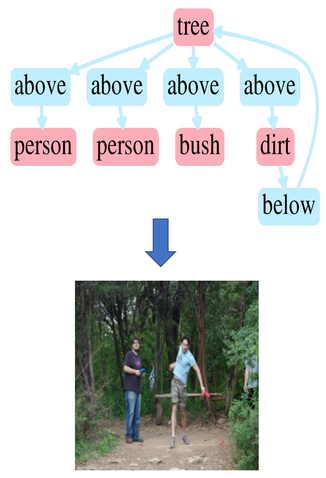
Heuristics for Image Generation from Scene Graphs

Subarna Tripathi, Anahita Bhiwandiwalla, Alexei Bastidas, Hanlin Tang
Intel AI Lab

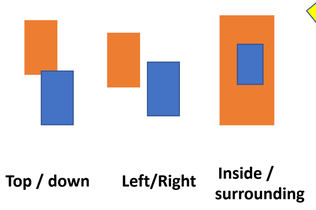
Summary



Scene Graph to Image
- Scene Graphs provide structured descriptions of complex scenes
- Objects as nodes, and edges as relationship between objects
- Compact alternative for communicating 'gist'

Challenges

1. Existing SG datasets:
   - Visual Genome with incomplete and incorrect annotations
   - Artificial dataset such as COCO stuff with limited spatial relationships



2. Lack of metrics
   - measure scene graph compliance
3. Scene Information
   - Better integration into the model

We Propose

1. Heuristics based **Relationship augmentation**
   - thing vs thing
   - thing vs stuff
   - stuff vs stuff
2. Scene Compliance **Metrics**
   - Relation Score
   - Mean Opinion Relation Score (MORS)
3. Auxiliary network to encode **scene context** for the generator and the discriminator

Scene Graph Compliance Metrics

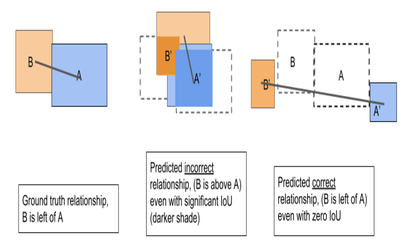


Figure: Relation Score vs Intersection over Union (IoU). Left: Example ground truth relationship between two objects **A** and **B**. Center: predicted relationship between the objects A and B in which B is above A. Note that the high IoU here does not guarantee compliance with the intended geometric relationship. Intersection with ground truth is highlighted with darker shades. Right: A predicted relationship between the objects A and B in which the two objects have no overlap. Note that despite the IoU being 0 here, the predicted relationship could still be compliant with the ground truth relationship.

Metric	JJ [9]	SC	DA
Average IoU	45.9%	48.3%	51.0%
Relation Score	51.2%	53.6%	59.8%

Table: Relationship compliance. Relation score (the higher the better) on COCO stuff test set. SC= Scene Context, DA = heuristic based data augmentation

Relation Score
for spatial relationships

Relation category	Relations
Semantic	covering, eating, standing on, carrying, looking at, walking on, sitting on, sitting in, standing in, holding, riding
Geometric	next to, above, beside, behind, by, laying on, hanging on, under, on, below, against, attached to, near, on top of, at, in front of, around, along, on side of, parked on
Possessive	has, belonging to, inside, with, over, covered in, have, in, wears, wearing
Miscellaneous	and, for, of, made of

Table: Relationship categories observed in Visual Genome

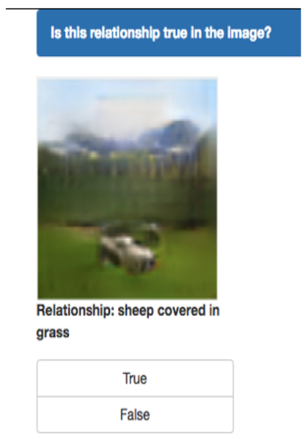


Figure: Example scene graph from the Visual Genome, with relationships between objects colored according to their category: Geometric (green), Possessive (orange) and Miscellaneous (grey).

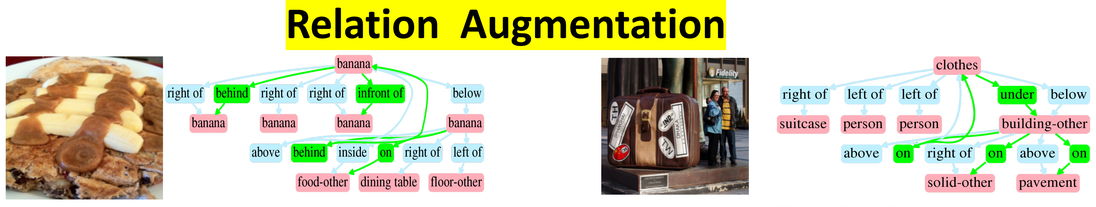
Relation category	JJ [9]	Ours
Semantic	0.60	**0.78**
Geometric	0.64	**0.68**
Possessive	0.62	**0.80**
Miscellaneous	0.78	**0.86**
Overall MORS	0.64	**0.74**
Avg IOU	0.223	**0.234**

Figure: Mean Opinion Relation Score (MORS) on 100 images and relationship pairs generated from the colored scene graphs in Visual Genome test set. The score is broken by relation category. Each image was rated by five workers. IoU corresponds to all predicted boxes in the test set.

Mean Opinion Relation Score (MORS)
for complex *semantic* relationships

Methodology - Key Contributions

Relation Augmentation

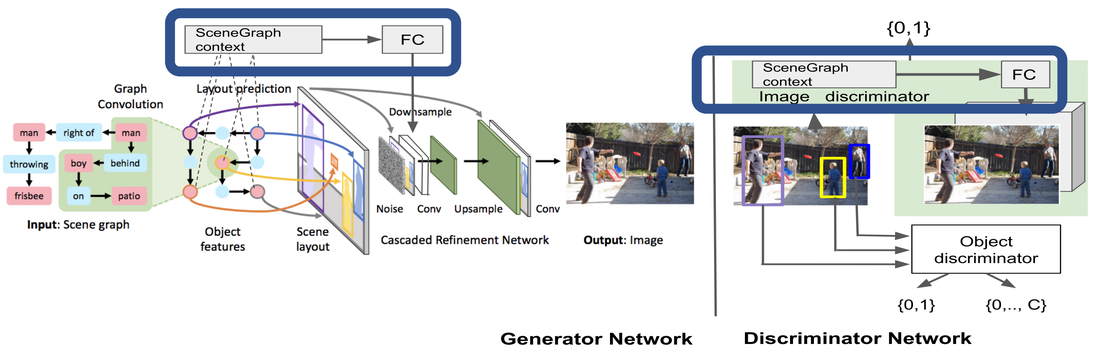


Instance-Instance (behind and in front of)

Instance-stuff / stuff-stuff (on and under)

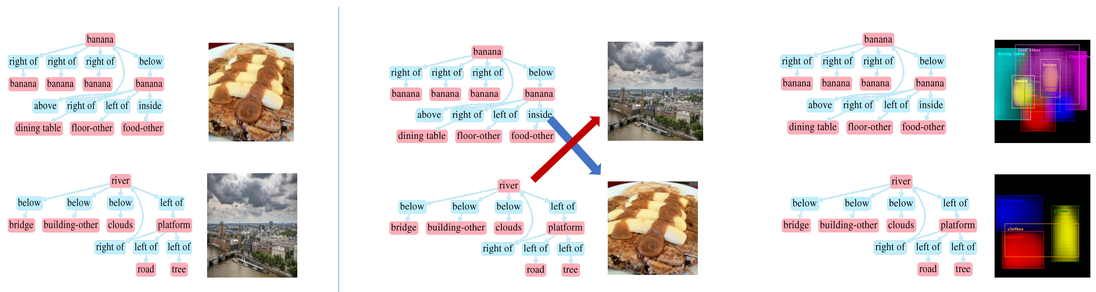
Figure: Heuristics based data augmentation outperforms the layout prediction for incomplete scene graph inputs. Model learned without the data augmentation notably struggles with placing objects apart when the input scene graph does not specify explicit relationships between those (Row 1). Data augmentation helps in learning better inter-class objects placements (e.g. person and cloths) (Row 2)

Scene Context Net



* Figure adapted from [9]

Matching-Aware GAN

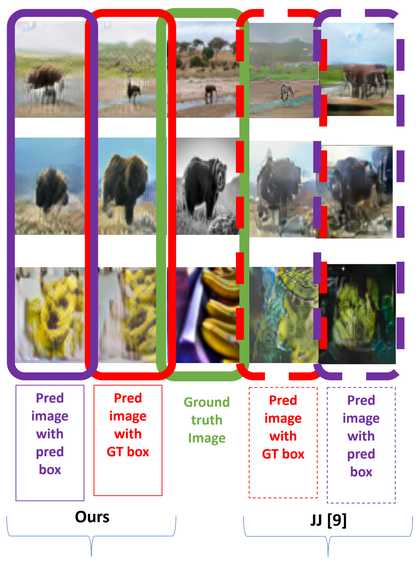


Subjective Quality Evaluation

AB-X (COCO caption)	JJ [9]	Ours
GT box and GT mask	39.5%	60.5%
Predicted box and mask	63.8%	36.2%

AvB	JJ [9]	Ours
GT box and GT mask	36.9%	63.1%
Predicted box and mask	67.8%	32.2%

COCO stuff



Study	JJ [9]	Ours
AB-X (Caption)	42.6%	57.4%
AvB	42.7%	58.3%

Table: **Visual Genome** qualitative study results. Our model outperformed Johnson et al. when workers were asked to select which image was more realistic (AvB), and which image better matched a provided pseudo-caption (AB-X).

* Ref [9]: Justine Johnson, Agrim Gupta, Fei-Fei Li, Image Generation from Scene Graphs, CVPR 2018

Experimental Details

Experiment	Dataset	Image Type	# Images	Total # Trials	% Correct (Control Trials)
AvB	COCO	GT Mask, GT BB	991	4960	97.2%
AvB	COCO	Pred Mask, Pred BB	300	1380	91.6%
AB-X	COCO	GT Mask, GT BB	991	3862	94.5%
AB-X	COCO	Pred Mask, Pred BB	300	1380	93.2%
AvB	Visual Genome	Pred Mask, Pred BB	1018	2964	83.3%
AB-X	Visual Genome	Pred Mask, Pred BB	200	500	83.6%
MORS	Visual Genome	Pred Mask, Pred BB	200	1000	86.4%

Scan me

© 2019 Intel Corporation. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.