Imagining Hidden Supporting Objects using Volumetric Conditional GANs and Differentiable Stability Scores

Hector Basevi
h.r.a.basevi@bham.ac.uk
Aleš Leonardis
a.leonardis@bham.ac.uk

School of Computer Science
University of Birmingham
Birmingham, United Kingdom

Abstract

Objects supporting the physical stability of an unstructured heap of items are often heavily or *completely* occluded by the objects that they are supporting. Identifying plausible supporting object candidates and their poses from visual information is challenging because there may be many candidates and it is not practical to exhaustively verify each one using physical simulation. We present a generative system which predicts the complete volumetric structure of a heap of objects from visible depth and semantic information. We leverage 3D conditional Wasserstein generative adversarial networks to perform this task and inject differentiable context about physical stability from a second network trained to score the physical stability of object heaps. We demonstrate that our system is capable of generating physically stable heaps from visual information, and that the use of both generative models *and* context about physical stability are crucial in replicating the true distribution of hidden objects. We train and evaluate our system using a novel simulation-based dataset[1] which we also present in this work.

1 Introduction

Visual scene understanding [18, 40] involves interpreting the content of an image and inferring properties of the underlying scene. Current research focuses on scene elements or objects which are directly visible in the image but may be partially occluded [35]. In general, imagining objects which may be present but are *completely* occluded is challenging because the potentially large number of possibilities are expensive to evaluate and difficult to use for other tasks.

One scenario in which this challenge is mitigated involves imagination of hidden objects which provide *physical support* to visible objects, either by supporting them from below or from the side, and prevent them from moving or falling. Imagination of these hidden supporting objects can be made on the basis of the physical support relations which need to be present to stabilise the visible objects, *in addition* to whether the hidden object can be contained in the occluded region of empty space. Imagination of hidden supporting objects is also relevant to planning for manipulation tasks, as the configuration of hidden supporting

[1] https://hectorbasevi.github.io/imagining-hidden-supporting-objects

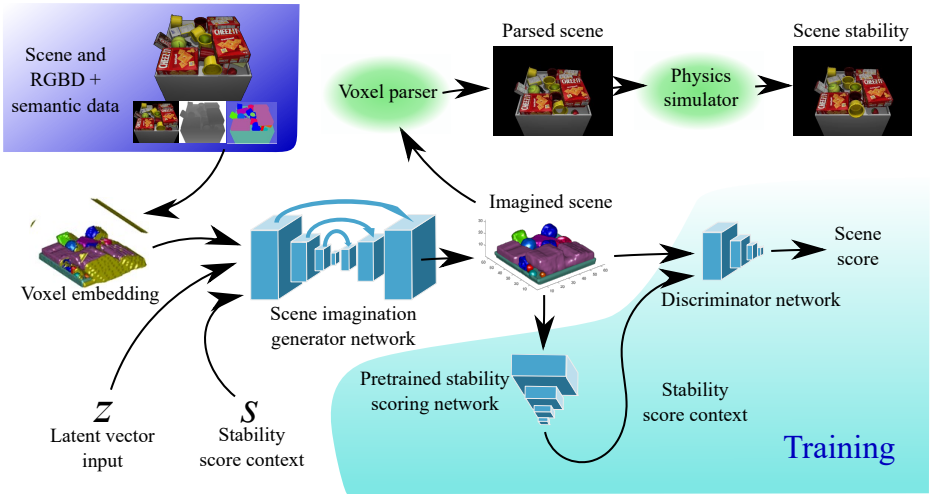


Figure 1: Our proposed system. Semantic information is embedded into a voxel representation using depth information, leaving occluded regions undefined (gold colour in the image). This voxel embedding is fed into a scene imagination generator network, which also takes a latent vector and stability score as context. The generator produces an imagined voxel representation in which all regions are defined. During training this output is then fed into both a discriminator network and a *pre-trained stability scoring network*. The stability scoring network is trained to produce higher scores for stable scenes than unstable scenes using the Wasserstein distance. This score is also fed into the discriminator network. The discriminator learns to distinguish between real pairs of voxel representations and stability scores, and synthetic pairs. The imagined voxel representation can be parsed into a set of object classes and poses using an ICP-based parser [3], and this object-based representation can be fed into a physics simulator to evaluate the stability of the imagined scene.

objects may affect which interactions with the scene are possible without causing objects to fall [24].

Imagining hidden supporting objects can treated as a data-driven problem by recognising that the distribution of the training data set implicitly contains information about physical support and assuming that learning directly from the data should be sufficient to produce plausible supporting objects. The problem can be treated either as a supervised regression task or as a conditional sampling task using a generative model. Alternatively, the learning process can incorporate an explicit learning signal about scene stability to guide training towards physically stable solutions where learning only from the data may be insufficient.

In this work we evaluate the effectiveness of these two approaches for imagining hidden supporting objects. We do so by designing a suitable prediction task and solution design. Our task involves predicting a set of *all* scene objects, and this set of scene objects should be both consistent with the visual information *and physically stable*. We construct a sophisticated and novel dataset for this task consisting of visual, semantic, geometric and physical information. We design a novel objective and training scheme which exploits an *explicit stability learning signal obtained from a neural network pre-trained via physical information*, as shown in fig. 1. We adopt a standardised neural network architecture for fair comparison and compare behaviour against regression and generative adversarial training objectives not

using stability information.

We find that there are trade-offs between the approaches. Our system trained via a regression objective provides the best consistency with regard to the visual information, but produces the fewest hidden objects and poor physical stability. Our generative model trained via an adversarial objective produces more hidden objects but does not improve the physical stability of solutions. Our generative model trained with context from a neural network which scores the stability of the output of our generative model provides the most hidden objects and best stability.

To the best of our knowledge this is the first work to tackle the imagination of completely hidden objects providing physical support using a learnt generative approach informed by an explicit stability learning signal. The main contributions of this work are:

1. A novel generative neural network and training scheme which learns to predict visible objects and imagine hidden objects with context from a pre-trained neural network which scores the stability of the imagined scenes.

2. A novel dataset for imagination of hidden objects consisting of rich visual, semantic, geometric and physical information.

3. An analysis and comparison of the qualitative and quantitative behaviour of regression, generative adversarial, and stability-guided generative adversarial learning and their trade-offs.

2 Related works

There are a number of works and visual benchmarks on prediction tasks where physical stability is a factor. These include the stability of tower structures [10, 17] and predicting future states in multiple scenarios involving instability and collisions [2]. However, these benchmarks centre around understanding the physical behaviour of visible entities whereas the most important entities in our benchmark are *hidden*. Benchmarks involving hidden objects focus on object permanence [26] or non-rigid occluders [32] rather than physical support of rigid objects.

Physical stability has been used to facilitate other tasks, such as constructing structures [19] and scene parsing [6]. Our application is conceptually similar but our design, embodying stability information via a neural network which provides learning supervision to a task network, is more efficient. It avoids the exhaustive pose sampling used in Li *et al*. [19] in their generation phase and is computationally tractable for complex objects and scenes unlike Du *et al*. [6], which makes heavy use of expensive physical simulation in their training phase through REINFORCE [34]. Efforts are being made to reduce the computational cost of obtaining gradients via physical simulation [8].

Scene completion systems produce a full representation of a scene from partial (visible) information, but do not in general consider physical properties of the scene [7, 9, 12, 31]. When physical properties are considered, this information is used to improve the prediction of partially visible objects rather than imagine hidden objects [29, 30].

Scene generation systems sample scenes unconditionally from a target distribution in image space [13], in voxel space [37], or in object pose space [39]. Our system samples all scene objects simultaneously in a voxel space, conditioned on visual *and physical stability* context.

A recent promising strand of research in neural rendering using volumetric radiance fields [21] has begun to explore unconditional scene generation [23] and conditional object completion [58], but has not yet been extended to generation or completion of scenes containing many objects.

3 The task and dataset

Comparing different approaches to imagining hidden objects requires a physical scenario which is realistic but has a limited amount of complexity. We choose to examine unstructured heaps of realistic objects, chosen from the YCB object set [4]. This set consists of 14 objects including cups, plates, bowls, boxes, and fruit. We synthetically generate scenes by choosing a random object from the set and simulating the effect of dropping it on top of the table and any pre-existing objects via a physics engine [5]. Once the objects have come to rest the process is repeated to produce object heaps containing between 1 and 25 objects. Simulation is necessary to collect information at the end of the generation process which is occluded such as the poses of objects which are hidden, and information which cannot be collected in the real world such as locations and magnitudes of contact forces between objects.

We generated a dataset of 3900 scenes consisting of 1400 scenes containing between 1 object and 14 objects drawn without replacement and 2500 scenes containing between 1 object and 25 objects drawn with replacement. These scenes were then subdivided into training (80 %), validation (10 %), and testing (10 %) sets. Having generated a set of object instances and poses for each scene, we used the scene description to produce RGBD images, semantic and instance maps, semantic volumetric occupancy, and contact forces between pairs of objects. All scenes are physically stable by construction. We choose to work in a relatively small data regime for scene imagination to minimise the possibility that the generative model can solve the problem trivially by memorising the training data and replicating the nearest training example.

We generated a separate dataset via the same method to train a system which learns to score scene stability. Such a system requires unstable as well as stable scenes so we generated potentially unstable scenes relating to counterfactual questions: "Would the scene have been unstable if an object was in a different pose?", and "Would the scene have been unstable if an object was missing?". To simulate the former we apply a random pose perturbation to a single object from a stable scene, resulting in one potentially stable or unstable perturbed scene per scene object. To simulate the latter we remove a single object from a stable scene, again resulting in one potentially stable or unstable perturbed scene per scene object. The goal of these questions is to allow the stability prediction system to learn when moving objects affects scene stability, and when *adding or removing objects* affects scene stability. The stability dataset consists of 79 800 scenes in total. The datasets are accessible from the project webpage[2].

[2] https://hectorbasevi.github.io/imagining-hidden-supporting-objects

4 Methods

4.1 The task

Each scene (see fig. 1) consists of a static table, and a set of objects S, where each object can be defined in terms of its class c, rotation r, and translation t: $s_i = \{c_i, r_i, t_i\}$. The visual data resulting from this scene consists of an RGB image I_S and a depth image D_S. Our goal is to imagine a set of objects $\hat{S}$ such that $\hat{S}$ is consistent with I_S and D_S, and $\hat{S}$ is also *physically stable*. We perform the process of inferring or imagining $\hat{S}$ within a semantic voxel space for each class of object because of its advantages in the context of this task. Firstly, voxel representations encode geometric information explicitly which can aid inference involving physics as this benefits from information about geometry and collisions (overlapping voxels). Secondly, perturbations to object poses *and the addition and removal of objects* can be performed by modifying sets of voxel values. These are relatively simple operations and involve a representation of fixed size, in contrast to sequences of object poses. In this case, altering an object pose involves different operations to adding or removing an object, and adding or removing objects alters the length of the sequence.

Given I_S, we assume that a semantic segmentation map C_S can be created. State-of-the-art systems for semantic segmentation provide high quality results [36] but to avoid the choice of semantic segmentation system confounding results we operate from ground truth C_S. Given C_S and D_S, we embed the semantic segmentation onto the depth surface to produce a partially segmented voxel representation V_S^p. This representation contains semantic information about the visible surfaces and free space, but lacks information about occluded regions of the scene. The central task of this work is to learn a function g_θ which samples fully segmented voxel representations V_S^f from the conditional distribution:

$$g_\theta(V_S^p) \sim P(V_S^f \mid V_S^p) \tag{1}$$

Given a fully segmented voxel representation V_S^f, we convert back into a set of objects S using a fixed parsing algorithm which is based on ICP [3]. The voxel segmentation is semantic rather than instance-based and so the parsing algorithm is applied to the voxel segmentation of each semantic class in an iterative manner. The parsing algorithm iteratively identifies and removes object instances from the voxel segmentation of the class until the segmentation is well explained by a (possibly empty) set of object instances. Please see supplementary material for additional details.

The representation S allows analysis of visual consistency through rendering, and physical behaviour through simulation. Note that parsing and subsequent simulation operations are only performed for analysis and are not used during the training process. Please see supplementary material for additional details.

4.2 The proposed system

The distribution $P(V_S^f \mid V_S^p)$ may be multimodal if there are several unique sets of hidden objects which are plausible given the visual information. We learn to sample from this distribution using a volumetric conditional [22] Wasserstein Generative Adversarial Network [1]. This system consists of a generator network $g_\theta(V_S^p, z)$ and a discriminator network $d_\phi(V_S^f, V_S^p)$. The generator takes as input a latent vector z and both networks use V_S^p as conditioning information. These networks are jointly trained to solve the objective:

$$\min_{\theta}\max_{\phi}\left(E_{V_S^f,V_S^p\sim P(V_S^f,V_S^p)}[d_\phi(V_S^f,V_S^p)]-E_{z\sim P(z)}[d_\phi(g_\theta(V_S^p,z),V_S^p)]\right) \tag{2}$$

The discriminator is also subject to a Lipschitz continuity constraint [11].

We wish to also incorporate information about scene stability into the learning process. We do so via a stability scoring network $s_\psi(V_S^f)$ which is trained to separate stable and unstable scenes via the Wasserstein distance to solve the objective:

$$\min_{\psi}\left(E_{V_S^f\sim P(V_S^f\,unstable)}[s_\psi(V_S^f)]-E_{V_S^f\sim P(V_S^f\,stable)}[s_\psi(V_S^f)]\right) \tag{3}$$

This maximises the separation of the scores for stable and unstable scenes, such that the scores for stable scenes are higher. The scoring network is subject to a Lipschitz continuity constraint [11].

After training $s_\psi(V_S^f)$ we calculate a normalisation operation so that the output of normalised $s_\psi(V_S^f)$ is of zero mean and has unit standard deviation when evaluated on its training data. This normalisation makes the output of $s_\psi(V_S^f)$ human-interpretable. Our full system modifies both the generator and discriminator networks to use the normalised output of $s_\psi(V_S^f)$ as additional conditioning, and the full loss becomes:

$$\begin{aligned}\min_{\theta}\max_{\phi}\Big(&E_{V_S^f,V_S^p\sim P(V_S^f,V_S^p)}[d_\phi(V_S^f,V_S^p,s_\psi(V_S^f))]\\&-E_{z\sim P(z)}[d_\phi(g_\theta(V_S^p,z,s_\psi(V_S^f)),V_S^p,s_\psi(g_\theta(V_S^p,z,s_\psi(V_S^f))))]\Big)\end{aligned} \tag{4}$$

Treating the stability signal as context rather than a separate regression objective allows the networks to learn to ignore the stability signal if that is advantageous, as it may be in the early stages of training. When used in the wild the stability context information provided to the generator can be sampled from the unit normal distribution due to the output normalisation of $s_\psi(V_S^f)$, or a value associated with either stability or instability. This allows an element of control over the stability of the imagined scenes.

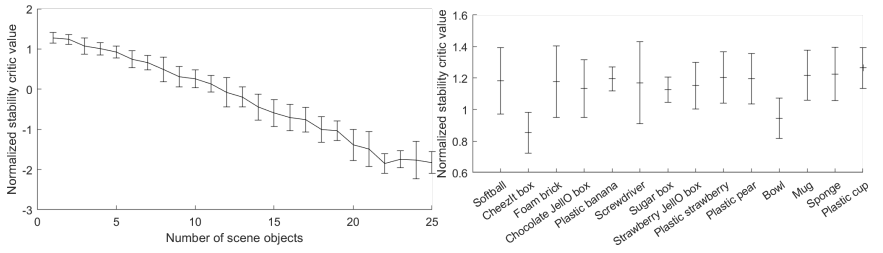
We train our full generative system incorporating stability context via eq. (4), and an ablated generative system without stability context via eq. (2). We also train a non-generative baseline via the regression objective:

$$\min_{\theta}\left\|g_\theta(V_S^p)-V_S^f\right\| \tag{5}$$

4.3 Architectures

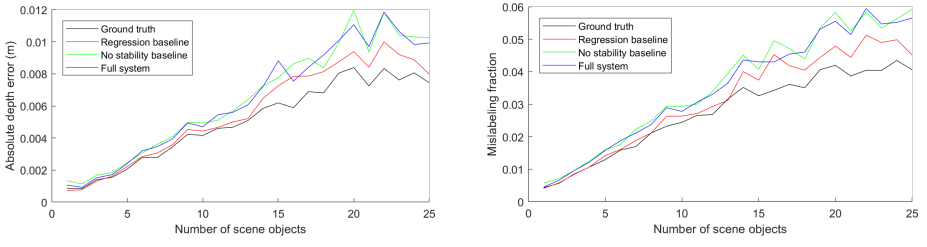
We implement all generators using a single U-Net [27] design consisting of 11 convolutional layers. The generators only differ in their input layers as the latent z and stability score $s_\psi(V_S^f)$ inputs require additional input channels. Both the discriminator network and stability scoring network are implemented via a modified DCGAN [25] discriminator. Unlike DCGAN, we choose not to use batch normalisation [14] due to the variance between the statistics of different batches, and instead use weight normalisation [28]. All networks use leaky rectified linear units [20] in hidden layers. We use the training hyperparameters proposed by Gulrajani *et al*. [11], including the scheme of 5 critic training batches per generator training batch, and the parameters of the Adam optimiser [15].

State-of-the-art computer vision architectures are undergoing a shift towards transformer-based [33] models, including for GANs [16]. Our aim is to compare different learning schemes and so we opt for mature architectures to avoid confounding results.



(a) Stability by scene complexity (b) Stability by object type

Figure 2: Mean stability scores over all scenes grouped according to the number of objects in each scene, and according to the presence of specific objects [4].



(a) Depth error (b) Mislabelling fraction

Figure 3: Depth error and pixel mislabelling fraction for ground truth and all explanation models, evaluated using object instance poses produced by the parsing process. Ground truth error corresponds to error resulting from the voxelisation and parsing processes applied to ground truth scene information.

5 Experiments

5.1 Stability scoring

The stability scoring network learns to separate stable and unstable scenes via a Wasserstein loss. After training for 25 epochs we examined normalised scores with respect to scene size, and object types as shown in fig. 2. We found that increasing scene size resulted in decreasing stability score until size 22 (fig. 2a), after which no further decrease was observed. This is consistent with the expectation that increasing heap complexity creates more potential for instability. We also found that most objects were associated with similar stability values with the exception of the two largest objects in the set: the CheezIt box, and the Bowl (fig. 2b). This makes sense, as these two objects are the most likely to be supporting other objects due to their size and shapes, and so the stability of scenes containing these objects should be sensitive to their existence and poses.

5.2 Visual consistency

We examined the visual consistency of the different systems in terms of the depth error and semantic segmentation mislabelling fraction of the imagined scenes with respect to the ground truth values, as shown in fig. 3. Here we observed that the baseline system trained via regression achieved a lower depth error and mislabelling fraction than the other systems,

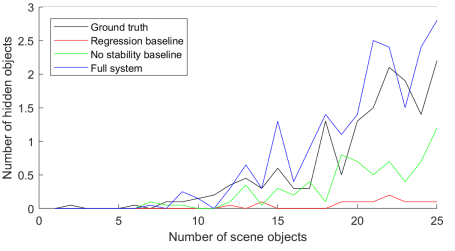
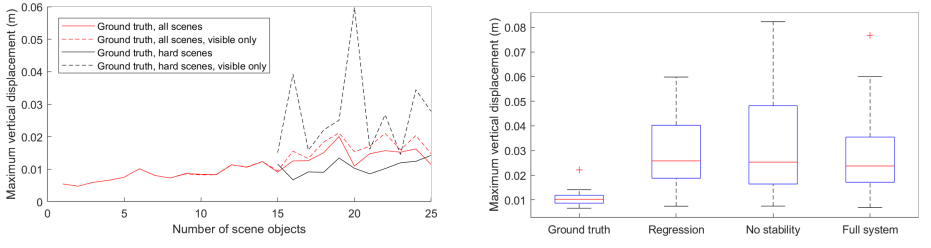


Figure 4: Mean number of hidden objects in ground truth scenes and mean number of imagined hidden objects produced by our full system and baselines. Scenes are grouped by the number of objects in the ground truth version of the scene.



(a) Instability from removing hidden objects

(b) Instability in imagined scenes

Figure 5: Instability for all explanation models tested on hard scenes. Lower object displacement indicates higher stability. Ground truth error corresponds to error resulting from the voxelisation and parsing processes applied to ground truth scene information.

while both generative systems performed similarly. This demonstrates one half of a trade-off between reconstructing the visible parts, and imagining the hidden parts of a scene (see section 5.3). We have also plotted error for ground truth poses which were voxelised and then parsed back to a pose set. These errors should be considered a bound on performance because both the voxelisation and parsing processes lose information and this process is applied to analysis of all systems. One potential solution for future work would be to train a neural network system to perform the parsing process. We do not compare directly against the voxelised ground truth scene because the goal is to produce scenes which are visually and physically *plausible* and there may be multiple plausible scenes for a given set of visual information.

5.3 Imagined objects

We examined the number of hidden objects present in the imagined scenes, and the effect on stability of removing these hidden objects, as shown in fig. 4. Here we see a large difference between regression and generative training. The regression baseline does not imagine hidden objects. The generative baseline does imagine hidden objects, but fewer than the ground truth. Only the system conditioned on stability produces results similar to the ground truth. These differences are most acute for scene sizes above 15 objects.

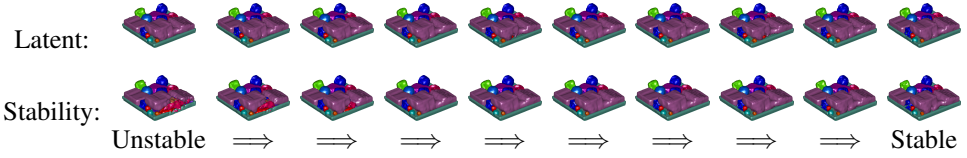


Figure 6: The effect of varying the latent input vector, and of varying the stability input signal. The first row shows the effect of different random samples of the latent vector. The second row shows the effect of varying the stability score, where scores go from very unstable to very stable moving from left to right. This figure is best viewed electronically.

5.4 Imagined scene stability

We identified the scenes containing hidden objects which were important to scene stability, in the sense that their removal caused other objects to move (see fig. 5a), and tested the systems on this subset (see fig. 5b). We found that voxel conversions and parsing operations on the ground truth data induced a small amount of instability, but far below that of any of the systems. Instability can result from objects having insufficient support due to poor placement or a lack of supporting objects, and can also result from multiple objects being placed such that they partially occupy the same space. All of these cases are physically implausibly. Both baselines exhibit higher vertical displacements than the full system (the distance that objects fall), suggesting that stability context information is useful for biasing generation towards stable scenes.

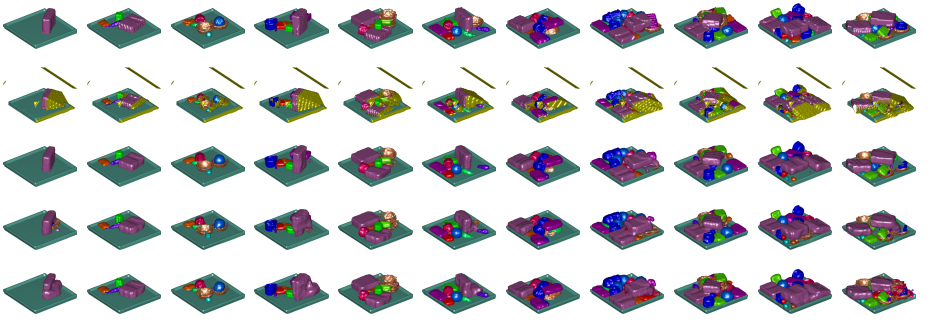
5.5 Generator conditioning

Finally, we examined the effect of both the latent vector and stability score on the output of the full system. In fig. 6 we show examples of the effect of different input values on the imagined scene. We see that the latent vector seems to encode local noise and has little effect on the large-scale structure. Conversely, changing the stability score causes large-scale changes to supporting objects. This demonstrates that stability is important to scene imagination.

5.6 Qualitative performance

We include a selection of reconstructed scenes in fig. 7 to illustrate qualitative performance. We observe the same general trends: the regression baseline performs best for simple scenes but does not produce supporting objects. The generative baseline without stability context is qualitatively similar to the full system but the differences become clear under quantitative examination and stability simulation. Note that in many cases the systems imagine a *different* set of hidden objects to those present in the ground truth scene.

The generative systems have a tendency to produce object fragments. GANs have a known weakness in global consistency; one potential solution would be to use a state-of-the-art GAN architecture *in conjunction* with training in a big data regime.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Figure 7: Examples of scenes imagined by the three explanation models. Each column represents a different scene. The first row shows ground truth scenes. The second row shows occluded regions in gold colour. The third row shows the regression baseline. The fourth row shows the generative baseline without stability supervision. The fifth row shows the full system. In general the regression baseline does not produce hidden objects for scenes where they are not needed for stability *and* for scenes where they *are* needed for stability. The generative baseline without stability supervision produces hidden objects, but which are often different to those produced by the full system. Quantitative examination shows that the hidden objects produced by the generative baseline result in less stable scenes than those produced by the full system (see section 5.4). This figure is best viewed electronically and a larger version of this figure, containing two scenes per page, can be found in supplementary material.

6 Conclusions

In this work we presented a novel task and dataset for imagining hidden objects in scenes. We found that a generative approach is important to imagining hidden objects, but that there is a trade-off with visual consistency. We also demonstrated that learning from stability is important to producing imagined hidden objects.

The fidelity of the voxel representation is an area in which the current work can be improved. This could take the form of an increase in resolution, or the use of a sparse representation to assign capacity only where it is needed. Adopting an object instance-based representation (such as a set of object poses) offers the greatest potential but also brings with it challenges such as making object shapes implicit, and providing stability supervision on adding and removing objects *without extensive object candidate sampling and the associated computational cost*. We hope that this work and dataset stimulates future research in the community.

7 Acknowledgements

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. The research in this paper was supported in part by the Engineering and Physical Sciences Research Council (grant number EP/S032487/1).

References

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

[2] Daniel Bear, Elias Wang, Damian Mrowca, Felix Jedidja Binder, Hsiao-Yu Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin A Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B Tenenbaum, Daniel L K Lamins, and Judith E Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[3] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791.

[4] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics Automation Magazine*, 22(3):36–52, 2015.

[5] Erwin Coumans et al. Bullet physics library. *Open source: bulletphysics. org*, 15:49, 2013.

[6] Yilun Du, Zhijian Liu, Hector Basevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, and Jiajun Wu. Learning to exploit stability for 3d scene parsing. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 1726–1736. Curran Associates, Inc., 2018.

[7] Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J. Brostow. Structured prediction of unobserved voxels from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5431–5440, 2016.

[8] Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, and Olivier Bachem. Brax - a differentiable physics engine for large scale rigid body simulation. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[9] Georgia Gkioxari, Nikhila Ravi, and Justin Johnson. Learning 3d object shape and layout without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1695–1704, June 2022.

[10] Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *European Conference on Computer Vision*, pages 702–717, 2018.

[11] 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.

[12] Ruiqi Guo, Chuhang Zou, and Derek Hoiem. Predicting complete 3d models of indoor scenes. *CoRR*, abs/1504.02437, 2015.

[13] Drew A Hudson and C. Lawrence Zitnick. Compositional transformers for scene generation. *Advances in Neural Information Processing Systems*, 2021.

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning-Volume 37*, pages 448–456. JMLR. org, 2015.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[16] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. In *Proceedings of the International Conference on Learning Representations*, 2022.

[17] 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.

[18] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2036–2043, 2009. doi: 10.1109/CVPR.2009.5206718.

[19] Wenbin Li, Aleš Leonardis, and Mario Fritz. Visual stability prediction for robotic manipulation. In *2017 IEEE International Conference on Robotics and Automation*, pages 2606–2613. IEEE, 2017.

[20] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.

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

[23] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

[24] Joni Pajarinen, Jens Lundell, and Ville Kyrki. POMDP manipulation planning under object composition uncertainty. *CoRR*, abs/2010.13565, 2020.

[25] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *The International Conference on Learning Representations*, 2015.

[26] Ronan Alexandre Riochet, Mario Ynocente Castro, Mathieu Bermard, Adam Lerer, Rob Fergus, Veronique Izard, and Emmanuel Dupoux. Intphys: A benchmark for visual intuitive physics reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. doi: 10.1109/TPAMI.2021.3083839.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[28] Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 901–909, 2016.

[29] 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.

[30] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. *Indoor Segmentation and Support Inference from RGBD Images*, pages 746–760. Springer Berlin Heidelberg, 2012.

[31] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] Tomer D Ullman, Eliza Kosoy, Ilker Yildirim, Amir Arsalan Soltani, Max Siegel, Joshua B. Tenenbaum, and Elizabeth S Spelke. Draping an elephant: Uncovering children's reasoning about cloth-covered objects. In *CogSci The Annual Meeting of the Cognitive Science Society*, 2019.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[34] R Williams. A class of gradient-estimation algorithms for reinforcement learning in neural networks. In *Proceedings of the International Conference on Neural Networks*, pages II–601, 1987.

[35] 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*, pages 153–164. Curran Associates, Inc., 2017.

[36] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[37] Ming-Jia Yang, Yu-Xiao Guo, Bin Zhou, and Xin Tong. Indoor scene generation from a collection of semantic-segmented depth images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[38] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[39] Yizhou Zhao, Kaixiang Lin, Zhiwei Jia, Qiaozi Gao, Govind Thattai, Jesse Thomason, and Gaurav S. Sukhatme. Luminous: Indoor scene generation for embodied ai challenges. In *NeurIPS 2021 Workshop on CtrlGen*, 2021.

[40] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, Joshua B. Tenenbaum, and Song-Chun Zhu. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. ISSN 2095-8099. doi: https://doi.org/10.1016/j.eng.2020.01.011.