International Journal of Computer Vision (2018) 126:899–901
https://doi.org/10.1007/s11263-018-1108-0

The Reasonable Effectiveness of Synthetic Visual Data

Adrien Gaidon[1] · Antonio Lopez[2] · Florent Perronnin[3]

Published online: 6 July 2018

The recent successes in many visual recognition tasks, such as image classification, object detection, and semantic segmentation can be attributed in large part to three factors: (i) advances in end-to-end trainable deep learning models (LeCun 2015), (ii) the progress of computing hardware, and (iii) the introduction of increasingly larger labeled datasets such as PASCAL VOC (Everingham et al. 2010), KITTI (Geiger et al. 2012), ImageNet (Russakovsky et al. 2015), MS-COCO (Lin et al. 2014), and Cityscapes (Cordts et al. 2016), among others. In fact, recent results (Sun et al. 2017; Hestness et al. 2017) indicate that the reliability of current visual models might not be limited by the algorithms themselves but by the type and amount of supervised data available. Therefore, to tackle more challenging tasks, such as video scene understanding, progress is needed not only on the algorithmic and hardware fronts but also on the data front, both for learning and quantitative evaluation. However, acquiring and densely labeling a large visual dataset with ground truth information (e.g. semantic labels, depth, optical flow) for each new problem is not a scalable alternative.

Observing the parallel progress of the computer graphics community, computer vision (CV) researchers have recently revived the use of synthetic visual datasets (Butler et al. 2012; Handa et al. 2016; Mayer et al. 2016; Ros et al. 2016; Gaidon et al. 2016; Richter et al. 2017; de Souza et al. 2017; Li et al. 2017; Johnson-Roberson et al. 2017) and simulators (Chen et al. 2015; Shah et al. 2017; Dosovitskiy et al. 2017; Savva et al. 2017) to train and benchmark both CV algorithms and sensorimotor models, respectively. The underlying idea is that datasets of RGB images and videos come with accurate and automatically generated ground truth such as 3D/2D object bounding boxes tracked in time, as well as per-pixel depth, optical flow, semantic classes and instance information; together with privileged information such as crash damage and rule infractions when running navigation simulations.

This special issue of IJCV provides an overview of that rapidly expanding research area with selected papers exploring novel ways to generate and use synthetic visual data for fundamental CV problems and their applications. The reasons for this renaissance include improved photo-realism, better and easier digital authoring tools (e.g., game engines), large libraries of 3D models, and common hardware (e.g., GPUs) to efficiently handle both the generation and use of such visual data.

As evidenced in this special issue, using synthetic visual data is a promising avenue for a variety of applications ranging from optical flow to human pose estimation. Furthermore, there is a large variety of generation strategies, from real-world images mixed with 3D models to full on creation of dynamic virtual worlds. There is no free lunch, and using synthetic data trades off the manual data acquisition and labeling costs for other generation challenges and a “sim2real” domain gap.

The nine articles accepted for publication in this special issue describe and experimentally validate exciting new research directions, such as clarifying the importance of (photo-)realism, overcoming the sim2real gap, learning from virtual humans, analysis-by-synthesis, and recognition in rarely observed conditions that can be simulated. In a way, we are turning the brute-force approach of manual labelling of ground truth and costly sessions of data acquisition, into the generic scientific problem of how to train and test visual/sensorimotor models with synthetic data so that we can ensure that they can ultimately operate in real-world conditions.

Guest Editorial, IJCV Special Issue on Synthetic Visual Data.

✉ Florent Perronnin
florent.perronnin@naverlabs.com

Adrien Gaidon
adrien.gaidon@tri.global

Antonio Lopez
antonio@cvc.uab.es

[1] Toyota Research Institute, Ann Arbor, USA

[2] Computer Vision Center, UAB, Barcelona, Spain

[3] Naver Labs Europe, Meylan, France

We now provide a brief summary of each paper:

- In "Sim4CV—a photo-realistic simulator for computer vision applications", Müller, Casser, Lahoud, Smith, and Ghanem advocate the use of modern game engines for the generation of virtual worlds, as they do not only provide photo-realism but also lifelike physics simulation. Building on top of the Unreal Engine 4 game engine, they provide a customizable simulation environment and validate it on two scenarios: UAV-based tracking of moving objects and autonomous driving.
- In "Configurable 3D scene synthesis and 2D image rendering with per-pixel ground truth using stochastic grammars", Jiang, Qi, Zhu, Huang, Lin, Yu, Terzopoulos, and Zhu focus on the problem of indoor scene understanding. The authors propose a complete learning-based pipeline to generate massive quantities of complex 3D synthetic indoor scenes. The pipeline is configurable in that it enables the customization of important attributes of the generated scenes. 2D images and their ground-truths can be subsequently derived from these 3D environments.
- In "What makes good synthetic training data for learning disparity and optical flow estimation", Mayer, Ilg, Fischer, Hazirbas, Cremers, Dosovitskiy, and Brox argue that existing synthetic datasets for disparity and optical flow such as Sintel (Butler et al. 2012) are sufficiently large for evaluation purposes but too small for training purposes. They consider several ways to generate training data for such tasks including using existing scene data as in Sintel, manually designing new scenes or creating randomized scenes in a procedural manner. An important conclusion of this work is that data does not need to be realistic to make for a good training set.
- In "Augmented reality meets computer vision: efficient data generation for urban driving scenes", Abu Alhaija, Mustikovela, Mescheder, Geiger, and Rother propose to create realistic urban scenes by combining real and synthetic data. Their approach consists in introducing 3D models of objects into real-world images, thus bypassing the problem of modeling complex 3D environments. This idea is validated on the problem of semantic instance-level car segmentation and car detection in outdoor driving scenarios.
- In "Semantic foggy scene understanding with synthetic data", Sakaridis, Dai, and Van Gool argue that the problem of foggy scene understanding has been insufficiently researched given its high practical value, for instance for the problem of autonomous driving. The authors propose to add synthetic fog in real-world images that contain clear-weather outdoor scenes. They also propose a semi-supervised learning approach that leverages clean yet unannotated real-world images and their foggy counterparts.
- In "Image-based synthesis for deep 3D human pose estimation", Rogez and Schmid consider the problem of 3D pose estimation from a single 2D image. The authors propose to generate synthetic data by combining 3D motion capture data with real-word images annotated with 2D poses. Given a candidate 3D pose, their algorithm selects for each joint an image whose 2D pose locally matches the projected 3D pose. It is thus possible to generate a virtually infinite number of realistic "stitched" images. This data is clustered into a large number of pose classes and the problem of pose estimation is treated as one of classification—an approach which is only feasible because of the sheer amount of available labeled data.
- In "3D interpreter networks for viewer-centered wireframe modeling", Wu, Xue, Lim, Tian, Tenenbaum, Torralba, and Freeman consider the problem of recovering the 3D structure of an object from a single image by jointly estimating the object's 3D skeleton and viewpoint. To do so, the authors propose to learn from both 2D-labeled real images and synthetic 3D objects. More precisely, the 3D interpreter network is trained on real images to estimate 2D keypoint heatmaps. It then learns from synthetic data to estimate the 3D structure from heatmaps.
- In "Synthesizing a scene specific pedestrian detector and pose estimator for static video surveillance", Hattori, Lee, Boddeti, Beainy, Kitani, and Kanade consider the surveillance scenario where a newly installed surveillance camera must bootstrap a pedestrian detector and pose estimator without access to any real-world data in the given location. The authors propose to generate synthetic data by leveraging the camera's calibration parameters as well as the scene geometry, both of which are supposed to be known.
- In "Virtual training for a real application: accurate object-robot relative localization without calibration", Loing, Marlet, and Aubry consider the problem of finely localizing and orienting a building block with respect to a robotic arm, both of which are seen from external uncalibrated cameras. This would be of practical value in uncontrolled construction environments for instance. It is shown experimentally that the proposed coarse-to-fine approach can achieve millimetric relative localization without a single real-world training image.

We hope that readers will enjoy this selection of works.

References

Butler, D., Wulff, J., Stanley, G. B., & Black, M. J. (2012). A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European conference on computer vision*.

Chen, C., Seff, A., Kornhauser, A., & Xiao, J. (2015). DeepDriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the international conference on computer vision*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the conference on computer vision and pattern recognition*.

de Souza, C., Gaidon, A., Cabon, Y., & López, A. (2017). Procedural generation of videos to train deep action recognition networks. In *Proceedings of the conference on computer vision and pattern recognition*.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In *Proceedings of the conference on robot learning*.

Everingham, M., Gool, L. V., Williams, C., Winn, J., & Zisserman, A. (2010). The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*, 303–338.

Gaidon, A., Wang, Q., Cabon, Y., & Vig, R. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the conference on computer vision and pattern recognition*.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the conference on computer vision and pattern recognition*.

Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., & Cipolla, R. (2016). Understanding real world indoor scenes with synthetic data. In *Proceedings of the conference on computer vision and pattern recognition*.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., et al. (2017). Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S., Rosaen, K., & Vasudevan, R. (2017). Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *Proceedings of the IEEE international conference on robotics and automation*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Li, X., Wang, K., Tian, Y., Yan, L., & Wang, F.-Y. (2017). The ParallelEye dataset: Constructing large-scale artificial scenes for traffic vision research. In *Proceedings of the IEEE international conference on intelligent transportation systems*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al.. (2014). Microsoft COCO: Common objects in context. In *Proceedings of the European conference on computer vision*.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., et al. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the conference on computer vision and pattern recognition*.

Richter, S., Hayder, Z., & Koltun, V. (2017). Playing for benchmarks. In *Proceedings of the international conference on computer vision*.

Ros, G., Sellart, L., Materzyska, J., Vázquez, D., & López, A. (2016). The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the conference on computer vision and pattern recognition*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Savva, M., Chang, A., D. A., Funkhouser, T., & Koltun, V. (2017). MINOS: Multimodal indoor simulator for navigation in complex environments. arXiv:1712.03931.

Shah, S., Dey, D., Lovett, C., & Kapoor, A. (2017). AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In *Proceedings of the field and service robotics*.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 843–852). IEEE.