Model-driven Simulations for Computer Vision

VSR Veeravasarapu [1], Constantin Rothkopf [2], Visvanathan Ramesh [1]

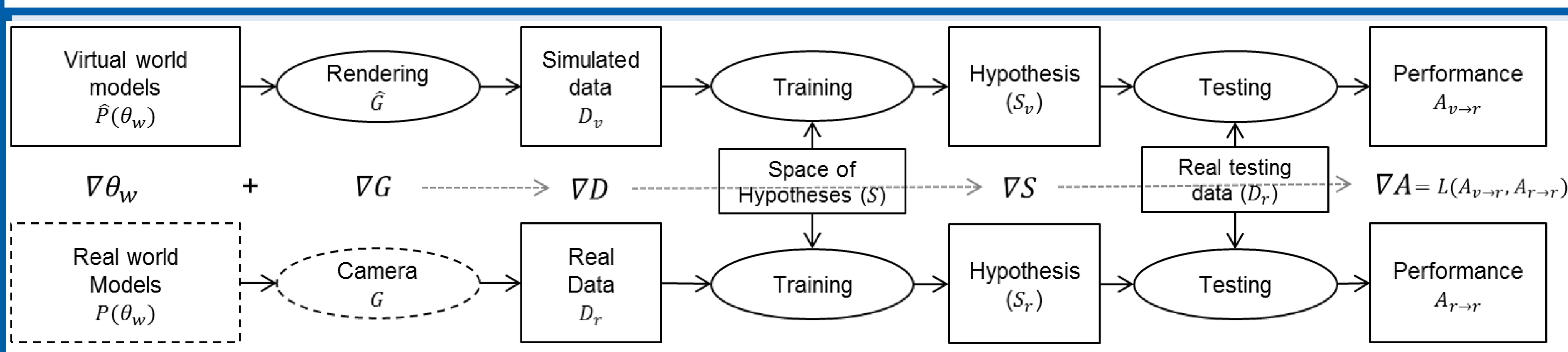
[1] Johann Wolfgang Goethe University, Frankfurt am Main, Germany

[2] Technical University of Darmstadt, Darmstadt, Germany

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Introduction

- Utilizing computer graphics (CG) generated data to train or validate modern computer vision (CV) systems has gained a recent attention due to the scarcity of large scale and well-annotated real world datasets.
- However, some works found that the models trained "only" on simulated data have less generalization capabilities on real data due to the issue of domain-shift. This opened up several fundamental questions about the role of several factors (for instance choice of rendering algorithm) that play a major role in the transfer from virtual to reality.



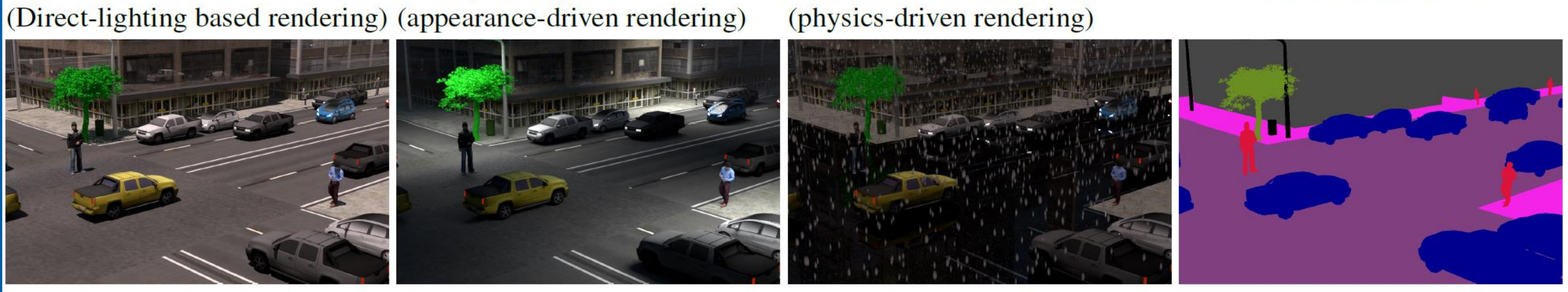
Systems Characterization Perspective [1]

- $\Delta A \approx \mathcal{F}\,(\hat{P}, \hat{G}, S, D_r, L)$
- Here, we take a case study in traffic scenario to empirically analyze the performance degradation due to different choices of $\hat{G}$ (rendering algorithm and it's parameters) when CV systems trained with virtual data are transferred to real data.

Parametric Rendering tool

- CG based rendering algorithms are three types: (i) Local illumination models, (ii) Real time rendering models and (iii) Physically Accurate rendering models.
- Our tool is implemented on top of BLENDER [2] to facilitate the selection of a rendering engine ranging from classical to modern methods and render the data along with required annotations (semantic labels for this work).

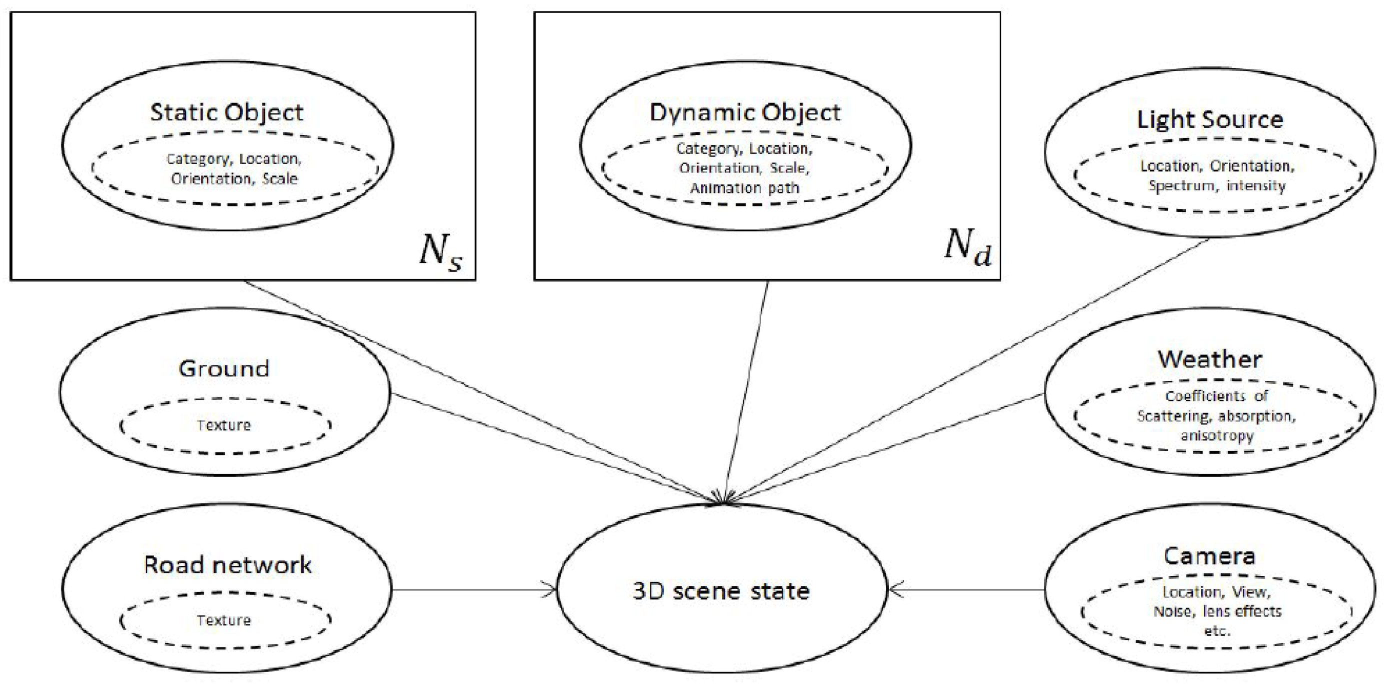
(a) Lambertian (Direct-lighting based rendering) (b) Ray tracing (appearance-driven rendering) (c) Monte-Carlo rendering (physics-driven rendering) (d) Semantic labels



(e) Day light (f) Night (g) Rain (h) Semantic labels

Scene Generative Model

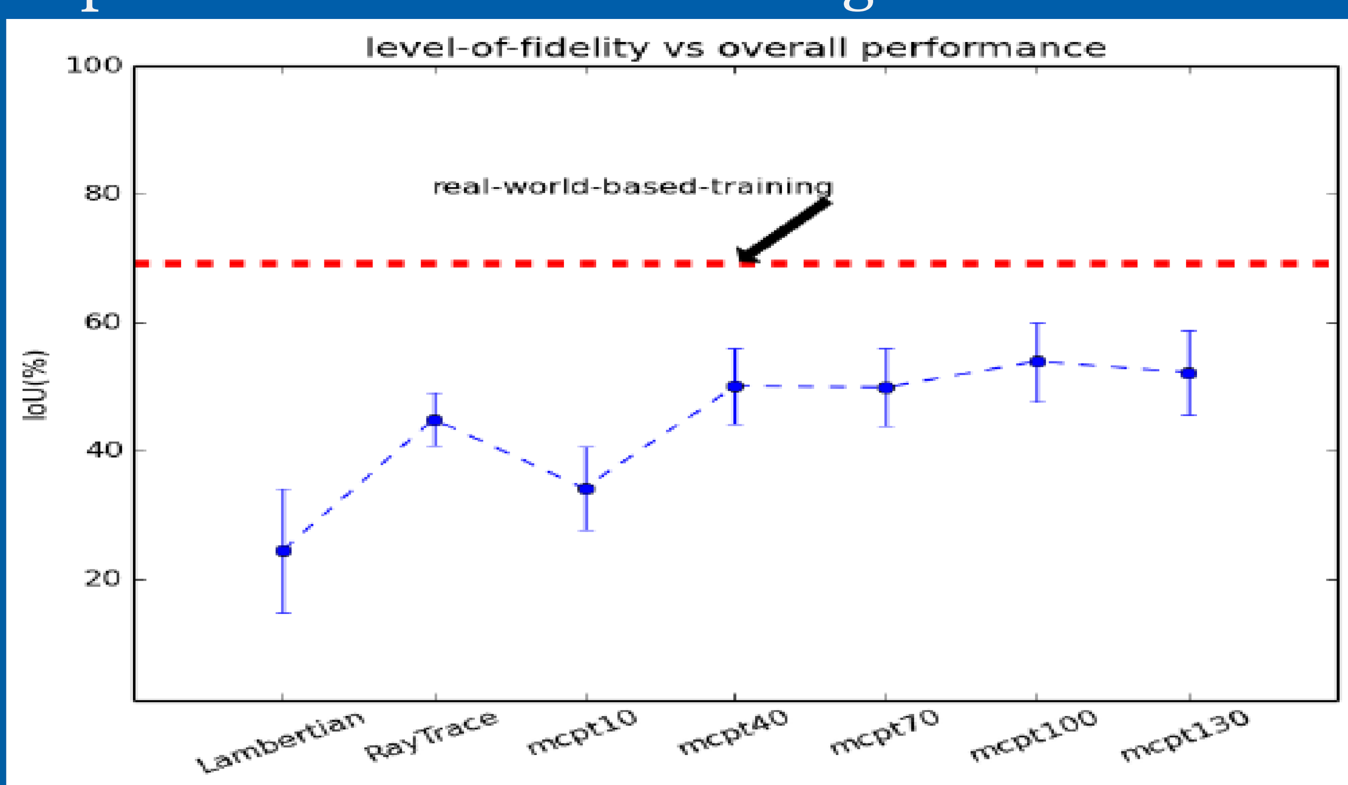
- Our scene generative model is based on Marked point processes coupled with 3D CAD objects and Factor potentials [3].



DeepLab for Pixel-level Semantic Segmentation

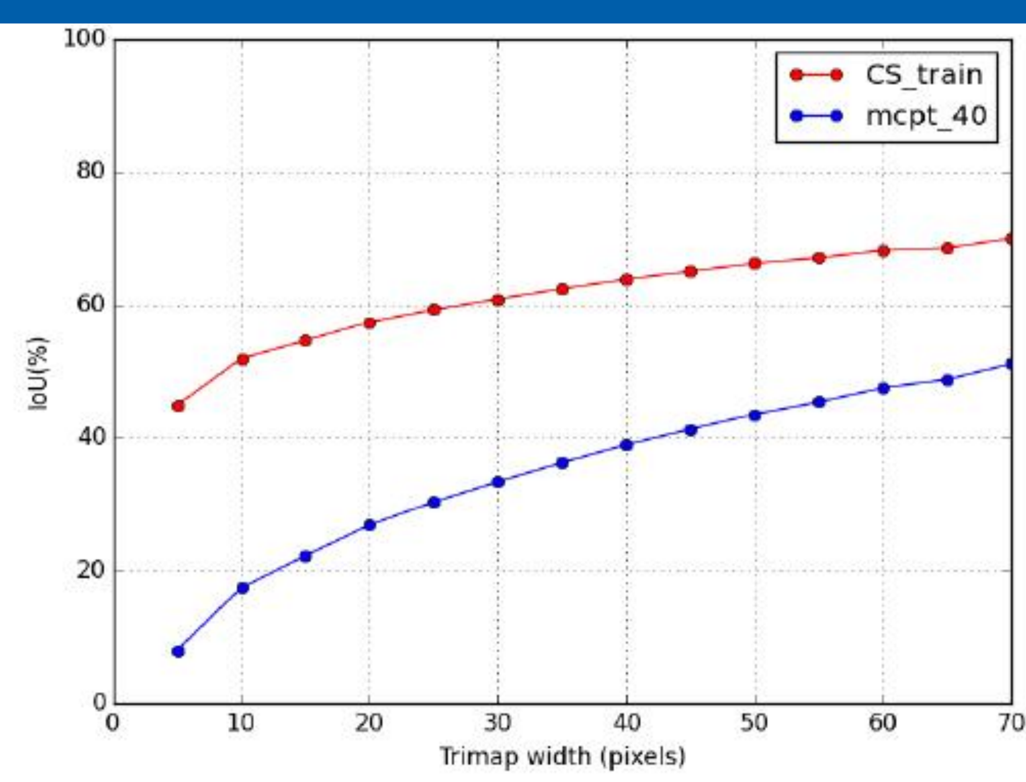
- We use a state-of-the-art CV System, DeepLab (DL) [4] for traffic scene semantic segmentation.
- We simulate 7 sets of CG data rendered with different options of rendering engines and their parameters and analyze the real world performance of DL models trained on the sets.

Experimental Results and Insights



(a) mean IoU vs rendering settings

(a) Input image and corresponding labels and trimaps of 10 and 30 pixel-width



(b) IoU vs Trimap width

- **Impact of Photorealism (Local vs Global illumination rendering)**
  - 20 % improvement
- **Impact of Physical accuracy (Real-time vs Physically accuracy)**
  - 5 % improvement
- **Need of Extreme levels of realism (Samples-per-pixel)**
  - 2% improvement
- **Locations of major errors**
  - Virtual data is statistically more deviated around object boundaries.
- **Things vs Stuff [5]**
  - Per-class performance on Things (Vehicles and Pedestrians etc.) more biased rather than that of the Stuff (Ground, Road, and Sky etc.).
- **Data augmentations:**
  - In our experiments just 10% real world dataset was enough to reach the levels of full real world training.
  - This significantly reduces the number of real world samples required at development phase.

References

1. *Ramesh, Visvanathan. "Performance characterization of image understanding algorithms." PhD diss., University of Washington, 1995.*
2. *www.blender.org*
3. *Veeravasarapu, V. S. R., Constantin Rothkopf, and Ramesh Visvanathan. "Adversarially Tuned Scene Generation." arXiv preprint arXiv:1701.00405 (2017).*
4. *Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." arXiv preprint arXiv:1606.00915 (2016).*
5. *Heitz, Geremy, and Daphne Koller. "Learning spatial context: Using stuff to find things." In European conference on computer vision, pp. 30-43. Springer Berlin Heidelberg, 2008.*

Acknowledgments

This work was funded by the German Federal Ministry of Education and Research (BMBF), project 01GQ0840 and 01GQ0841 (Bernstein Focus: Neurotechnology Frankfurt).

www.uni-frankfurt.de