMR), representing—to the best of our knowledge—the first entirely synthetically generated large-scale VMMR image dataset. Based on the findings of preliminary studies and prior work on VMMR, the dataset contains 32 400 images of 162 classes (259 200 with variations), rendered via the Cycles path tracing engine in the OCTANE simulation, under different data- and physically-based parameters modeling the imaging process. Annotated detail variations include car paint colors and metalness, road surface conditions, approx. daytime, and imaging process parameters.

Applications of the generated data for a VMMR ML task on the real-world CompCars dataset indicate that synthetic data enable an ML performance comparable to real datasets. Specifically, some challenges with manual annotation are successfully avoided; however, some challenges specific to the synthetic generation remain that will be detailed below.

Outlook

The presented dataset contains several known limitations that were outlined in Tab. VII. Most prominently, the intra-class variation should be improved by including more vehicle model variants and realistic vehicle light functions.

To effectively resolve the challenge in VMMR of keeping the dataset "up to date" with new models being released, the systematic extension would be required, primarily by reducing the manual effort of preparing the vehicle models and parameters for simulation. To enable the use even in critical or "high-risk" applications, such as law enforcement, a considerably more substantial evaluation of the achievable reliability in the trained ML models is required.

The approach of using a data- and physically-based simulation enables a detailed sensitivity analysis to parameters (both from the perspective of VMMR and from the perspective of evaluating the effects of synthetic data) that was conducted and supported only to a limited degree. A more comprehensive and quantitative variation of camera, environment, and lighting parameters would enable a more exhaustive and comparative study of effects, including a comparison of rasterization-based approaches, path/ray tracing-based approaches, the addition of style transfer, with different domain variations of real-world data, to establish a basis for quantifying the "sim-to-real" gap with respect to other types of domain gaps.

TABLE VII
COMPARISON OF THE KNOWN ADVANTAGES AND LIMITATIONS OF SYNSET BLVD. IN RELATION TO CCSV.

Features of Synset Blvd.		Limitations of Synset Blvd.
Less effort expanding without domain divergence		Smaller number of classes and thus of total images
Balanced class distribution	*different*	Less intra-class variance due to single vehicle model variants
Decreased geographic bias	from	Fixed license plate modulation per vehicle model (except MLP subset)
No privacy issues	CCSV	No complex shadows/reflections/occlusions
No mislabeling due to human errors		No low-light/nighttime images
Variations are capable of parametrization	*in common*	Opaque/blurred windshield modulation might enlarge domain gaps
High coverage of front camera perspectives	with	Only frontal perspective images
Results comparable to other cross-dataset results	CCSV	No extreme weather conditions (snow, raindrops, fog, ...)

References

[1] W. Pan, X. Zhou, T. Zhou, and Y. Chen, "Fake license plate recognition in surveillance videos," *Signal, Image and Video Processing*, pp. 1–9, 2022.

[2] H. C. Sánchez, N. H. Parra, *et al.*, "Are We Ready for Accurate and Unbiased Fine-Grained Vehicle Classification in Realistic Environments?" *IEEE Access*, vol. 9, pp. 116 338–116 355, 2021, Alonso, Ignacio Parra and Nebot, Eduardo and Fernández-Llorca, David. DOI: 10.1109/ACCESS.2021.3104340.

[3] European Commission, *Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, Apr. 2021.

[4] European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, GDPR)*, Apr. 2016.

[5] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[6] J. Tobin, R. Fong, A. Ray, *et al.*, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Schneider, Jonas and Zaremba, Wojciech and Abbeel, Pieter, IEEE, 2017, pp. 23–30.

[7] A. Prakash, S. Boochoon, M. Brophy, *et al.*, "Structured domain randomization: Bridging the reality gap by context-aware synthetic data," in *2019 International Conference on Robotics and Automation (ICRA)*, Acuna, David and Cameracci, Eric and State, Gavriel and Shapira, Omer and Birchfield, Stan, IEEE, 2019, pp. 7249–7255.

[8] J. Tremblay, A. Prakash, D. Acuna, *et al.*, "Training Deep Networks With Synthetic Data: Bridging the Reality Gap by Domain Randomization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Brophy, Mark and Jampani, Varun and Anil, Cem and To, Thang and Cameracci, Eric and Boochoon, Shaad and Birchfield, Stan, Jun. 2018.

[9] R. Khirodkar, D. Yoo, and K. Kitani, "Domain randomization for scene-specific car detection and pose estimation," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2019, pp. 1932–1940.

[10] T. S. Kim, B. Shim, M. Peven, *et al.*, "Learning From Synthetic Vehicles," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, Qiu, Weichao and Yuille, Alan and Hager, Gregory D., Jan. 2022, pp. 500–508.

[11] J. Krause, M. Stark, *et al.*, "3D Object Representations for Fine-Grained Categorization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Deng, Jia and Fei-Fei, Li, Jun. 2013.

[12] L. Yang, P. Luo, *et al.*, "A large-scale car dataset for fine-grained categorization and verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Change Loy, Chen and Tang, Xiaoou, Jun. 2015.

[13] F. Tafazzoli, H. Frigui, and K. Nishiyama, "A Large and Diverse Dataset for Improved Vehicle Make and Model Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jul. 2017.

[14] L. Lu, P. Wang, and H. Huang, "A Large-Scale Frontal Vehicle Image Dataset for Fine-Grained Vehicle Categorization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1818–1828, 2022. DOI: 10.1109/TITS.2020.3027451.

[15] A. Amirkhani and A. H. Barshooi, "DeepCar 5.0: Vehicle Make and Model Recognition Under Challenging Conditions," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[16] J. Sochor, A. Herout, and J. Havel, "BoxCars: 3D Boxes as CNN Input for Improved Fine-Grained Vehicle Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[17] J. Sochor, J. Špaňhel, and A. Herout, "Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance," *IEEE transactions on intelligent transportation systems*, vol. 20, no. 1, pp. 97–108, 2018.

[18] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, Springer, 2016, pp. 102–118.

[19] M. Johnson-Roberson, C. Barto, R. Mehta, *et al.*, "Driving in the Matrix: Can Virtual Worlds Replace Human-Generated Annotations for Real World Tasks?" *arXiv preprint arXiv:1610.01983*, 2016, Sridhar, Sharath Nittur and Rosaen, Karl and Vasudevan, Ram.

[20] G. Ros, L. Sellart, J. Materzynska, *et al.*, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vazquez, David and Lopez, Antonio M., Jun. 2016.

[21] Y. Yao, L. Zheng, X. Yang, *et al.*, "Simulating content consistent vehicle datasets with attribute descent," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceed-*

ings, Part VI 16, Naphade, Milind and Gedeon, Tom, Springer, 2020, pp. 775–791.
[22] https://unity.com.
[23] https://www.unrealengine.com.
[24] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual Worlds as Proxy for Multi-Object Tracking Analysis," in *CVPR*, 2016.
[25] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," *arXiv preprint arXiv:2001.10773*, 2020.
[26] https://www.octane.org.
[27] https://www.ogre3d.org.
[28] https://www.cycles-renderer.org.
[29] K. He, X. Zhang, *et al.*, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Ren, Shaoqing and Sun, Jian, Jun. 2016, pp. 770–778.
[30] M. Contributors, *OpenMMLab's Image Classification Toolbox and Benchmark*, https://github.com/open-mmlab/mmclassification, 2020.
[31] J. Deng, W. Dong, R. Socher, *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Li, Li-Jia and Li, Kai and Fei-Fei, Li, Ieee, 2009, pp. 248–255.
[32] H. Jang, D. McCormack, and F. Tong, "Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images," *PLoS biology*, vol. 19, no. 12, e3001418, 2021.
[33] D. Chugai and O. Chugai, http://texturelib.com.
[34] https://polyhaven.com.
[35] https://doschdesign.com.
[36] https://www.cgtrader.com/3d-models/car/luxury-car/silver-jeep-3d-model.
[37] Kraftfahrt-Bundesamt, *Dezente Farben nach wie vor gefragt*, https://www.kba.de/DE/Statistik/Fahrzeuge/Neuzulassungen/Farbe/2021/2021_n_farbe_kurzbericht_pdf.pdf, Accessed: 2023-02-20, 2021.
[38] D. Russell, https://github.com/DeclanRussell/NvidiaAIDenoiser, 2017.
[39] G. A. C. ADAC, *Autokatalog (car catalog)*, https://www.adac.de/rund-ums-fahrzeug/autokatalog/.
[40] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, *YOLOX: Exceeding YOLO Series in 2021*, 2021. arXiv: 2107.08430 [cs.CV].
[41] *Car license plates dataset*. [Online]. Available: https://makeml.app/datasets/cars-license-plates.
[42] Z. Liu, H. Mao, C.-Y. Wu, *et al.*, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Feichtenhofer, Christoph and Darrell, Trevor and Xie, Saining, Jun. 2022, pp. 11 976–11 986.
[43] E. D. Cubuk, B. Zoph, *et al.*, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Shlens, Jonathon and Le, Quoc V., Jun. 2020.