CLEVRTEX: A TEXTURE-RICH BENCHMARK FOR UNSUPERVISED MULTI-OBJECT SEGMENTATION

LAURYNAS KARAZIJA IRO LAINA CHRISTIAN RUPPRECHT

WWW.ROBOTS.OX.AC.UK/~VGG/RESEARCH/CLEVRTEX

UNIVERSITY OF OXFORD

Summary (TL;DR)

Recently, multiple models aimed at unsupervised multi-object segmentation have been proposed. They rely on simpler synthetic data. Scaling to real-world data remains an open problem.

- We introduce a new dataset and benchmark, CLEVRTEX, featuring complex, diverse shapes and 60 photo-mapped materials.
- Create an extra out-of-distribution test set featuring different shapes and 25 new materials.
- Benchmark recent models on CLEVR[7] and CLEVRTEX.
- Design dataset variants controlling for different aspects of scene complexity and probe current approaches for shortcomings.

Key Findings

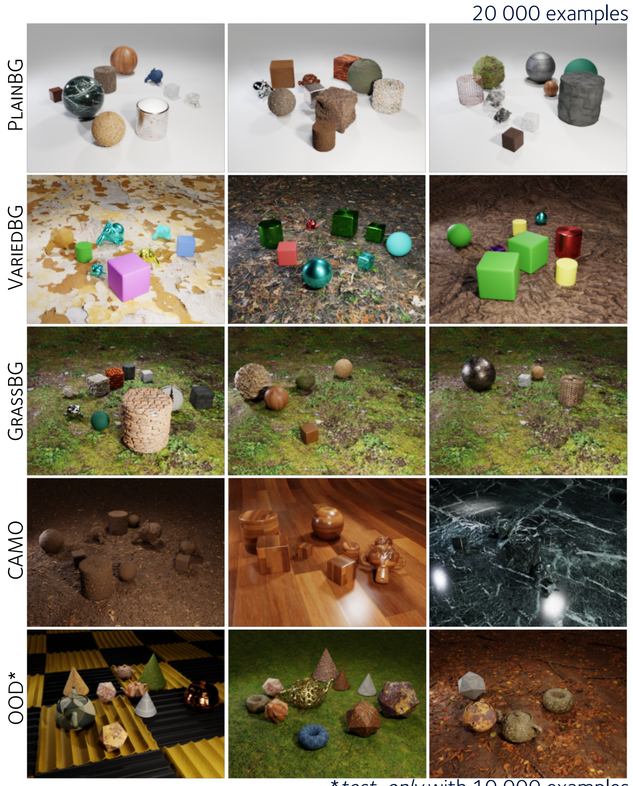
- Despite impressive performance on CLEVR, no method achieves satisfactory performance on CLEVRTEX.
- No approaches exploit global context cues for objects.
- Most models *rely* on consistency in object appearances and, thus, struggle with textures.
- All methods overfit overall scene appearance, missing small objects or recognising background patterns as foreground.

CLEVRTEX Dataset

50 000 examples

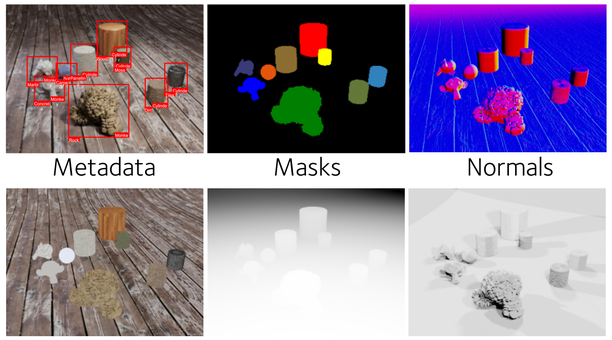
Variants

20 000 examples

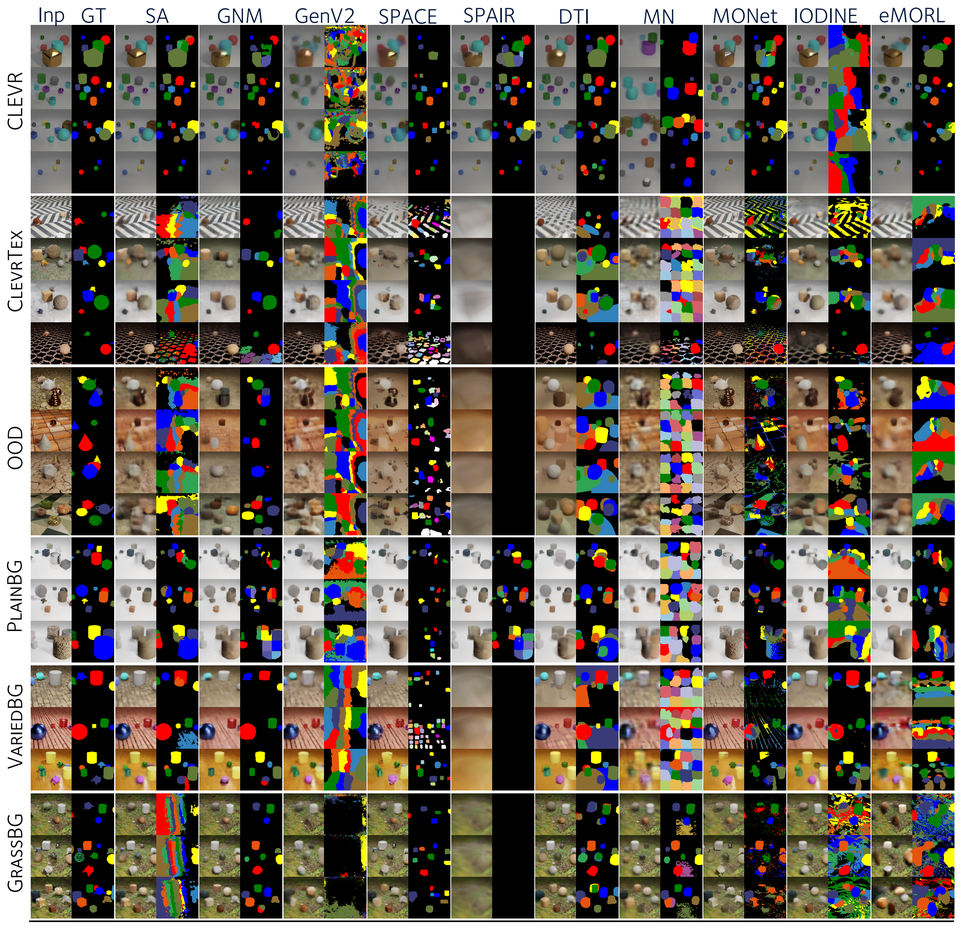


test-only with 10 000 examples

Annotations



Benchmark Results



Model	CLEVR[7] ↑mIoU (%)	CLEVR[7] ↓MSE	CLEVRTEX ↑mIoU (%)	CLEVRTEX ↓MSE	OOD ↑mIoU (%)	OOD ↓MSE	CAMO ↑mIoU (%)	CAMO ↓MSE
SPAIR[2]	65.95± 4.02	55± 10	0.0 ± 0.0	1101± 2	0.0 ± 0.0	1166± 5	0.0 ± 0.0	668± 3
SPACE[8]	26.31±12.93	63± 3	9.14± 3.46	298± 80	6.87± 3.32	387± 66	8.67± 3.50	251± 61
GNM[6]	59.92± 3.72	43± 3	42.25± 0.18	383± 2	40.84± 0.30	626± 5	17.56± 0.74	353± 1
MN[11]	56.81± 0.40	75± 1	10.46± 0.10	335± 1	12.13± 0.19	409± 3	8.79± 0.15	265± 1
DTI[10]	48.74± 2.17	77± 12	33.79± 1.30	438± 22	32.55± 1.08	590± 4	27.54± 1.55	377± 17
GenV2[4]	9.48± 0.55	158± 2	7.93± 1.53	315±106	8.74± 1.64	539±147	7.49± 1.67	278± 75
eMORL[3]	50.19±22.56	33± 8	12.58± 2.39	318± 43	13.17± 2.58	471± 51	11.56± 2.09	269± 31
MONet[1]	30.66±14.87	58± 12	19.78± 1.02	146± 7	19.30± 0.37	231± 7	10.52± 0.38	112± 7
SA[9]	36.61±24.83	23± 3	22.58± 2.07	254± 8	20.98± 1.59	487± 16	19.83± 1.41	215± 7
IODINE[5]	45.14±17.85	44± 9	29.16± 0.75	340± 3	26.28± 0.85	504± 3	17.52± 0.75	315± 3

Results average of 3 runs, shown ±σ.

References

[1] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. arXiv preprint arXiv:1901.11390, 2019.
[2] Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3412–3420, 2019.
[3] Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations. In Proceedings of the 38th International Conference on Machine Learning, pages 2970–2981. PMLR, 2021.
[4] Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. Genesis-v2: Inferring unordered object representations without iterative refinement. In Advances in Neural Information Processing Systems, 2021.
[5] Klaus Greff, Raphaeel Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, DanielZoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In International Conference on Machine Learning, pages 2424–2433. PMLR, 2019.
[6] Jindong Jiang and Sungjin Ahn. Generative neurosymbolic machines. In Advances in Neural Information Processing Systems, volume 33, pages 12572–12582, 2020.
[7] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2901–2910, 2017.
[8] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. SPACE: Unsupervised object-oriented scene representation via spatial attention and decomposition. In International Conference on Learning Representations, 2020.
[9] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In Advances in Neural Information Processing Systems, volume 33, pages 11525–11538, 2020.
[10] Tom Monnier, Elliot Vincent, Jean Ponce, and Mathieu Aubry. Unsupervised Layered Image Decomposition into Object Prototypes. ICCV, 2021.
[11] Dmitriy Smirnov, Michael Gharbi, Matthew Fisher,Vitor Guizilini, Alexei A Efros, and Justin Solomon. Marionette: Self-supervised sprite learning. In Advances in Neural Information Processing Systems, 2021.

Engineering and Physical Sciences Research Council

EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems

NEURAL INFORMATION PROCESSING SYSTEMS