LEARNING FROM SYNTHETIC VEHICLES

Tae Soo Kim *Bohoon Shim* *Michael Peven* *Weichao Qiu*
Alan Yuille *Gregory D. Hager*

Johns Hopkins University
Department of Computer Science

ABSTRACT

In this paper, we release the Simulated Articulated VEhicles Dataset (SAVED) which contains images of synthetic vehicles with moveable vehicle parts. SAVED consists of images that are much more relevant for vehicle-related pattern-recognition tasks than other popular pretraining datasets such as ImageNet. Compared to a model initialized with ImageNet weights, we show that a model pretrained using SAVED leads to much better performance when recognizing vehicle parts and orientation directly from an image. We also find that a multi-task pretraining approach using fine-grained geometric signals available in SAVED leads to significant improvements in performance. We release SAVED and instructions on how to simulate a custom dataset here[1].

***Index Terms*—** Synthetic data, neural networks, vehicle pose estimation, vehicle parts, cars

1. INTRODUCTION

Annotating large scale datasets may be prohibitively expensive or practically impossible depending on the problem domain. For example, while annotating a presence of a common object in an image may be suitable for large scale data collection with a crowd sourced workforce, collecting detailed 3D information of object parts from real images at scale is far more challenging. When access to a large annotated dataset is limited, practitioners typically rely on pretraining a deep network model on a large scale, but unrelated, dataset and later finetune the poorly initialized model using a small set of annotated samples from the target domain. Our experiments show that this standard practice often leads to models with sub-optimal performance.

In this work, we explore the use of synthetic data to address this challenge. There is growing evidence that a dataset with both real and synthetic images can successfully train deep network models for various vision problems [1, 2, 3, 4]. Given the advancements in graphical renderers such as Unreal Engine 4 and Blender [5] coupled with software developments such as UnrealCV [6], researchers now have direct access to a simulation engine that can generate large scale datasets. Compared to real datasets, the cost of annotating a very large dataset is inconsequential. Moreover, most simulation engines provide information about surface normals and depth maps of the rendered scene in addition to RGB images. Many of such rendering parameters and other auxiliary information that can be extracted from the simulator provide a rich set of 'free' annotations that are difficult to obtain in natural images.

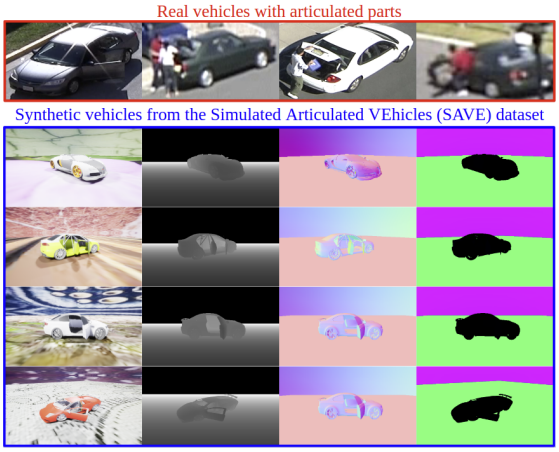

Fig. 1: SAVED is the first large scale dataset of synthetic vehicles with articulated vehicle parts such as doors and trunks. Top: vehicles with articulated parts from the DIVA-Doors dataset, Bottom: simulated instances with domain randomized RGB examples, depth maps, surface normals and semantic segmentation labels (from left to right).

In this paper, we present a synthetic dataset, Simulated Articulated VEhicles Dataset (SAVED), which is used to train a model for localizing vehicle parts (doors and trunks) and estimating vehicle orientation. Though there are real [7, 8, 9] and simulated [10, 11] datasets with vehicle annotations, the granularity of annotations are insufficient for fine-grained analysis of a vehicle's state. Using SAVED, we present the first approach to recognize parts of vehicles and their states (i.e. opened-doors and closed-trunks) from natural images. We demonstrate performance improvements of image classi-

[1] https://taesoo-kim.github.io/

fication through pretraining on a large simulated dataset before finetuning on a smaller set of real images. Interestingly, we show that providing additional supervision with geometric signals during pretraining leads to better performance.

In summary, the following are contributions of this paper:

1. The Simulated Articulated VEhicles Dataset (SAVED): A large scale dataset of rendered synthetic vehicle images with fine-grained vehicle part and 3D geometry annotations.
2. The first model trained using simulated data to recognize vehicle parts and orientation.
3. Experimental evidence that intermediate supervision with geometric signals (i.e. surface normals and depth maps) is critical when pretraining a model using a simulated dataset.

2. RELATED WORK

Learning from simulation. Researchers have successfully trained various computer vision models using simulated data for applications in stereo-vision [12], semantic segmentation [13, 14] and 3D pose estimation [3, 15, 16, 4]. For such tasks, groundtruth annotations on real images are insufficient to train deep neural networks. Using a simulation engine with a software such as UnrealCV [6], groundtruth data that is otherwise difficult to obtain can be generated in large amounts with significantly less effort. SAVED has the same advantage and it provides accurate 3D information of vehicle parts.

Related simulated vehicle datasets. The most notable simulated datasets with vehicle annotations are SYNTHIA [11] and V-KITTI [10]. Viewpoints are limited in these datasets because samples are captured from the point of view of a driver. Hence, the virtual vehicles found in the two datasets are not well suited for training fine-grained models for reasoning about vehicle parts. With more diverse camera viewpoints and articulated vehicle parts, SAVED is a better dataset to train models for problems such as 3D pose estimation and recognizing parts of vehicles.

Simulation to real transfer. Several studies have shown that classifiers trained using simulated images often require methods for simulation-to-real transfer to perform well on real images [17]. We show in our experiments that the use of geometric signals during pretraining with simulation data helps mitigate the issue of domain shift. When there are small number of labeled real instances, we show that a simple approach of pretraining using simulated instances and then later finetuning with real examples leads to best results. We also show that domain randomization techniques [18] as well as intermediate supervision [4] are important when training with synthetic datasets.

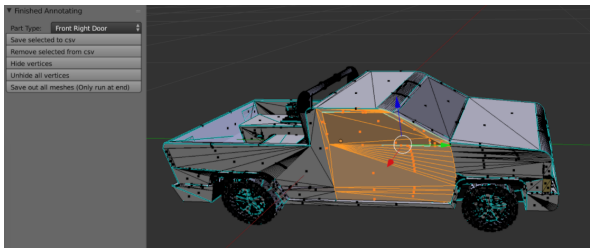

Fig. 2: The custom Blender plugin for annotating vehicle parts such as doors and trunks.

3. THE SIMULATED ARTICULATED VEHICLES DATASET (SAVED)

We describe the details of our Simulated Articulated VEhicles Dataset (SAVED). In contrast to existing real or synthetic datasets of vehicles, the simulated instances in SAVED have moveable parts such as doors, trunks and hoods with ground truth annotations on how much the vehicle part is rotated around its axis. As illustrated in Figure 1, SAVED provides per-pixel depth, surface normal and semantic part labels.

We use Unreal Engine 4 as our renderer of choice and use UnrealCV [6] to interact with the virtual environment to simulate and capture data. We simulate vehicles by rendering the 3D CAD models provided by the ShapeNet dataset [19]. The synthetic vehicles found in ShapeNet do not provide vehicle part annotations as standard. Thus, we manually annotated doors, trunks and hoods of vehicles in order to articulate them as needed using the simulator.

For this purpose, we built a custom Blender plugin (depicted in Fig. 2) to label the sections of the mesh as its corresponding part. To maximize diversity of simulated vehicle appearance in the dataset, we search for similar vehicles via hierarchical clustering over features extracted from rotation invariant 3D shape descriptors using spherical harmonics [20]. We annotated 103 vehicle meshes corresponding to the center of the largest clusters.

Table 1 compares SAVED to other vehicle datasets. Our dataset contains the most number of images captured from a diverse set of camera viewpoints. SAVED is the first dataset with annotations on vehicle parts: we provide the extent to which each door is rotated in degrees. Next, we describe our approach for training with simulated data.

4. LEARNING FROM SYNTHETIC VEHICLES

We use extra geometric information about the scene such as depth maps and surface normals as auxiliary tasks in addition to the main task for the model to optimize for during pretraining with synthetic data. We observe in our experiments that a model initialized using simulation data with this simple

	SAVED (Ours)	KITTI [9]	V-KITTI [10]	SYNTHIA [11]	Pascal 3D+ [7]	EPFL [8]
Real/Simulated	Sim	Real	Sim	Sim	Real	Real
# annotated samples	586,340	80,000	80,000	200,000	6704	2137
Background	Random Texture	Outdoor	Sim. Outdoor	Sim. Outdoor	Indoor+Outdoor	Indoor
Orientation label	yes	no	yes	yes	yes	yes
Azimuth label	yes	no	yes	yes	yes	no
Depth and normal labels	D+N	D	D	D	no	no
Vehicle part Label	yes	no	no	no	no	no

Table 1: Compared to existing datasets with vehicle annotations, SAVED provides vehicle part information and the most comprehensive set of 3D geometry information.

multi-task training approach leads to much better classification performance on real images.

Multi-task approach with geometric signals. We describe our approach for a general classification scenario but our method can be generalized trivially to other problems such as detection and pose estimation. Let $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ be a synthetic training dataset with pairs of a rendered image of a vehicle $x_n \in \mathbb{R}^{H \times W \times C}$ and some corresponding ground truth task label $y_n \in \{1, \ldots, M\}$. The objective is to learn a classifier $\hat{y} = F(x)$ such that the following classification loss $\mathcal{L}_{cls}$ is minimized:

$$\mathcal{L}_{cls} = -\sum_n \sum_m y_n^m log(\hat{y_n^m}) \tag{1}$$

We use a deep neural network for F. This is a standard formulation for optimizing a deep neural network using a cross-entropy loss.

One of the biggest benefits of synthetic data is that a simulation engine has a representation for the 3D geometry of the virtual scene that is readily available. Let $Z = \{z_1, z_2, \ldots, z_N\}$ be a set of some geometric representations such as surface normals or depth maps extracted from the simulator. The intuition behind this approach is that various tasks regarding a vehicle such as vehicle part detection and pose estimation are fundamentally related to its geometry. We use an encoder-decoder framework to jointly predict the task label $\hat{y_n}$ and the geometry $\hat{z_n}$ from x_n. We refactor the classifier F such that:

$$\hat{y} = \text{softmax}(f_{cls}(f(x))) = F(x) \tag{2}$$

where $f(x) \in \mathbb{R}^D$ is an output of an encoder that maps an image to a feature representation of some dimension D and f_{cls} is a linear classification layer that maps feature vectors with D dimension to the output space with M outputs. The output of the encoder $f(x)$ then becomes the input the decoder g to predict the geometric signal $\hat{z}$:

$$\hat{z} = g(f(x)) \tag{3}$$

Then, we define the geometric loss $\mathcal{L}_g$ over all samples as:

$$\mathcal{L}_g = \sum_n d(z_n, \hat{z_n}) \tag{4}$$

where d produces a large scalar when differences between z_n and $\hat{z_n}$ are large. L1 or L2 norms are suitable functions for d, we use the L2 norm in our experiments. The final objective $\mathcal{L}_{final}$ to minimize is then:

$$\mathcal{L}_{final} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_g\mathcal{L}_g \tag{5}$$

During pretraining, we optimize the entire encoder-decoder to minimize $\mathcal{L}_{final}$. During finetuning with real images, we discard the decoder g and use the encoder f initialized using synthetic training examples. We set λ_{cls}, λ_g to 0.5 in our experiments.

5. EXPERIMENTS

We compare the effect of pretraining using a popular ImageNet [21] dataset to SAVED for the tasks of vehicle part recognition and orientation estimation. We show that using synthetic data with geometric signals during pretraining leads to much better results for both tasks. Model specifications and training details are included in the supplementary material.

5.1. Recognizing Vehicle Parts

Using the vehicle part ground truth from SAVED, we train a model for recognizing opened vehicle doors directly from an image.

Dataset. There are no datasets with real images that have ground truth annotations on vehicle door states. To test the ability to train an 'opened-door' detector using simulated images, we manually annotate a small set of real images from the DIVA dataset[2], visualized in Figure 1. We collected 3950 images for training and 732 for validation. Please see the supplementary material for details on how we collected the training data and our experimental setup.

Results. Table 2a shows the performance of a ResNet-101 [22] model pretrained on ImageNet finetuned to our task on DIVA-Doors. A naive approach for training with simulation data is to simply augment the existing real training set with additional synthetic data points. This naive joint training approach only leads to a minor improvements of 0.3 points

[2]https://actev.nist.gov/

	Pretrain	Train	Val Acc
R101-E	ImgNet (R)	DIVA-Doors (R)	75.0
R101-E	ImgNet (R)	Joint (R+S)	75.3
R101-E	SAVED (S)	-	51.8
R101-E	SAVED (S)	DIVA-Doors (R)	80.5

(a) Results from an encoder only model (Res101-E) trained with only classification loss. R: real images. S: simulated images.

	Pretrain	Train	Val Acc
R101-E	ImgNet (R)	DIVA-Doors (R)	75.0
R101-ED-N	SAVED (S)	-	52.8
R101-ED-N	SAVED (S)	DIVA-Doors (R)	85.6
R101-ED-D	SAVED (S)	-	52.3
R101-ED-D	SAVED (S)	DIVA-Doors (R)	83.7

(b) Results comparing encoder-decoder models that trained with auxiliary geometric signals using surface normals (Res101-ED-N) or depth maps (Res101-ED-D).

Table 2: Results on articulated vehicle recognition. We show that pretraining with synthetic auxiliary geometric signals greatly improves model performance on real images.

over the train-on-real-test-on-real baseline. Instead, we observe a much more substantial performance gain of 5.5 points when we follow the pretrain-on-sim-then-finetune-on-real paradigm.

When the model uses geometric signals during pretraining, we observe significantly improved classification results in Table 2b. The model, which uses surface normals (R101-ED-N) to compute the geometric loss during pretraining, has an accuracy of 85.6%, a significant improvement (+10.6%) over the model trained without any simulation data. A model pretrained using surface normals as the multi-task signal outperforms the model that uses depth maps (R101-ED-D) for this application.

5.2. Vehicle Orientation Estimation

Dataset. We use the EPFL [8] dataset which is a small dataset with 20 image sequences of 20 car types at a show. We follow the settings in [24, 25] and report the mean-absolute-error (MeanAE) for evaluated models. Please refer to supplementary material for details.

Results. We implement existing state-of-the-art methods reported for this dataset and report our results in Table 3 as baselines. We choose surface normals as our source for the geometric signal during pretraining. Compared to the DIVA-Door experiments, we see a direct simulation-to-real transfer for this dataset where a model pretrained using only synthetic images without observing a single real image performs on par with baseline models. When both models ([23] and [24]) are pretrained using SAVED and then finetuned using real images, we see significant relative improvements of 51.2% and 36.0% respectively.

	Pretrain	Train	MeanAE
[23]	ImgNet	EPFL	23.8
Our Impl. of [23]	ImgNet	EPFL	24.4
Our Impl. of [23] + N	SAVED	-	23.4
Our Impl. of [23] + N	SAVED	EPFL	11.9
[24]	ImgNet	EPFL	9.86
Our Impl. of [24]	ImgNet	EPFL	10.1
Our Impl. of [24] + N	SAVED	-	12.3
Our Impl. of [24] + N	SAVED	EPFL	6.46
[25]*	ImgNet	EPFL	6.04

Table 3: Results on the EPFL dataset. We improve the existing state-of-the-art models using our approach and pretraining on SAVED. + N indicates that the model is pretrained with surface normals as the geometric auxiliary signal. Lower is better. * We were unable to replicate [25]

6. CONCLUSION

The presented SAVED dataset is the first dataset of synthetic vehicles with articulated vehicles parts with 3D geometry annotations. Using SAVED to pretrain deep neural networks, we showed that we can recognize vehicle parts such as opened doors directly from real images. Using our multi-task formulation with geometric auxiliary signals, we obtained models that generalize to real images much more effectively. In the case of vehicle orientation estimation, a model trained using only synthetic images transferred directly to real images. We wish SAVED contributes to development of new methods for training with synthetic images and approaches for more finegraine analysis of vehicles.

Acknowledgements. This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345 and by the National Science Foundation under Grant No. 1763705. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, expressed or implied, of IARPA, DOI/IBC, the U.S. Government, or the National Science Foundation.

7. REFERENCES

[1] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba, "Virtualhome: Simulating household activities via programs," in *CVPR*, 2018.

[2] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To,

Eric Cameracci, Shaad Boochoon, and Stan Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *CVPR Workshops*, 2018, pp. 969–977.

[3] Jiteng Mu, Weichao Qiu, Gregory D. Hager, and Alan L. Yuille, "Learning from synthetic animals," in *CVPR*, 2020.

[4] C. Li, M. Z. Zia, Q. Tran, X. Yu, G. D. Hager, and M. Chandraker, "Deep supervision with intermediate concepts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[5] Blender Online Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[6] Weichao Qiu, Fangwei Zhong, Yi Zhang, Siyuan Qiao, Zihao Xiao, Tae Soo Kim, and Yizhou Wang, "Unrealcv: Virtual worlds for computer vision," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017.

[7] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese, "Beyond pascal: A benchmark for 3d object detection in the wild," in *WACV*, 2014.

[8] M. Ozuysal, V. Lepetit, and P. Fua, "Pose estimation for category specific multiview object localization," in *CVPR*, 2009.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012.

[10] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *CVPR*, 2016.

[11] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez, "The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes," 2016.

[12] Yi Zhang, Weichao Qiu, Qi Chen, Xiaolin Hu, and Alan L. Yuille, "Unrealstereo: Controlling hazardous factors to analyze stereo vision," in *3DV*, 2018.

[13] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun, "Playing for data: Ground truth from computer games," in *ECCV*, 2016.

[14] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *CVPR*, 2016.

[15] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen, "Synthesizing training images for boosting human 3d pose estimation," in *3DV 2016*, pp. 479–488.

[16] Grégory Rogez and Cordelia Schmid, "Mocap-guided data augmentation for 3d pose estimation in the wild," in *NIPS*, Red Hook, NY, USA, 2016.

[17] David Vazquez, Javier Marin, Antonio Lopez, Daniel Ponsa, and David Geronimo, "Virtual and real world adaptation for pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 797–809, 2014.

[18] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid, "Learning from synthetic humans," in *CVPR*, July 2017.

[19] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu, "ShapeNet: An Information-Rich 3D Model Repository," Tech. Rep., 2015.

[20] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3d shape descriptors," in *Symposium on Geometry Processing*, 2003.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] Kota Hara and Rama Chellappa, "Growing regression tree forests by classification for continuous object pose estimation," *Int. J. Comput. Vis.*, 2017.

[24] Kota Hara, Raviteja Vemulapalli, and Rama Chellappa, "Designing deep convolutional neural networks for continuous object orientation estimation," *CoRR*, 2017.

[25] Xiaofeng Liu, Yang Zou, Tong Che, Peng Ding, Ping Jia, Jane You, and B.V.K. Vijaya Kumar, "Conservative wasserstein training for pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.