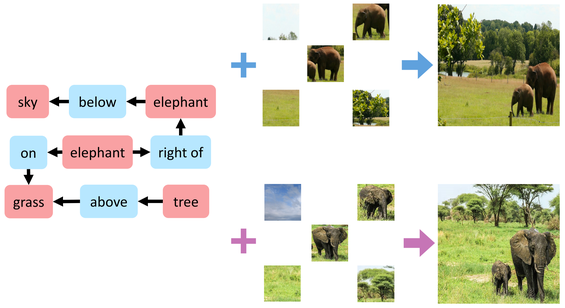
# PasteGAN: A Semi-Parametric Method to Generate Image from Scene Graph

Yikang Li[1*], Tao Ma[2*], Yeqi Bai[3], Nan Duan[4], Sining Wei[4], Xiaogang Wang[1]
[1]The Chinese University of Hong Kong, [2]Northwestern Polytechnical University,
[3]Nanyang Technological University, [4]Microsoft

## Motivation

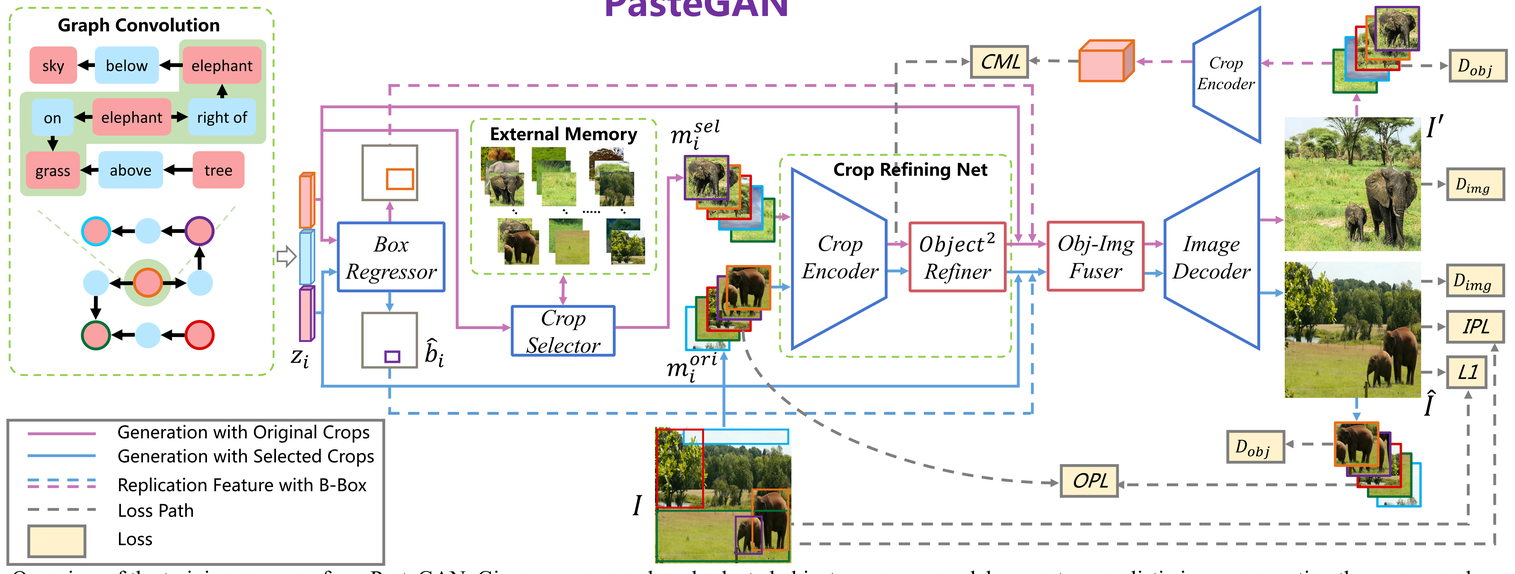



- Despite some exciting progress on high-quality image generation from structured (scene graphs) or free-form (sentences) descriptions, most of them only guarantee the image-level semantical consistency, i.e. the generated image matching the semantic meaning of the description.
- They still lack the investigations on synthesizing the images in a more controllable way, like finely manipulating the visual appearance of every object.

## Contribution

- A semi-parametric method, PasteGAN, is proposed to generate realistic images from a scene graph, which uses the external object crops as anchors to guide the generation process.
- A scene-graph-guided Crop Refining Network and an attention based Object-Image Fuser are proposed to reconcile the isolated crops into an integrated image, to make the objects in crops appear on the final image in the expected way.
- A crop selector is also introduced to automatically pick the most-compatible crops from our object tank by encoding the interactions around the objects in the scene graph.

## PasteGAN




Overview of the training process of our PasteGAN. Given a scene graph and selected object crops, our model generates a realistic image respecting the scene graph and the appearance of selected object crops. The training process involves two branches, bottom branch aims to reconstruct the ground-truth image using the original crops $m_i^{ori}$ (the bottom branch), the top branch focuses on generating the diversified images with the crops $m_i^{sel}$ retrieved from the external memory tank (the top branch). L1, CML, IPL and OPL mean image reconstruction loss, crop matching loss, image perceptual loss and object perceptual loss respectively.

## Qualitative Results


## Experiment Results

❑ Statistics of COCO-Stuff and VG dataset. # Obj. denotes the number of object categories. # Crops denotes the number of crops in the external memory.

| Dataset | COCO | VG |
|---|---|---|
| Train | 74 121 | 62 565 |
| Val. | 1 024 | 5 506 |
| Test | 2 048 | 5 088 |
| # Obj. | 171 | 178 |
| # Crops | 411 682 | 606 319 |

❑ Ablation Study using Inception Score (IS ) and Fréchet Inception Distance (FID) on COCO-Stuff dataset.

| Method | IS ↑ | FID ↓ |
|---|---|---|
| Real Images | $16.3 \pm 0.4$ | - |
| w/o *Crop Selection* | $7.1 \pm 0.3$ | 96.75 |
| w/o *Object² Refiner* | $8.3 \pm 0.3$ | 61.28 |
| w/o *Obj-Img Fuser* | $8.7 \pm 0.2$ | 56.14 |
| full model | $9.1 \pm 0.2$ | 50.94 |
| full model (GT) | $10.2 \pm 0.2$ | 38.29 |

❑ Comparison with existing methods on COCO-Stuff (COCO) Visual Genome (VG) Datasets. Please check our paper for detailed experiment results and references.

| Method | Inception Score ↑ COCO | Inception Score ↑ VG | Diversity Score ↑ COCO | Diversity Score ↑ VG | FID ↓ COCO | FID ↓ VG |
|---|---|---|---|---|---|---|
| Real Imgs | $16.3 \pm 0.4$ | $13.9 \pm 0.5$ | - | - | - | - |
| sg2im | $6.7 \pm 0.1$ | $5.5 \pm 0.1$ | $0.02 \pm 0.01$ | $0.12 \pm 0.06$ | 82.75 | 71.27 |
| PasteGAN | $\mathbf{9.1 \pm 0.2}$ | $\mathbf{6.9 \pm 0.2}$ | $\mathbf{0.27 \pm 0.11}$ | $\mathbf{0.24 \pm 0.09}$ | **50.94** | **58.53** |
| sg2im (GT) | $7.3 \pm 0.1$ | $6.3 \pm 0.2$ | $0.02 \pm 0.01$ | $0.15 \pm 0.12$ | 63.28 | 52.96 |
| layout2im | $9.1 \pm 0.1$ | $8.1 \pm 0.1$ | $0.15 \pm 0.06$ | $0.17 \pm 0.09$ | - | - |
| PasteGAN (GT) | $\mathbf{10.2 \pm 0.2}$ | $\mathbf{8.2 \pm 0.2}$ | $\mathbf{0.32 \pm 0.09}$ | $\mathbf{0.29 \pm 0.08}$ | **38.29** | **35.25** |