SynCity – Syntetisk data för träning av djupa neuronnät

Publik rapport

Författare: Jonas Unger
Datum: 2018-07-30
Projekt inom Maskininlärning för fordonsindustri - FFI

FFI Fordonsstrategisk Forskning och Innovation
VINNOVA Energimyndigheten TRAFIKVERKET FKG VOLVO SCANIA VOLVO

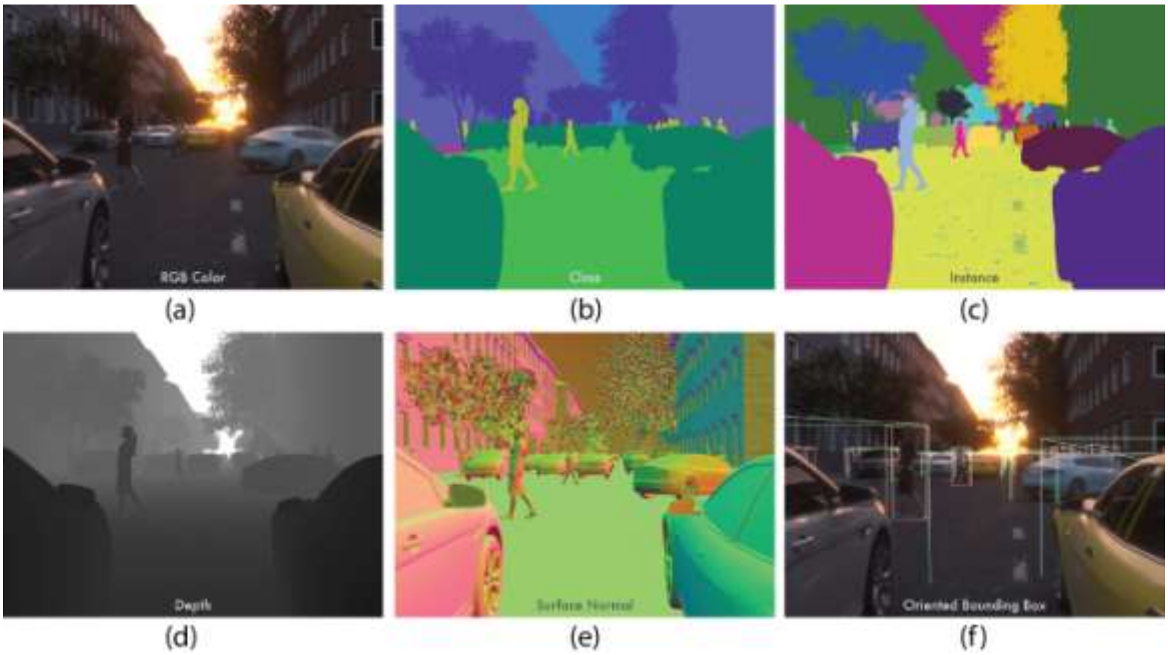
Table of contents

Kort om FFI
FFI är ett samarbete mellan staten och fordonsindustrin om att gemensamt finansiera forsknings- och innovationsaktviteter med fokus på områdena Klimat & Miljö samt Trafiksäkerhet. Satsningen innebär verksamhet för ca 1 miljard kr per år varav de offentliga medlen utgör drygt 400 Mkr.

För närvarande finns fem delprogram; Energi & Miljö, Trafiksäkerhet och automatiserade fordon, Elektronik, mjukvara och kommunikation, Hållbar produktion och Effektiva och uppkopplade transportsystem. Läs mer på www.vinnova.se/ffi.

1 Sammanfattning på svenska

Det här projektet är en konceptstuide som har haft som mål att utveckla nya metoder för att generera syntetiska data för träning av djupa neuronnät för tillämpningar inom datorseende för autonoma fordon, samt att göra tillgängligt ett ”state-of-the-art” dataset för träning av djupa neuronnät under open-source-licens. Det dataset som har utvecklats innehåller foto-realistiskt simulerade kamerabilder (simulerar fordonets bild-sensorer) samt tillhörande "ground truth" annoteringar vilka möjliggör träning av djupa neuron-nät för tillämpningar såsom semantisk segmentering och objekt-detektion. Figur 1 visar ett exempel på: (a) en simulerad bild från en kamerasensor, (b) per pixel-annoteringar där bilden kategoriseras som olika objektklasser för semantisk segmentering, (c) per-pixel annoteringar där bilden kategoriseras som olika objektklasser och instanser inom varje enskild klass, (d) avståndsinformation, (e) ytnormaler, och (f) ”bounding box” i 3D för objektdetektion. Sensor-simuleringen (a) används i kombination med en eller flera av referns-annoteringarna (b) – (f) för att träning, validering och/eller test av djupa neuronnät.



Figur 1: visar ett exempel på (a) en simulerad sensor-bild, (b) – (f),”ground truth” annoteringar som används för träning, validering och eller testning av djupa neuronnät.

Projektet drivs av forskare vid Linköpings Universitet och 7DLabs Inc. (USA) och har i diskussion med våra industriella partners tagit fram riktlinjer för vilka tillämpningar inom datorseende som är av störst vikt för autonoma fordon samt vilka aspekter av syntetiska data som är centrala för dessa. Baserat på detta har vi i en iterativ process förfinat våra metoder för automatisk generering av syntetiska världar och annoteringar för semantisk segmentering och objektdetektion

Arbetet med analys och utvärdering av syntetiska data träning har lett till en fördjupad förståelse om både hur data kan syntetiseras så effektivt som möjligt och hur strategier för träning bör utformas för att nå bästa prestanda vid träning av djupa neuronnät för datorseendetillämpningar. Projektet har utforskat hur beräkningskomplexitet och realism hos simulerad/syntetiska data påverkar resultatet vid träning av djupa neuronnät. Figur 2 visar ett exempel på detta där bilder av samma miljöer och med samma innehåll genereras med hjälp av olika typer av bildsyntes från

Figur 2: visar exempel på data genererad med olika metoder med varierande beräknings -komplexitet. Våra studier visar att noggrann, fotorealistisk simulering av bildsensorn är en mycket viktig komponent vid generering av syntetisk data.

vad som är möjligt att göra i realtid med t.ex. en motor för dataspel till noggrann simulering av sensorer, optik, geometriska objekt, materialegenskaper och ljustransport utförs med hjälp av s.k.
”path tracing”, vilket är en modern metod för noggrann och realistisk bildsyntes. Projektet har också undersökt hur domänskiftet mellan syntetiska och verkliga data kan överbryggas. Resultaten av analyserna visar att domänskiftet mellan den syntetiska data till ”verklig” data som fångats in med en kamera och annoterats för hand är lika stort som domänskiftet mellan två ”verkliga” dataset som fångats in med olika kameror under olika förhållanden. Våra studier och utvärderingar visar att noggrannheten i resultaten från datorseende-arkitekturer baserad på djupa neuronnät kan förbättras avsevärt om ett mindre verkligt dataset kompletteras med syntetiska data. Projektet har med tydlighet visat att syntetiska data är en möjliggörande faktor för utveckling och utvärdering av nya algoritmer för autonoma fordon.

Det publika datasetet kommer att presenteras vid konferensen ACM SIGGRAPH 2018 - Driving Simulation and Visualization, Vancouver Convention Centre den 14 augusti 2018 och kommer sedan att göras tillgängligt genom URL: http://vcl.itn.liu.se och https://7dlabs.com/.

2 Executive summary in English

This project is a proof-of-concept project and has had two goals. The first goal is to develop and evaluate new methods for generation of synthetic data for training of deep neural nets for applications in computer vision for autonomous vehicles. The second goal is to make available a state-of-the-art data set for training of deep neural networks. The data set, which has been developed within the project, consists of photo realistic simulations of camera images (simulating a vehicles image sensor(s)) and the corresponding ground truth annotations which enables training, validation and testing of deep neural networks designed for semantic segmentation and object detection. Figure 1 shows: (a) an example of a synthesized camera sensor image, (b) per-pixel annotations in which the image is segmented into a set of classes, (c) per-pixel annotations in which the image is segmented into classes and instances within each class, (d) depth information, (e) surface normal information, and (f) 3D bounding boxes around each object. The imaging sensor simulation (a) is used in combination with one or more of the ground truth annotations in (b) - (f) as training or validation data for deep neural networks for computer vision tasks. The goals of this project can be summarized in three main items:

1. Build a state-of-the-art synthetic training data set designed for automotive applications and release this to industry and academia under open source licenses.
2. Understand how the level of realism (accuracy) in the image synthesis and sensor simulation training and performance of deep neural networks for computer vision tasks.
3. Investigate how the domain shift between synthetic training data and data from physical sensors can be minimized or even bridged with accurate simulation in the data synthesis.

The project is driven by researchers at Linköping University and 7DLabs Inc. (USA). We have, through discussion with our industrial partners, analyzed which computer vision applications are most critical for autonomous vehicles and which aspects of the synthetic data are most central to these. Based on this analysis, we have, in an iterative process, improved and developed new tools for automatic generation of virtual worlds, synthetic data, and annotations, and integrated them into our systems pipeline for generation and analysis of synthetic data for training and validation of machine learning algorithms. This includes a so called procedural engine for efficient generation and population of virtual worlds and an image synthesis pipeline for accurate simulation of sensors, optics, geometries, lighting environments, material properties and geometries. To address item 2 above, we have conducted a series of experiments where we systematically vary the accuracy and by that the computational complexity in the image synthesis. An example of this is shown in Figure 2, where the simulation of the image sensor is based on different methods ranging from simplistic simulation suitable for real-time rendering to highly accurate simulation of the light transport in the scene using so called path tracing, which is a modern method for photo-realistic image synthesis and simulation of sensors and optics. The results from our studies show that accuracy and realism indeed play key roles in the generation of synthetic data for training and validation. To address item 3 above focused on better understanding the domain shift between synthetic and real data, as well as strategies for efficient training of deep neural networks using synthetic data, we have conducted a series of studies where we compare the performance of a set of representative state-of-the-art deep learning architectures trained using synthetic data to those trained using real data using cross-validations. By real data we refer to training data captured using cameras with ground truth annotations created by hand. From these studies one can conclude that the domain shift between the synthetic data generated using our methods and real data is at the same level as that which occurs between two different real data sets generated using different camera systems. From our studies, we have also seen that mixing of only small real data sets with our synthetically generated data leads to significant improvements in the performance of deep learning architectures for semantic segmentation and object detection. The evaluations show that synthetic data is an enabling factor in the development of new machine learning algorithms for autonomous vehicles.

The state-of-the-art synthetic data set generated within the scope of the project consists of 25,000 synthesized images with a range of corresponding ground annotations. The data set will be presented at the ACM SIGGRAPH conference in Vancouver in August 2018, and will after presentation be made available under an open source license for research and development through our web-pages at URLs: http://vcl.itn.liu.se and https://7dlabs.com/.

3 Background

Artificial Intelligence (AI) is currently developing at unprecedented speed and has a predicted impact in essentially all aspects of modern society. Behind the recent successes of AI lies the data driven approach of neural networks, what is known as deep learning based on layers of interconnected networks. One of the key components in learning approaches is the access to data used in the training phases. This data is in most applications and practical cases be hard to obtain. For instance, a self-driving car needs to be trained with enormous amounts of image sequences representing relevant traffic scenarios and environments.

For over a decade, we have developed a large set of algorithms and methods that allows us to accurately simulate how light interacts with surfaces in the scene to form an image as illustrated in Figure 1, and the machinery required to automatically build entire digital worlds in which the appearance of each element, e.g. houses, cars, the road, Using these technologies we are now, within Vinnova FFI and other research initiatives such as the Wallenberg Autonomous Systems and Software Program (WASP) directed towards the development technology for AI and machine learning, developing new algorithms and tools where we are using sensor modelling and photo realistic image synthesis algorithms to improve the performance of AI and deep learning systems. This allows us to use image and sensor data synthesis to automatically generate vast amounts of training data and develop new tools for introspection of machine learning and AI algorithms.

4 Purpose, research questions and method

Robust semantic segmentation, [1], of and object detection in images are two key challenges in a range of important applications such as autonomous driving, active safety systems and robot navigation. The goal of semantic segmentation is to identify which part of an image, Figure 1(a), corresponds to specific classes, 1(b), such as cars, trucks, trees, buildings etc. and as in 1(c) even distinguish between instances within the different classes. Recently, it has been shown that solutions based on deep neural networks, [1], can solve the related to computer vision tasks with high accuracy and performance. Although deep neural networks in many cases have proven to outperform traditional algorithms, their performance is limited by the training data used in the learning process. A fundamental problem, [2,3], is that there is a lack of both: the availability of training data with accurate per-pixel ground truth annotations, and robust methods for generating such data.

The difficulty is that the training data needs to consist of both the sensor input, i.e. Figure 1(a) and the reference, or ground truth response that the algorithm should learn represented by Figure 1(b-f). If real, captured image/video or other sensor data is used the ground truth annotations need to be created manually by hand, [4,5]. Hand annotation is a difficult, time consuming and unreliable task that does not scale to large data volumes with a typical size of hundreds of thousands or even millions of images.

The problem of generating accurate training data with high quality ground truth annotations has over the last 2-3 years led to the development of methods for creating synthetic data, [5,7], which have proven to increase the performance of deep learning algorithms for certain tasks. It is thus highly important to better understand what are the possibilities and limitations of using synthetic data, as well as what are the trade-offs between the computational complexity in the image synthesis and neural network performance. Previous approaches based on computer game engines have been limited in a way such that this type of analysis cannot easily be performed in a systematic and complete way.

5 Goal

Efficient solutions for generating accurate training data is a pertinent challenge that has the potential to accelerate the development of both: new deep learning algorithms, as well as tools that allows us to analyze their convergence, error bounds, and performance. The goals of this project can be summarized in three main items:

1. Build a state-of-the-art synthetic training data set designed for automotive applications and release this to industry and academia under open source licenses.
2. Understand how the level of realism (accuracy) in the image synthesis and sensor simulation affects the training and performance of deep neural networks for computer vision tasks. Using our image synthesis engine, we can systematically vary virtually all parameters controlling the complexity in the image synthesis and generate data ranging from cartoonish images to images where we can vary and investigate the impact of subtle effects such as the sensor noise model or the way complex lighting effects are simulated.
3. Investigate how the domain shift between synthetic training data and data from physical sensors can be minimized or even bridged with accurate simulation in the data synthesis.

In this project, [6], we use the extensive battery of computer graphics and computer vision libraries and techniques developed within our research to enable the use of state-of-the-art image synthesis designed for photo-realistic simulation of sensor characteristics, optics, detailed geometries, and scattering at surfaces to synthesize training data with pixel-perfect ground truth annotations and labelling. The image synthesis engine combines automatic world generation with accurate light transport simulation and scalable, cloud based computations capable of producing and handling hundreds of thousands or millions of photo-realistic images with known class distributions. Access to large amounts of high quality data has the potential to accelerate the development of both new machine learning algorithms as well as tools for analyzing their convergence, error bounds, and overall performance.

6 Results

This project has extended state-of-the-art in generation of synthetic data and the analysis of how synthetic data can be used to improve the performance of deep neural networks for computer vision. We have developed new tools for automatic generation of virtual worlds, synthetic data, and annotations, and integrated them into our systems pipeline for generation and analysis of synthetic data for training and validation of machine learning algorithms. This includes a so called procedural engine for efficient generation and population of virtual worlds and an image synthesis pipeline based on path tracing for accurate simulation of sensors, optics, geometries, lighting environments, material properties and geometries. Using this image production pipeline designed for generation of synthetic training data, we have developed a state-of-the-art data set for training, validation and testing of machine learning algorithms. The data set will be made publically available for research and development.

To address the research challenges described above, we have conducted a series of experiments where we systematically vary the accuracy and by that the computational complexity in the image synthesis. An example of this is shown in Figure 2, where the simulation of the image sensor is based on different methods ranging from simplistic simulation suitable for real-time rendering to highly accurate simulation of the light transport in the scene using so called path tracing, which is a modern method for photo-realistic image synthesis and simulation of sensors and optics. The results from our studies show that accuracy and realism indeed play key roles in the generation of
synthetic data for training and validation. The project has also investigated the domain shift between synthetic and real data, as well as strategies for efficient training of deep neural networks using synthetic data. We have conducted a series of studies where we compare the performance of a set of representative state-of-the-art deep learning architectures trained using synthetic data to those trained using real data using cross-validations. By real data we refer to training data captured using cameras with ground truth annotations created by hand. From these studies one can conclude that the domain shift between the synthetic data generated using our methods and real data is at the same level as that which occurs between two different real data sets generated using different camera systems. From our studies, we have also seen that mixing of only small real data sets with our synthetically generated data leads to significant improvements in the performance of deep learning architectures for semantic segmentation and object detection. The evaluations show that synthetic data is an enabling factor in the development of new machine learning algorithms for autonomous vehicles.

The state-of-the-art synthetic data set generated within the scope of the project consists of 25,000 synthesized images with a range of corresponding ground annotations. The data set will be presented at the ACM SIGGRAPH conference in Vancouver in August 2018, and will after presentation be made available under an open source license for research and development through our web-pages at URLs: http://vcl.itn.liu.se and https://7dlabs.com/.

7 Knowledge transfer and publication

7.1 Transfer of knowledge and results

How have and will the results from the project be spread?	Mark with X	Comment
Increased knowledge within the area	x	Through publications, the publicly available data set, and through presentations.
Transfer to other advanced technology projects	x	The research conducted within the project are general, and we are already investigating extensions of the results to the medical domain and digital pathology.
Transfer to product development projects	x	The company 7DLabs Inc. are already commercializing products and services around synthetic data for machine learning.
Introduceras på marknaden	x	7DLabs Inc. are working with commercial customers in different application domains, in which the results from this project will have an impact.
Användas i utredningar/regelverk/ tillståndsärenden/ politiska beslut		

7.2 Publications

At the time of writing, the project has generated two scientific publications which currently are under review. We have held several presentations, e.g. through the Wallenberg Autonomous Systems and Software Program (WASP). The project has been covered by IVA Aktuellt Nr 1, 2018, published by Kungliga Ingenjörsvetenskapsakademien.

The open source synthetic training data set will be presented at the ACM SIGGRAPH conference in Vancouver in August 2018, and will after presentation be made available under an open source license for research and development through our web-pages at URLs: http://vcl.itn.liu.se and https://7dlabs.com/.

8 Conclusion and continued research

This project has contributed to the development of technology for generation of synthetic data for machine learning and computer vision in the application domain of autonomous vehicles and self-driving cars. We have shown that carefully constructed high accuracy simulation of image data has great potential to be a key part of the challenge of generating the training/validation data required in the development of future machine learning and AI systems for autonomous vehicles. Synthetic data created in the correct way not only lead to a significant boost in performance, but can also be used for systematic validation and testing of an architecture. Synthetic data is a versatile tool that can be adapted to virtually any application domain, and using the procedural approach used in this project it enables systematic coverage and investigation of the problem space. We see a range of future directions in which continued research is necessary, and will continue to pursue these in the future. A key component of this project and its continuation is to produce PhDs and industry experts with a skill set at the intersection of advanced image synthesis and machine.

9 Project partners and contact persons

Linköping University	Dr. Jonas Unger
7DLabs Inc.	M.Sc. Magnus Wrenninge
Zenuity	Dr. Erik Rosén

References

[1] *A Krizhevsky, I Sutskever, GE Hinton: Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 2012.*

[2] *F. Yu, and V. Koltun, Multi-Scale Context Aggregation by Dilated Convolutions, International Conference on Learning Representations (ICLR), 2016.*

[3] *E. Shelhamer, J. Long, and T. Darrel, Fully Convolutional Networks for Semantic Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 4, Pages 640-651, April 2017.*

[4] *M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.*

[5] *S. R. Richter, V. Vineet, S. Roth, and V. Koltun: Playing for Data: Ground Truth from Computer Games. In proceedings of European Conference on Computer Vision (ECCV), 2016.*

[6] *A. Tsirikoglou, J. Kronander, M. Wrenninge, and J. Unger. Procedural modeling and physically based rendering for synthetic data generation in automotive applications. In arXiv:1710.06270, 2017.*

[7] *G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez, The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes, in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.*