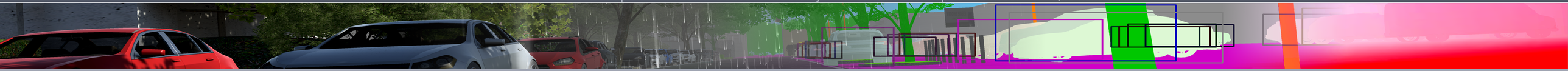
Virtual Worlds as Proxy for Multi-Object Tracking Analysis

Adrien Gaidon[1]*, Qiao Wang[2]*, Yohann Cabon[1], Eleonora Vig[3]
http://www.xrce.xerox.com/Research-Development/Computer-Vision/Proxy-Virtual-Worlds
[1]Xerox Research Centre Europe [2]Arizona State University [3]German Aerospace Center * equal contribution

CVPR 2016



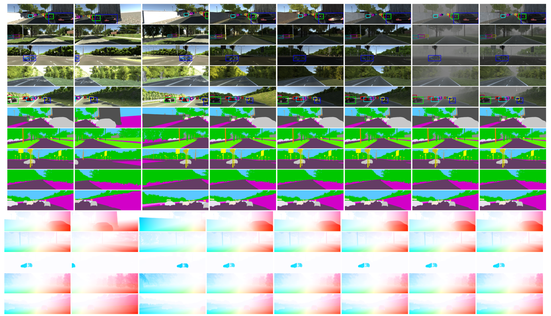
Motivation & Contributions

- **Problem: Labeled Video Data Bottleneck**
  - ☹ high data acquisition and labeling costs
  - ☹ limited variety (weather & imaging conditions, rare events, etc.)
  - ⇒ difficult to train for and benchmark robustness in the wild
- **Solution: Synthetic Data from Game Engines**
  - ☺ modern game engines (e.g. Unity): efficient, low cost, photo-realistic
  - ☺ full control over data generation (quantity & variety)
  - ⇒ large varied scenes with exact dense annotations
- **Contributions**

1. **Virtual KITTI**: fully-annotated photorealistic synthetic *video* dataset
2. Efficient semi-automatic **real-to-virtual world cloning method**
3. **Quantitative measures of usefulness** of virtual worlds as proxies for MOT: *virtual pre-training* and *transferability* across real-to-virtual gap
4. **Measuring impact of simulated weather and imaging conditions**

Virtual KITTI Dataset

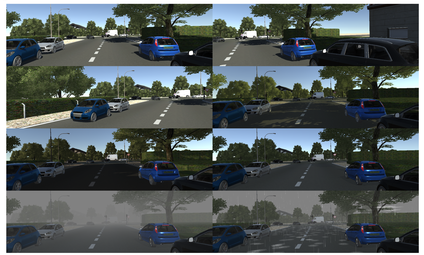
- **40 high-resolution photo-realistic synthetic videos** (17,008 frames) generated from **5 virtual worlds** created in Unity in urban settings under **different imaging and weather conditions**
- **Automatically and densely annotated**: *object detection*, *MOT*, *scene-level* and *instance semantic segmentation*, *optical flow*, *depth*



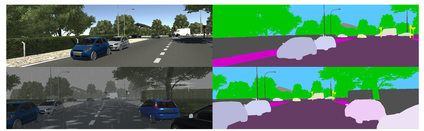
Generating Proxy Virtual Worlds

- **Small seed real-world data**: RGB, LIDAR, GPS/IMU, bounding boxes
- **Generating synthetic clones**:
  - automatic placement of roads and vehicles from seed sensor data
  - manual configuration of background: trees, buildings, light source, etc.

- **Changing conditions in synthetic videos**: camera ∢, lighting, weather



- **Generating ground-truth annotations**: MOT, segmentation, flow, etc.



Experiments – Multi-Object Tracking (quantitative)

- **Strong deep learning baselines for Multi-Object Tracking**
  - **Detector**: Fast R-CNN [2] with Edge Boxes proposals [3]
  - **Trackers**: DP-MCF [4] (57% MOTA on [1]), MDP [5] (76% MOTA)
- **Virtual pre-training improves real-world performance**
  - '**r**': training only on 5 real KITTI sequences
  - '**v**': training only on the corresponding 5 virtual clone sequences
  - '**v**→**r**': *virtual pre-training* on 5 clones + fine-tuning on 5 real-wold videos
  - evaluation on 7 held-out real-world KITTI videos

	MOTA↑	MOTP↑	MT↑	ML↓	IDS↓	FRAG↓	Prec↑	Rec↑
DP-MCF v	64.3%	75.3%	35.9%	31.5%	0	15	96.6%	71.0%
DP-MCF r	71.9%	79.2%	45.0%	24.4%	5	17	98.0%	76.5%
DP-MCF v→r	76.7%	80.9%	53.2%	12.3%	7	27	98.3%	81.1%
MDP v	63.7%	75.5%	35.9%	36.9%	5	12	96.0%	70.6%
MDP r	78.1%	79.2%	60.7%	22.0%	3	9	97.3%	82.5%
MDP v→r	78.7%	80.0%	51.7%	19.4%	5	10	98.3%	82.6%

- **Real-world models transfer across the real-to-virtual gap**
  - real-world trained models: ImageNet → Pascal VOC → KITTI object
  - performance comparison: "seed" real-world *vs.* virtual "clone" videos
  - MOTA difference $< 0.5\%$ → *minimal real-to-virtual performance gap*
  - lower virtual MOTP & MT → *inconsistent annotations in real videos*

	MOTA↑	MOTP↑	MT↑	ML↓	IDS↓	FRAG↓	Prec↑	Rec↑
DP-MCF r	81.7%	85.7%	73.9%	7.2%	26	48	96.2%	88.1%
DP-MCF v	82.2%	78.9%	63.9%	7.9%	21	45	98.2%	86.7%
MDP r	85.9%	84.8%	65.2%	22.1%	4	7	96.7%	90.3%
MDP v	86.0%	80.5%	60.3%	22.1%	0	4	99.1%	87.9%

- **Impact of weather and imaging conditions → overfitting**

MDP	MOTA↑	MOTP↑	MT↑	ML↓	IDS↓	FRAG↓	Prec↑	Rec↑
clone	86.0%	80.5%	60.3%	22.1%	0	4	99.1%	87.9%
+15deg	-5.9%	-0.3%	-7.4%	6.2%	0	0	0.1%	-5.4%
-15deg	-4.5%	-0.5%	-4.8%	5.7%	0	3	-0.5%	-4.0%
morning	-5.1%	-0.4%	-6.1%	3.1%	1	1	0.1%	-4.9%
sunset	-6.3%	-0.5%	-6.4%	4.3%	0	2	-0.3%	-5.5%
overcast	-4.0%	-1.0%	-7.2%	4.6%	0	0	-0.2%	-3.6%
fog	-57.4%	1.2%	-57.4%	40.7%	0	-2	-0.0%	-53.9%
rain	-12.0%	-0.6%	-15.3%	5.7%	1	3	-0.2%	-10.9%

Experiments – Multi-Object Tracking (qualitative)

- Comparing real-world pre-trained models on real videos *vs.* virtual clones

Related Works and Events

- ECCV & ACM-MM 2016 workshop on "Virtual/Augmented Reality for Visual Artificial Intelligence (VARVAI)" http://adas.cvc.uab.es/varvai2016

- Ros *et al.*, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes", *CVPR*, 2016.

- Mayer *et al.*, "A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation", *CVPR*, 2016.

References

[1] Geiger *et al.*, "Are we ready for autonomous driving? The KITTI vision benchmark suite", *CVPR* 2012
[2] Girshick, "Fast R-CNN", *ICCV* 2015
[3] Zitnick, Dollár, "Edge Boxes: locating object proposals from edges", *ECCV* 2014
[4] Pirsiavash *et al.*, "Globally-optimal greedy algorithms for tracking a variable number of objects", *CVPR* 2011
[5] Xiang *et al.*, "Learning to track : online multi-object tracking by decision making", *ICCV* 2015