Imagining hidden supporting objects in tabletop scenes

Hector Basevi
School of Computer Science
University of Birmingham
West Midlands, United Kingdom
h.r.a.basevi@cs.bham.ac.uk

Aleš Leonardis
School of Computer Science
University of Birmingham
West Midlands, United Kingdom
a.leonardis@cs.bham.ac.uk

Abstract

Understanding of complex disordered piles of objects can require consideration of object support relations contributing to the stability of the scene. Such supporting objects are often partially occluded or completely hidden. We explore how supporting hidden objects may be efficiently imagined. We evaluate existing state-of-the-art regression and generative adversarial machine learning frameworks and demonstrate that neither are sufficient to imagine *stable supporting* hidden objects. We propose a novel framework incorporating an explicit stability learning signal. The addition of this signal biases the imagined object distribution strongly towards objects with a large flat base and low centre of gravity, resulting in maximally supported scenes.

1 Introduction

Understanding of complex scenes is a difficult problem of particular relevance to embodied scenarios where the potential consequences of interactions with scenes must be considered. Occlusion is a major challenge for visual inference and is connected to physical inference because objects involved in physical support of a scene are often themselves occluded. In extreme cases supporting objects can be completely invisible.

Existing systems for physical scene understanding have only partially addressed this issue. Many algorithms for physical scene parsing and understanding focus only on situations where objects are partially or completely visible [1, 2, 3, 4], and those that consider hidden objects do so via a complex and computationally expensive iterative process of object proposal and hypothesis evaluation [5]. We explore whether imagination of hidden supporting objects can be performed in a feed-forward generative fashion using artificial neural networks. We further explore whether learning from data drawn from a distribution of stable scenes is sufficient, and whether a separate learning signal from a pre-existing stability estimation model can lead to more stable imagined objects and scenes.

Our contributions are as follows:

1. We present a novel scenario and dataset for this task.
2. We show that models trained by regression do not imagine hidden objects.
3. We show that conditional generative adversarial networks (cGANs) [6, 7] trained on distributions of stable scenes do imagine hidden objects.
4. We present a framework for training a cGAN which incorporates generator stability conditioning, and discriminator supervision from a pre-existing stability estimation model.
5. We show that incorporating a pre-existing stability estimation model improves scene stability.

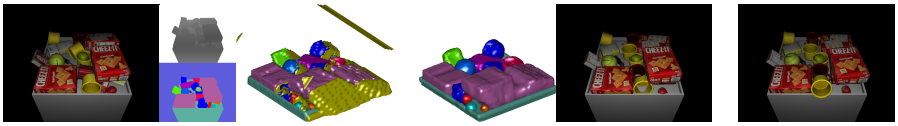
32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

2 Scenario

Support object imagination is tested by imagining hidden parts of physically stable tabletop scenes composed of household objects (see fig. 1). These scenes are generated using random combinations of objects in random but physically stable poses, which removes any semantic content from the data to encourage trained models to make use of stability information. Scenes contain between a single object and 25 objects. These more complex scenes contain significant occlusion and some contain hidden supporting objects (determined by removing the hidden objects and evaluating the consequences via simulation). The task is to produce a scene description from an RGB-D image containing visible objects and potential hidden objects. To prevent visual complexity from confounding the experiments, a 2D segmentation of the image into object types is assumed.

3 Object imagination models and design

Scenes are represented via voxel grids of voxel occupancy with a channel for each object type (see fig. 1). The task is to take as input a partial voxel grid, containing an input channel indicating voxels which are unknown because they lie behind the visible surfaces, and produce complete voxel grids, which are then parsed into object types and poses by a separate algorithm based on an iterative closest point algorithm.



(a) Visual data (b) Explanation (c) Simulation

Figure 1: Input visual data, an explanation for the scene, and the result of a physical simulation of the scene to evaluate its physical stability. The visual data consists of an RGB image, a depth image, a 2D semantic segmentation, and a 3D visible surface. The yellow region of the 3D representation corresponds to occluded/unexplained regions. The explanation consists of a 3D volumetric representation and a set of parsed object types and poses. This parsing can be physically simulated to evaluate stability via object displacement.

Three types of models are compared: regression via voxel error (VER), sampling via Conditional [[6] Improved Wasserstein GAN [8] (CWGAN), and a novel sampling model (S-CWGAN) conditioned on a separate stability signal with stability supervision provided by a pre-trained stability estimation model (see fig. 2).

- VER is trained to produce a ground truth complete voxel grid from a partial voxel grid via least squares regression.
- CWGAN is trained to sample complete voxel grids conditioned on a partial voxel grid using a conditional Wasserstein generative adversarial network framework.
- S-CWGAN is additionally conditioned on a desired scene stability, and the discriminator is conditioned on scene stability supervision produced by a scene stability estimation model.

VER and CWGAN are implemented identically to S-CWGAN, but with unnecessary S-CWGAN components removed (stability conditioning and adversarial components for VER, and stability conditioning and stability supervision for CWGAN).

The scene stability estimation model is trained as a WGAN critic on a separate dataset of scenes. These scenes consist of stable scenes and modified versions on which one of two types of perturbations is applied:

1. The pose of one of the scene objects is randomly perturbed.
2. One of the scene objects is removed.

The first type of perturbation enables the stability estimation model to identify instabilities arising from object intersections and unstable orientations. The second type of perturbation enables the

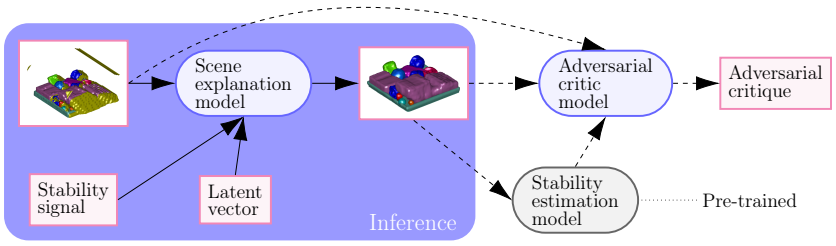


Figure 2: Schematic of S-CWGAN model including stability estimation signal.

stability estimation model to identify instabilities arising from complete lack of support. The resulting set of scenes are split into unstable and stable subsets for training based on object movement during simulation. As a WGAN critic, the stability estimation model is regularised to be Lipschitz-continuous which is expected to improve the quality of the learning signal for the object imagination models.

4 Results

The object imagination dataset consists of 3,900 scenes in a 8:1:1 training:validation:testing split. Each model was trained for 25 epochs, and evaluated using semantic, geometric, imagination, and stability metrics:

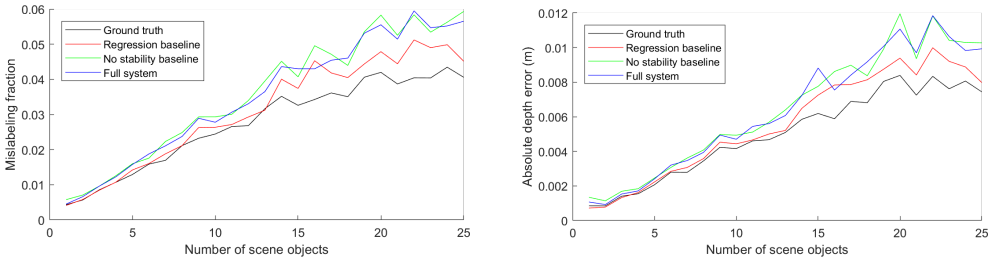
Semantic error: The fraction of pixels with semantic labels differing from the ground truth labels.

Geometric error: The average pixel depth error.

Imagination: The average number of hidden objects imagined.

Stability: The average scene stability for scenes where the ground truth contains hidden objects.

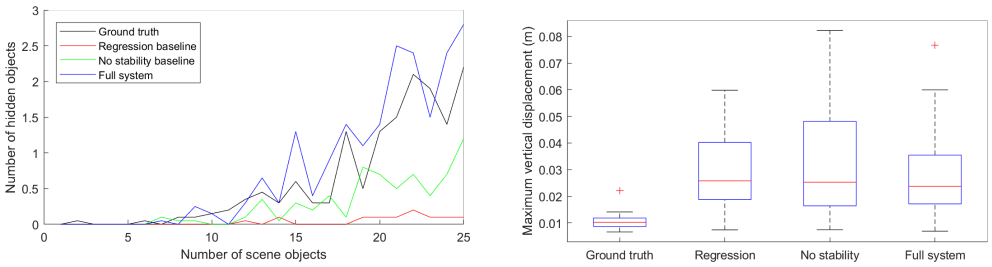
The results show that VER produces slightly better visible geometric and semantic fidelity than the sampling models (see fig. 3), but does not imagine hidden objects (see fig. 4). CWGAN does imagine hidden objects, but imagines fewer than the ground truth and the resulting scenes are unstable. S-CWGAN imagines similar numbers of objects to the ground truth, and imagined scenes are more stable than CWGAN.



(a) Semantic error (b) Geometric error

Figure 3: Depth and mislabeling fraction for ground truth and all explanation models.

S-CWGAN was probed to explore the effect of the random latent input, and the stability conditioning. We found that the latent vector appears to represent noise and minor deformations to the scene, rather than different collections of hidden objects. We hypothesise that this is a result of the cGAN framework and the strong learning signal provided by the stability estimation model. Sampling different input stability values has large effects on the imagined scenes, resulting in boxes on their sides (large flat contact area, low centre of mass) for high stability values and small object fragments for low stability values.



(a) Imagination (b) Stability

Figure 4: Number of hidden objects and maximum object displacement in simulation for all explanation models.

5 Conclusion

We have demonstrated that regression models (VER) are unsuited to imagining hidden objects, and that generative models trained on sets of stable scenes without an explicit stability learning signal (CWGAN) do generate hidden objects but generate fewer than originally present. Adding an explicit stability learning signal (S-CWGAN) results in generation of similar numbers of hidden objects to those originally present, and more stable scenes. The stability signal also induces a strong preference for imagining boxes on their sides, which are extremely stable.

Acknowledgments

We acknowledge MoD/Dstl and EPSRC for providing the grant to support the UK academics' involvement in a Department of Defense funded MURI project through EPSRC grant EP/N019415/1.

References

[1] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems*, pages 127–135, 2015.

[2] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 430–438. PMLR, 2016.

[3] Li Wenbin, Ales Leonardis, and Mario Fritz. *Visual Stability Prediction and Its Application to Manipulation*. IEEE Computer Society, 2017.

[4] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 153–164. Curran Associates, Inc., 2017.

[5] Tianjia Shao, Aron Monszpart, Youyi Zheng, Bongjin Koo, Weiwei Xu, Kun Zhou, and Niloy J. Mitra. Imagining the unseen: Stability-based cuboid arrangements for scene understanding. *ACM Transactions on Graphics*, 33(6), 2014.

[6] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.