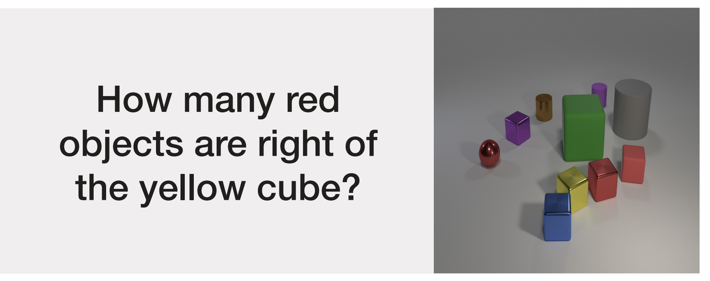
# Transparency by Design:
# Closing the Gap Between Performance and Interpretability in Visual Reasoning

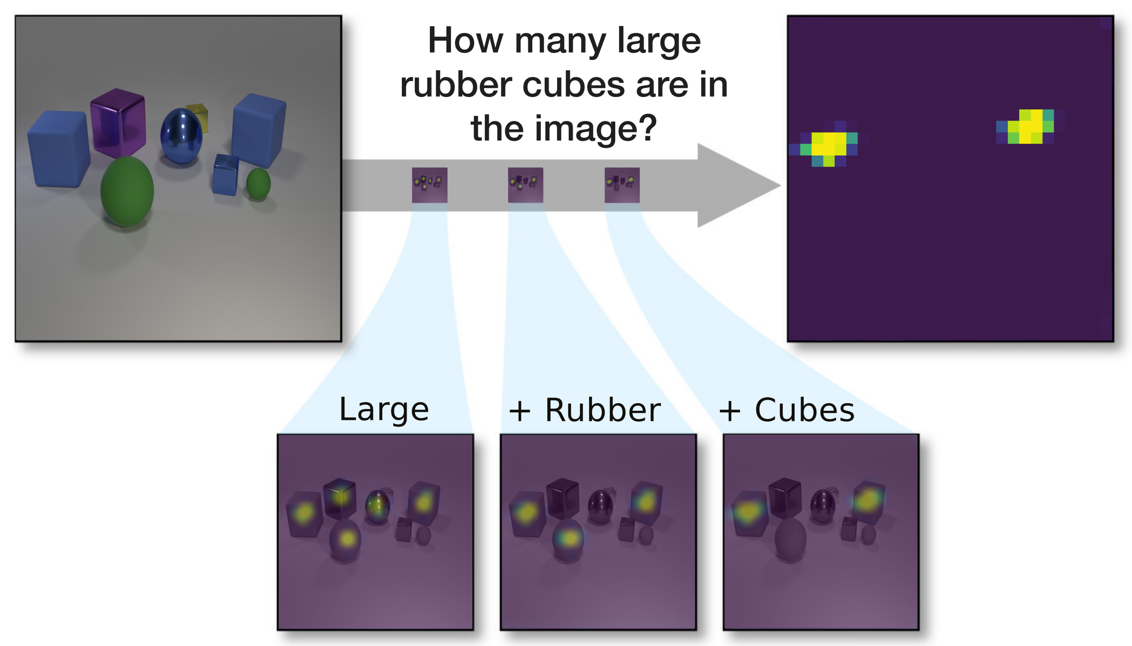
David Mascharka, Philip Tran‡, Ryan Soklaski, Arjun Majumdar | MIT Lincoln Laboratory†

## Overview

**Visual Question Answering involves determining the correct answer for a given question-image pair**

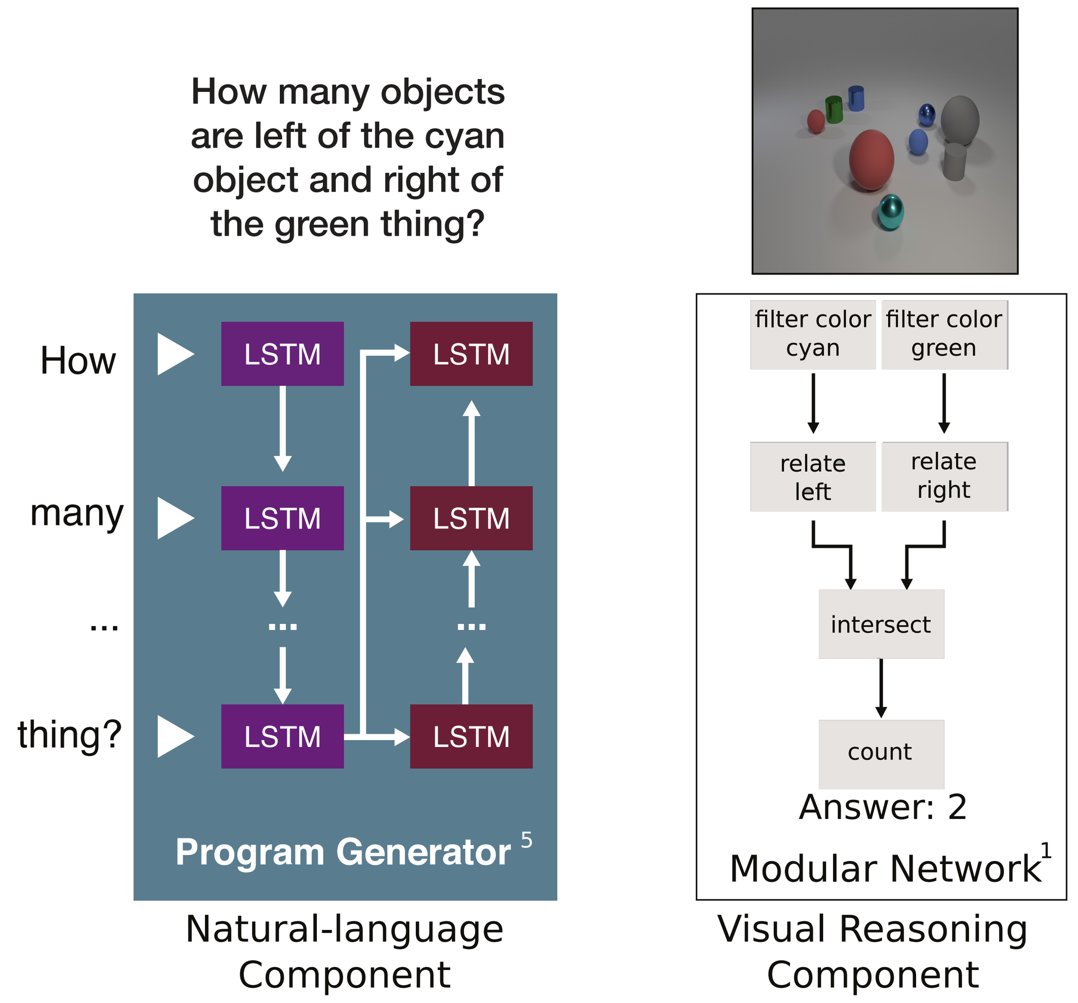



**Unlike existing methods, TbD-nets leverage attention masks that are explicitly grounded in visual primitives.**




## Related Work

- Andreas et al. [1] introduced a method that combines a natural-language parser with reusable neural "modules" to compose question-specific neural module networks (NMNs)
- Early NMNs [1, 3] produced interpretable outputs using visual attention masks, but struggled to achieve good performance
- By improving the natural-language parser and developing modules that process high-dimensional features rather than attentions, Johnson et al. [5] significantly improved performance at the cost of interpretability



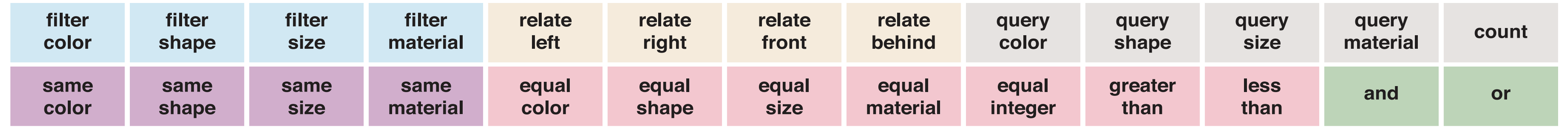

‡ This work was conducted while Philip was at MIT Lincoln Laboratory. Philip is now at Planck Aerosystems.

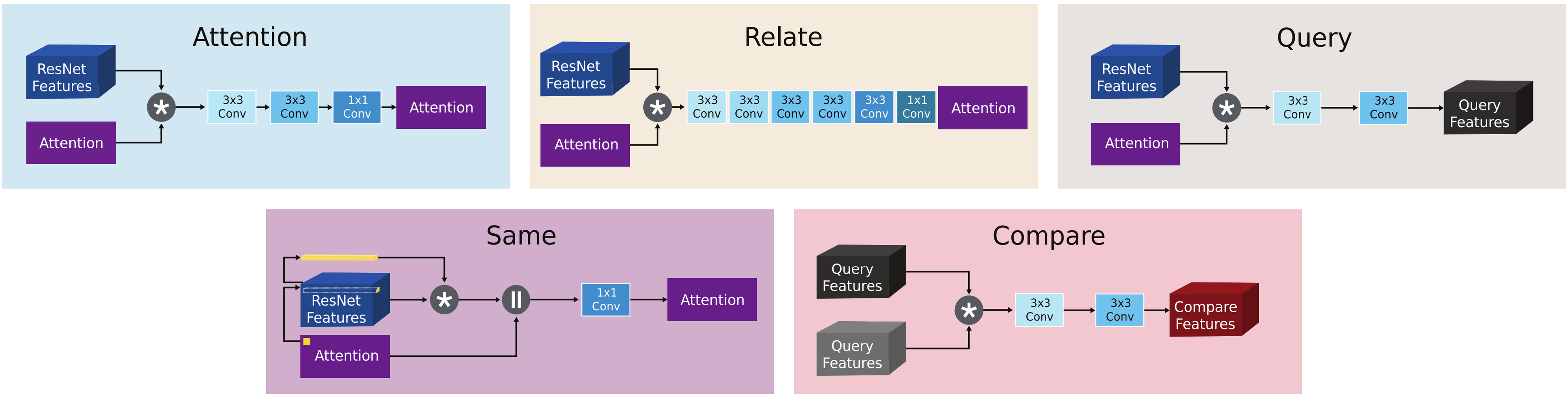
† **DISTRIBUTION STATEMENT A.** Approved for public release. Distribution is unlimited. This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.



## Transparency by Design Networks

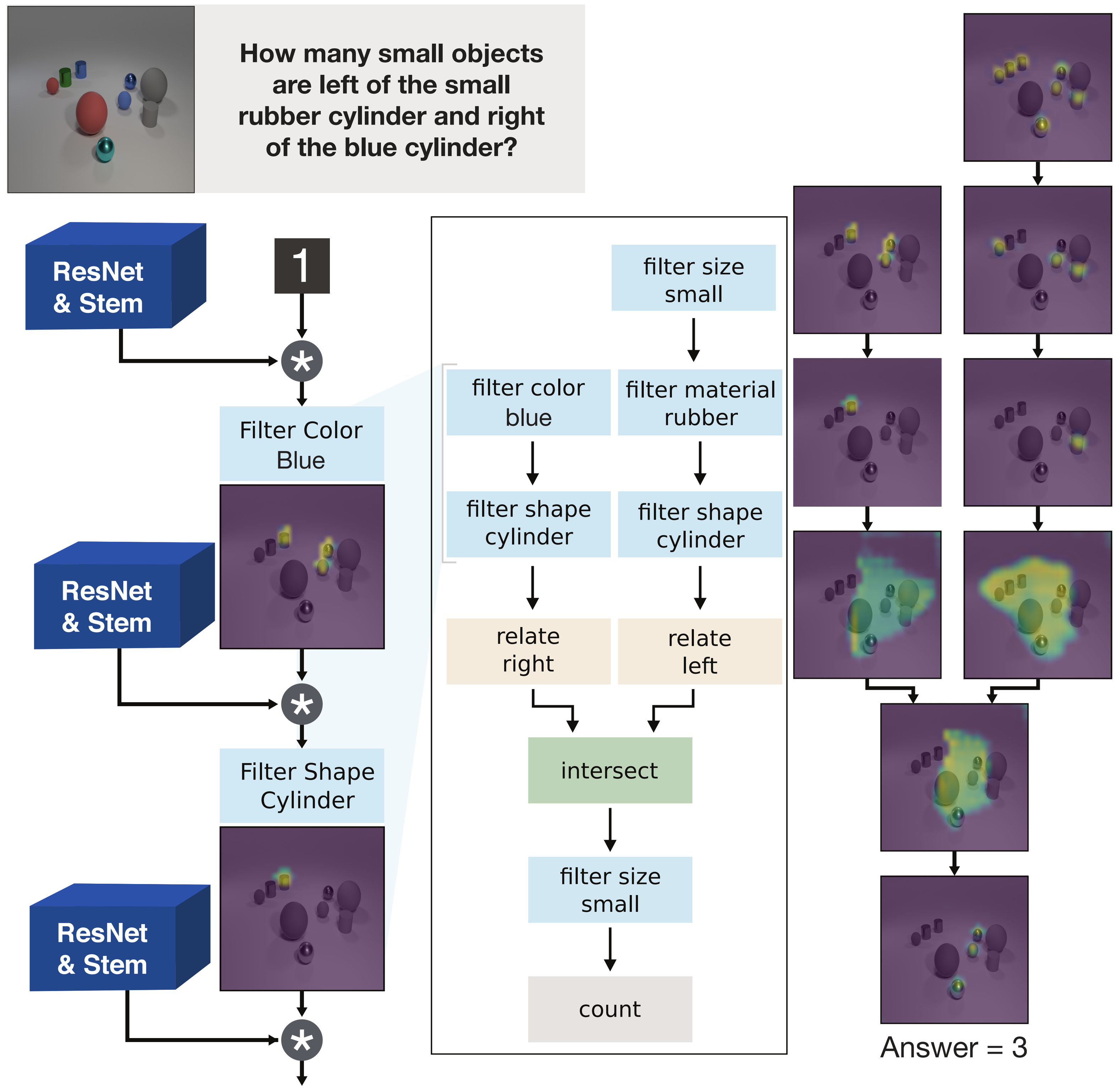
- Transparency by Design networks (TbD-nets) are built to achieve the performance of black-box models while surpassing the interpretability of initial NMNs by specializing each module type




- Our approach reuses the program generator from [5] and focuses on improving the visual reasoning component to yield highly performant and interpretable modules
- The visual reasoning component is comprised of modules which operate on and produce visual attentions
- Each module is designed to perform spatial transformations on visual attention to suit its specific task

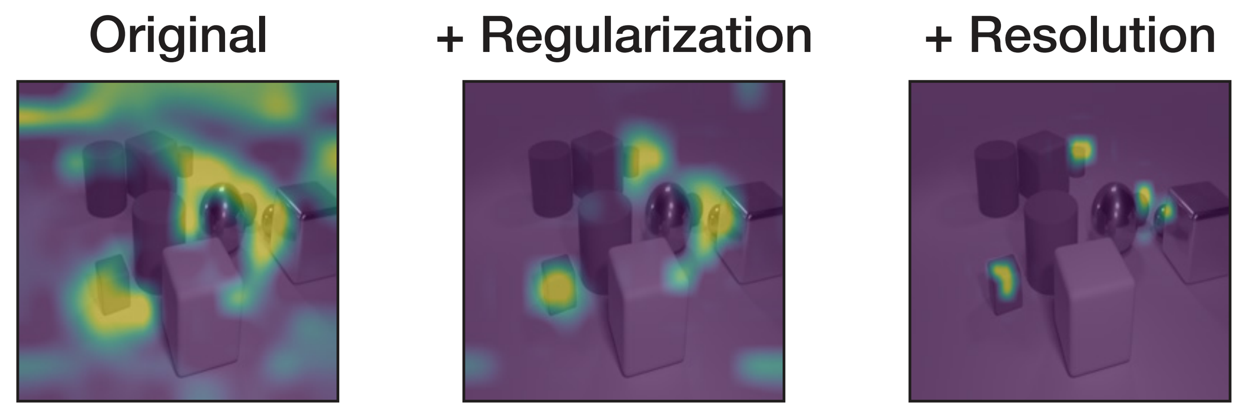
### TbD Visual Reasoning Component




## Results on Main Task

- We evaluate our model on the CLEVR dataset [4], a visual reasoning benchmark comprised of synthetic scenes containing 3D shapes
- We achieve state-of-the-art 99.1% accuracy on CLEVR with σ=0.07

| Model | Overall |
|---|---|
| NMN [1] | 72.1 |
| N2NMN [3] | 88.8 |
| Human [4] | 92.6 |
| PG + EE (700k) [5] | 96.9 |
| CNN + GRU + FiLM [6] | 97.6 |
| MAC [2] | 98.9 |
| TbD-net (Ours) | 98.7 |
| TbD + regularization | 98.5 |
| TbD + regularization + resolution | **99.1** |




### Quantifying Interpretability

| | Original | +Regularization | +Resolution |
|---|---|---|---|
| Correct-object recall | 0.86 | 0.92 | 0.99 |
| Correct-object precision | 0.41 | 0.90 | 0.98 |

- Adding regularization and increasing the spatial resolution reduces the noise in and improves localization of the attentions
- Specifically, we measure the center-of-mass overlap of the attentions with the ground-truth regions

## Results on Generalization Task

| | Train A | | Fine-tune B | |
|---|---|---|---|---|
| | A | B | A | B |
| PG + EE [5] | 96.6 | 73.7 | 76.1 | 92.7 |
| TbD + reg (Ours) | 98.8 | 75.4 | 96.9 | 96.3 |

- The Compositional Generalization Test (CoGenT) evaluates generalizability to new color/shape combinations
- While our model learns entangled representations of color and shape (Train A), we quickly recover performance fine-tuning on a small amount of data (Fine-tune B)

### Quantifying Entanglement

| | Predict Shape | | Predict Color | |
|---|---|---|---|---|
| | $P(\checkmark \mid A)$ | $P(\checkmark \mid B)$ | $P(\checkmark \mid A)$ | $P(\checkmark \mid B)$ |
| Train A | 0.90 | 0.22 | 0.91 | 0.84 |
| Fine-tune B | 0.77 | 0.81 | 0.90 | 0.86 |

- We find that our model's representation of shape is entangled with color (Predict Shape A), but its color representation is not entangled with shape (Predict Color A)
- Fine-tuning on a small amount of data rectifies the entanglement (Fine-tune B)

**Code available at github.com/davidmascharka/tbd-nets**

### References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. CoRR, abs/1511.02799, 2015.
[2] D. Hudson and C. Manning. Compositional attention networks for machine reasoning. International Conference on Learning Representations, 2018.
[3] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. CoRR, abs/1704.05526, 2017.
[4] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. CoRR, abs/1612.06890, 2016.
[5] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. In ICCV, 2017.
[6] E. Perez, H. de Vries, F. Strub, V. Dumoulin, and A. Courville. FiLM: Visual Reasoning with a General Conditioning Layer. ArXiv e-prints, December 2017.