Image Generation from Layout

Bo Zhao Lili Meng Weidong Yin Leonid Sigal
University of British Columbia
{bzhao03, menglili, wdyin, lsigal}@cs.ubc.ca

CVPR 2019 (Oral)

2019.09.24

발표자 : 김용규

Introduction

Github link : https://github.com/zhaobozb/layout2im

- 사용자가 그림을 그리면서 설명할 수 있도록 편의를 제공

- Artist가 그림 초안을 그려 볼 수 있음

- 사용자가 쉽게 생성한 그림을 가지고 그림을 통한 검색도 가능

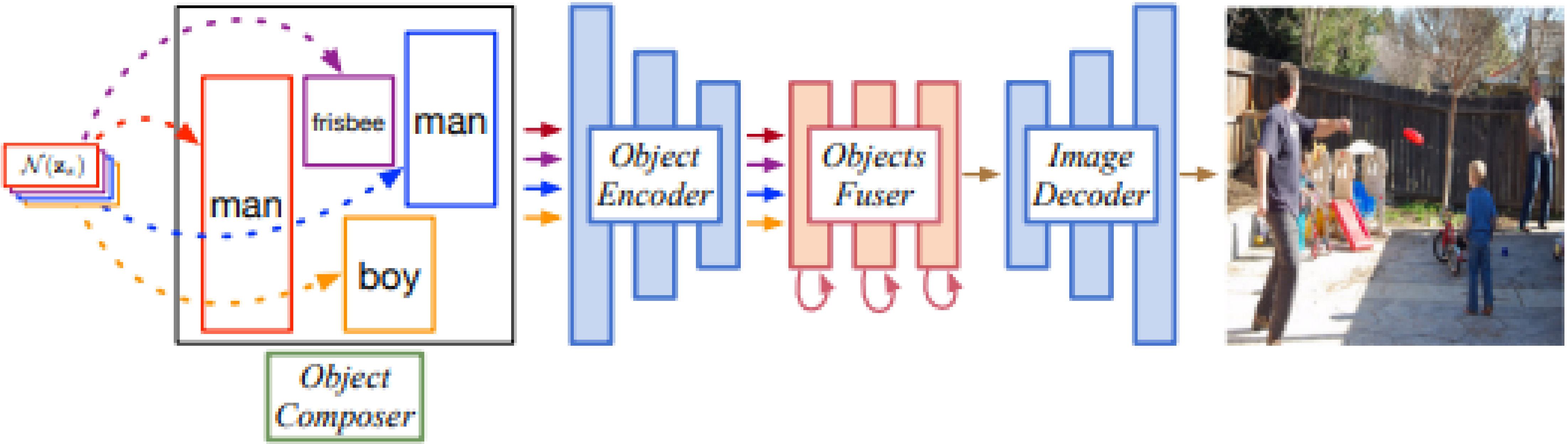
Introduction

Previous relevant work : Text-to-image approach

- 단순한 이미지에 대해서만 그럴듯하게 생성함
- 사람마다 기준이 단어(작은, 큰)로 인한 애매모호함
- 복잡한 이미지(Multiple Object)에서 생성이 어려움

Introduction

Layout2Im



- Coarse layout (bounding boxes + object categories)
- It is much more controllable and flexible to generate an image from layout than textual description

Introduction

Two challenges

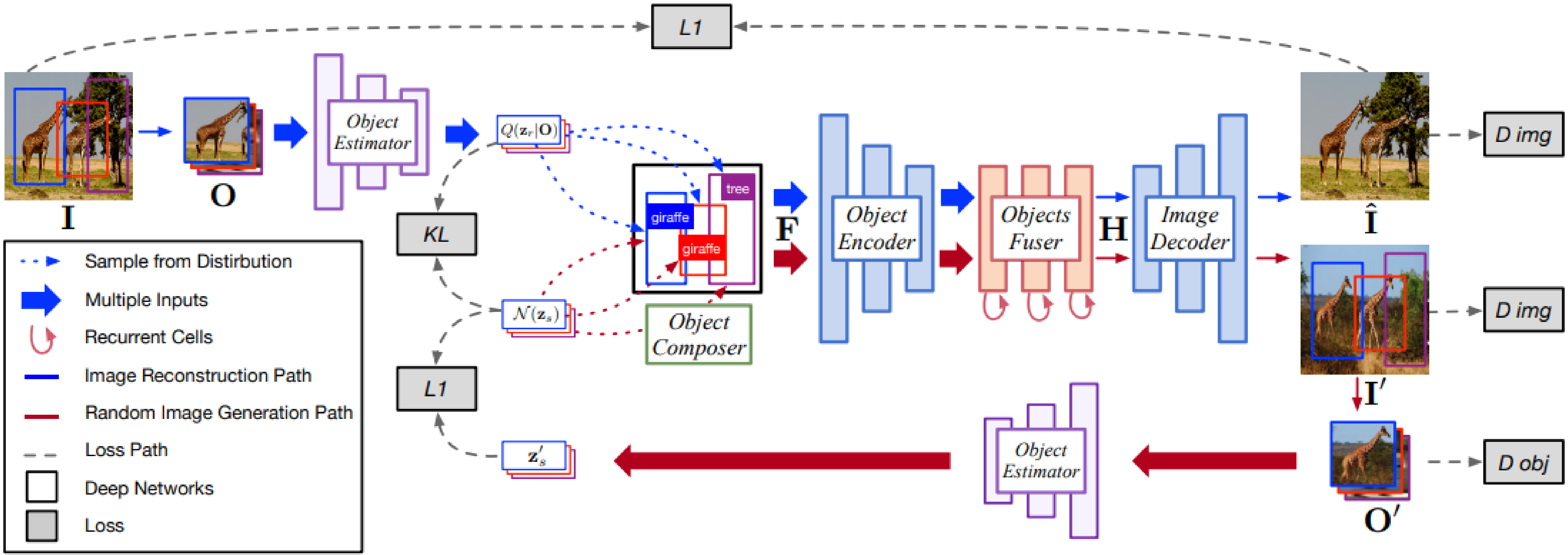
- Image generation from layout is a difficult one-to-many problem
  - interaction
- The information conveyed by a bounding box and corresponding label is very limited
  - Category & location 만으로 이미지가 결정되는게 아니라 interaction도 고려해야함
  - 공간적으로 가까운 물체는 bounding box가 겹칠 수 있음

Contribution

- Coarse layout(bounding boxes + object categories)로부터 유연한 이미지 생성
- Representation of objects를 category & appearance로 disentangle

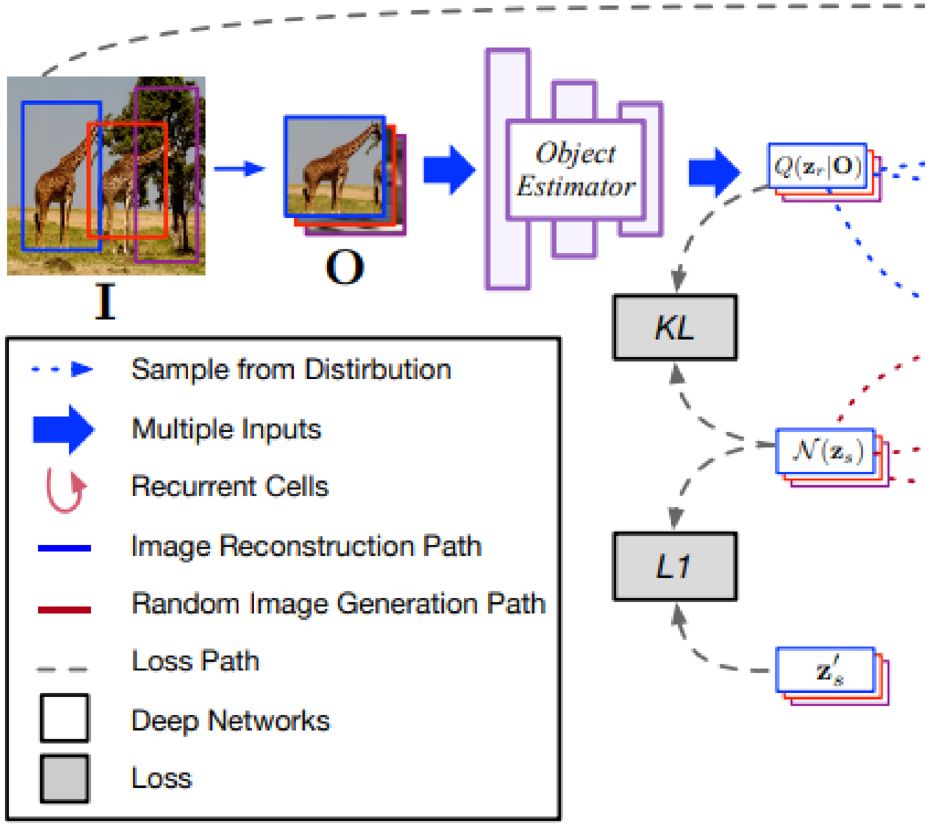
  : 같은 layout에서 다양한 Image 생성
- Segmentation mask 없이 COCO-Stuff and Visual Genome datasets 에서 좋은 성능

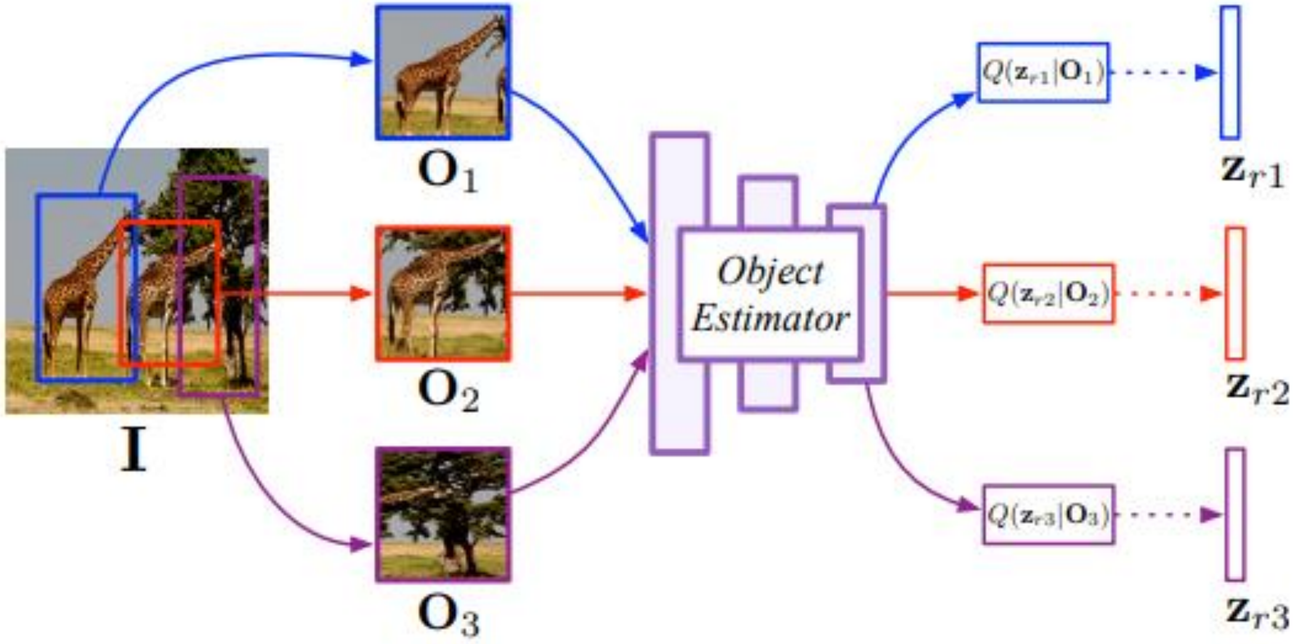
Method



L1
Object Estimator
$Q(\mathbf{z}_r|\mathbf{O})$
tree
giraffe
giraffe
I
O
KL
$\mathcal{N}(\mathbf{z}_s)$
Object Composer
F
Object Encoder
Objects Fuser
H
Image Decoder
Î
D img
D img
I′
D obj
O′
Object Estimator
L1
$\mathbf{z}'_s$
Sample from Distribution
Multiple Inputs
Recurrent Cells
Image Reconstruction Path
Random Image Generation Path
Loss Path
Deep Networks
Loss

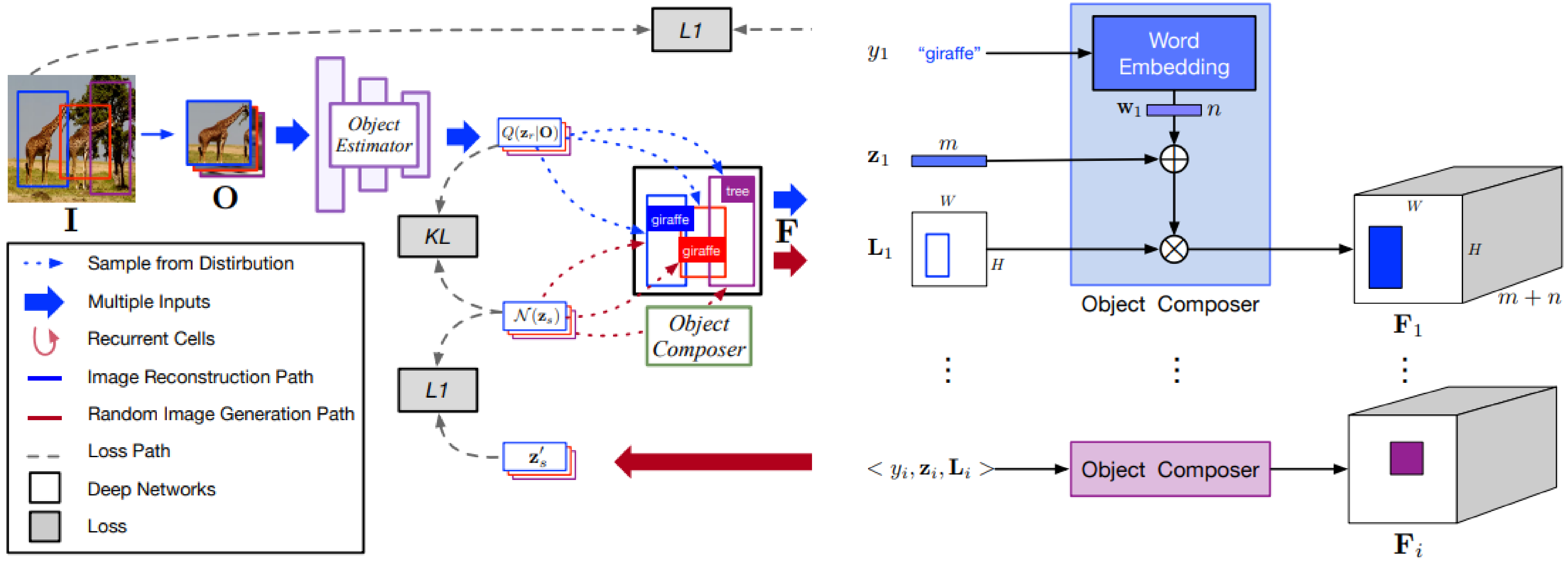
Method





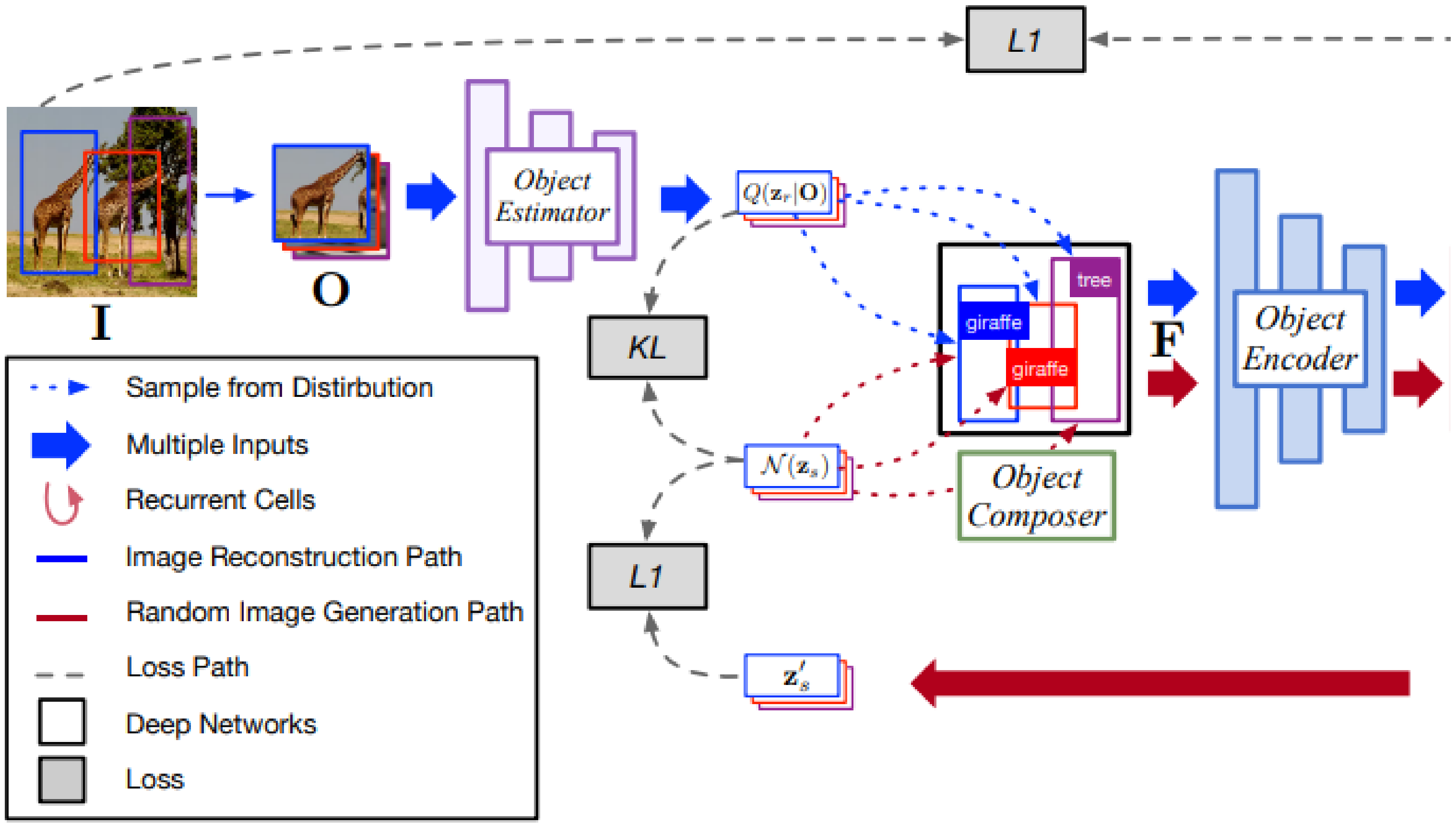
$$\mathbf{z}_{ri} \sim Q(\mathbf{z}_{ri}|\mathbf{O}_i) = \mathcal{N}(\mu(\mathbf{O}_i), \sigma(\mathbf{O}_i))$$

Method

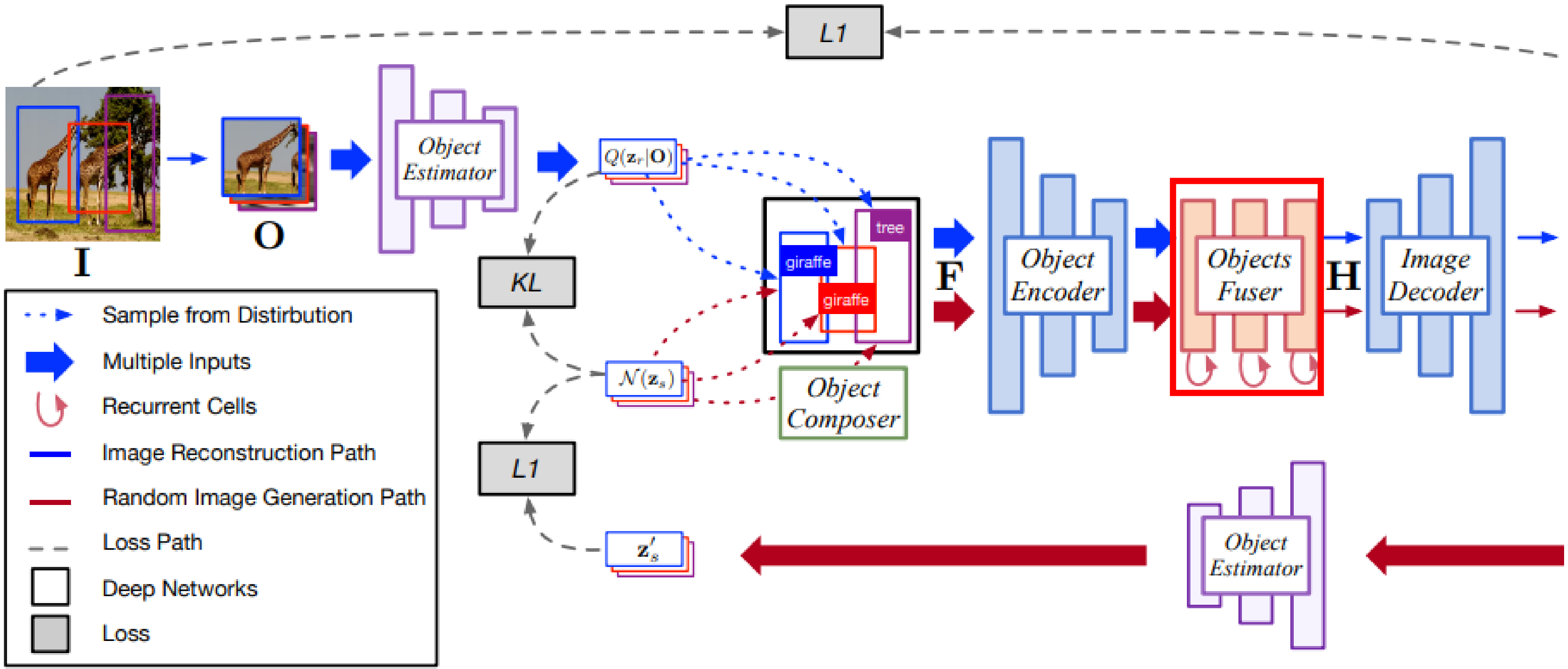


Word embedding : Identity of the object / Object latent code : appearance of a specific instance

Method


L1
I
O
Object Estimator
$Q(\mathbf{z}_r|\mathbf{O})$
KL
$\mathcal{N}(\mathbf{z}_s)$
L1
$\mathbf{z}'_s$
tree
giraffe
giraffe
Object Composer
F
Object Encoder
Sample from Distirbution
Multiple Inputs
Recurrent Cells
Image Reconstruction Path
Random Image Generation Path
Loss Path
Deep Networks
Loss

Method



L1
I
O
Object Estimator
$Q(\mathbf{z}_r|\mathbf{O})$
KL
$\mathcal{N}(\mathbf{z}_s)$
L1
$\mathbf{z}'_s$
tree
giraffe
giraffe
Object Composer
F
Object Encoder
Objects Fuser
H
Image Decoder
Object Estimator
Sample from Distirbution
Multiple Inputs
Recurrent Cells
Image Reconstruction Path
Random Image Generation Path
Loss Path
Deep Networks
Loss

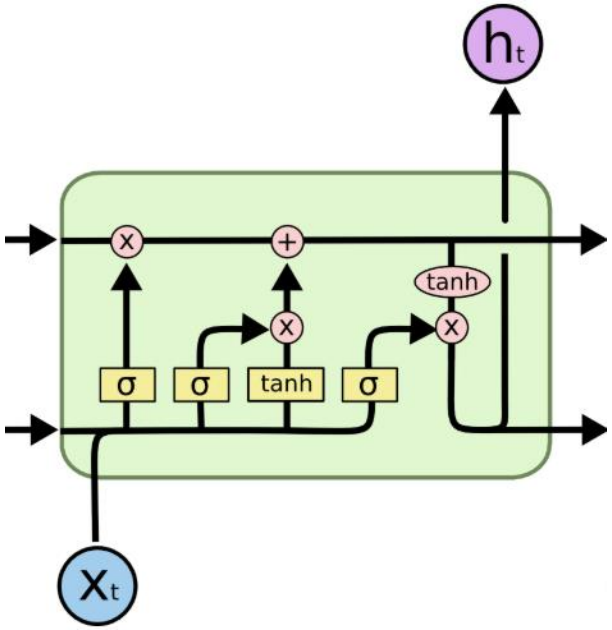
Method

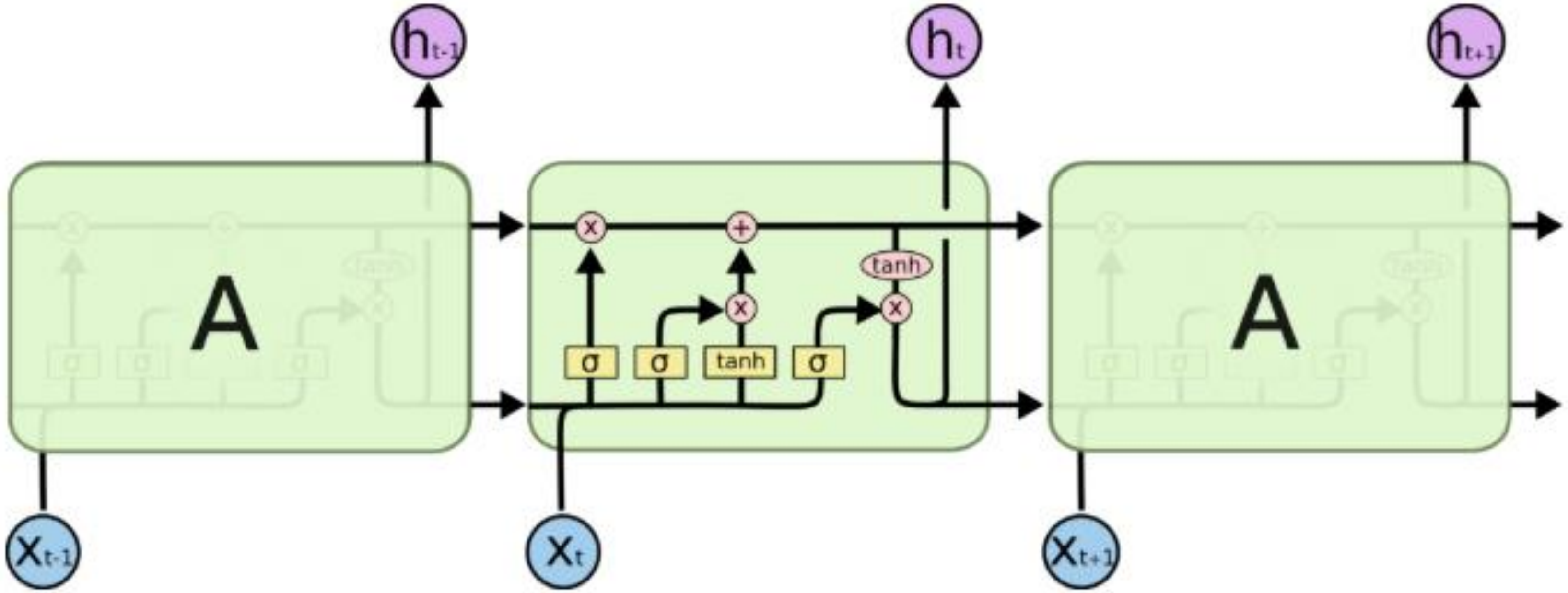
- 모든 object가 각각 원하는 위치에 존재

- 다른 object를 보고 object representation을 조정

- 배경 같이 정해지지 않은 지역(unspecified regions)을 채워야 함

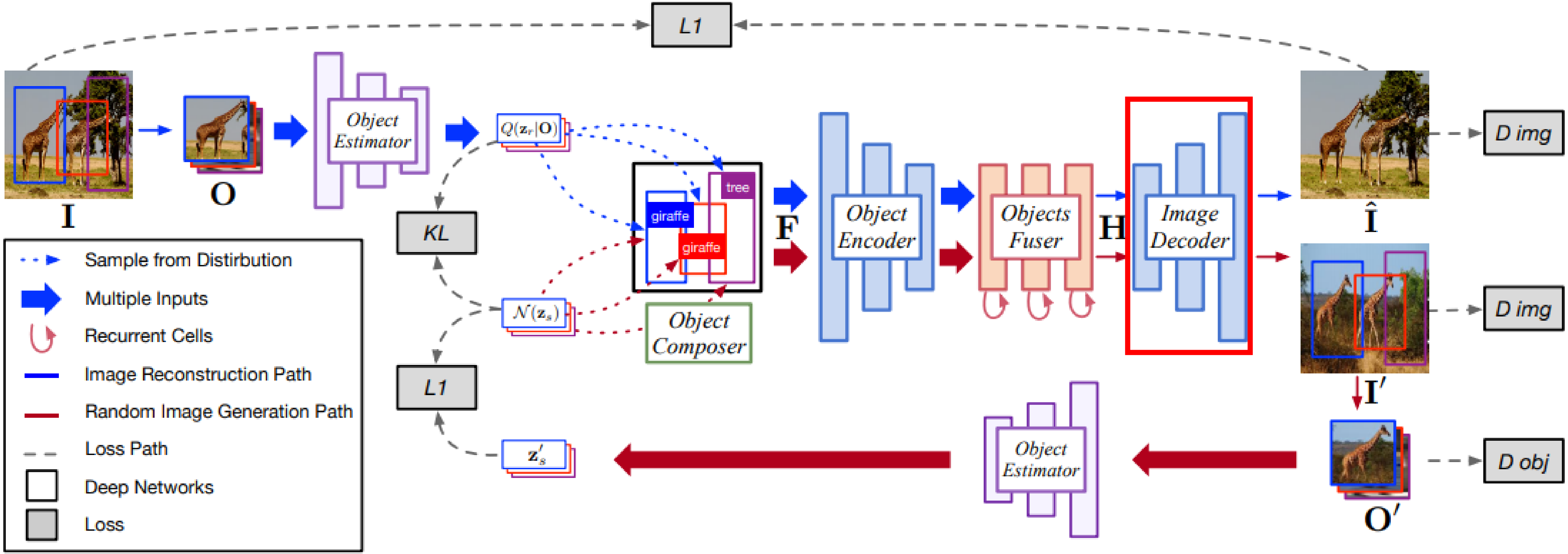
Method



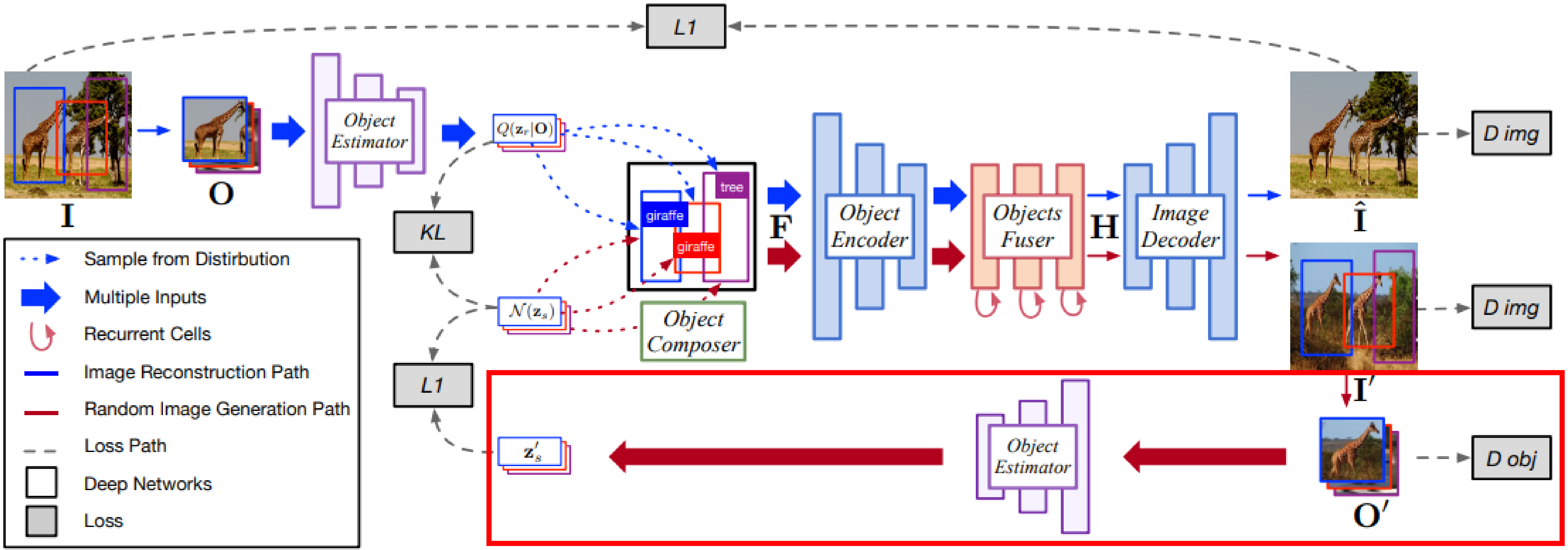


- Hidden state & Cell state : Feature map
- Convolutional layer
- Spatial information 유지
- Location & category information in H

Method

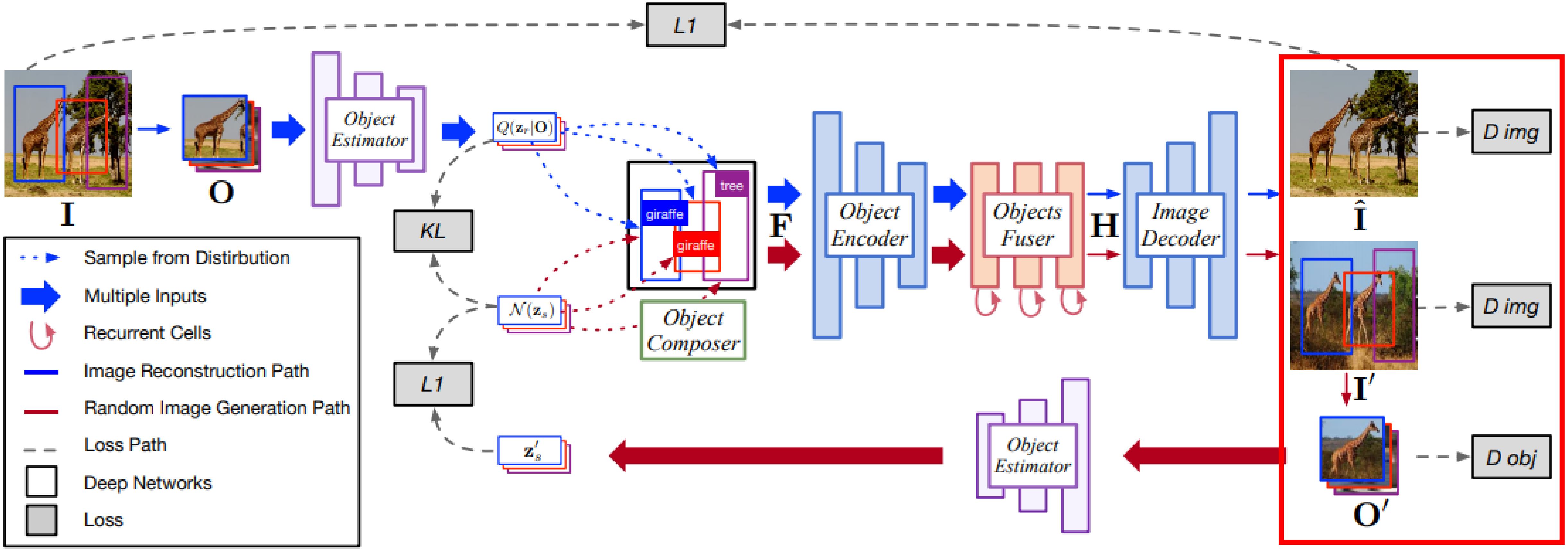

L1
I
O
Object Estimator
$Q(\mathbf{z}_r|\mathbf{O})$
KL
$\mathcal{N}(\mathbf{z}_s)$
L1
$\mathbf{z}'_s$
tree
giraffe
giraffe
Object Composer
F
Object Encoder
Objects Fuser
H
Image Decoder
Î
I′
O′
D img
D img
D obj
Object Estimator
Sample from Distirbution
Multiple Inputs
Recurrent Cells
Image Reconstruction Path
Random Image Generation Path
Loss Path
Deep Networks
Loss

Method



- Object latent code regression
  : Many-to-one 방지(다양한 이미지 생성 / mean vector 사용

Method



- Discriminator 역할 : (real image, real object, classification)

Method

- Total loss

- **KL Loss** $\mathcal{L}_{\mathrm{KL}} = \sum_{i=1}^{o} \mathbb{E}[\mathcal{D}_{\mathrm{KL}}(Q(\mathbf{z}_{ri}|\mathbf{O}_i)||\mathcal{N}(\mathbf{z}_r))]$ computes the KL-Divergence between the distribution $Q(\mathbf{z}_r|\mathbf{O})$ and the normal distribution $\mathcal{N}(\mathbf{z}_r)$, where o is the number of objects in the image/layout.
- **Image Reconstruction Loss** $\mathcal{L}_1^{\mathrm{img}} = ||\mathbf{I} - \hat{\mathbf{I}}||_1$ penalizes the $\mathcal{L}_1$ difference between ground-truth image $\mathbf{I}$ and reconstructed image $\hat{\mathbf{I}}$.
- **Object Latent Code Reconstruction Loss** $\mathcal{L}_1^{\mathrm{latent}} = \sum_{i=1}^{o} ||\mathbf{z}_{si} - \mathbf{z}'_{si}||_1$ penalizes the $\mathcal{L}_1$ difference between the randomly sampled $\mathbf{z}_s \sim N(\mathbf{z}_s)$ and the re-estimated $\mathbf{z}'_s$ from the generated objects $\mathbf{O}'$.
- **Image Adversarial Loss** $\mathcal{L}_{\mathrm{GAN}}^{\mathrm{img}}$ is defined as in Eq. (1), where x is the ground truth image $\mathbf{I}$, y is the reconstructed image $\hat{\mathbf{I}}$ and sampled image $\mathbf{I}'$.
- **Object Adversarial Loss** $\mathcal{L}_{\mathrm{GAN}}^{\mathrm{obj}}$ is also defined as in Eq. (1), where x is the objects $\mathbf{O}$ cropped from the ground truth image $\mathbf{I}$, y are $\hat{\mathbf{O}}$ and $\mathbf{O}'$ cropped from the reconstructed image $\hat{\mathbf{I}}$ and sampled image $\mathbf{I}'$.
- **Auxiliar Classification Loss** $\mathcal{L}_{\mathrm{AC}}^{\mathrm{obj}}$ from D_{obj} encourages the generated objects $\hat{\mathbf{O}}_i$ and $\mathbf{O}'_i$ to be recognizable as their corresponding categories.

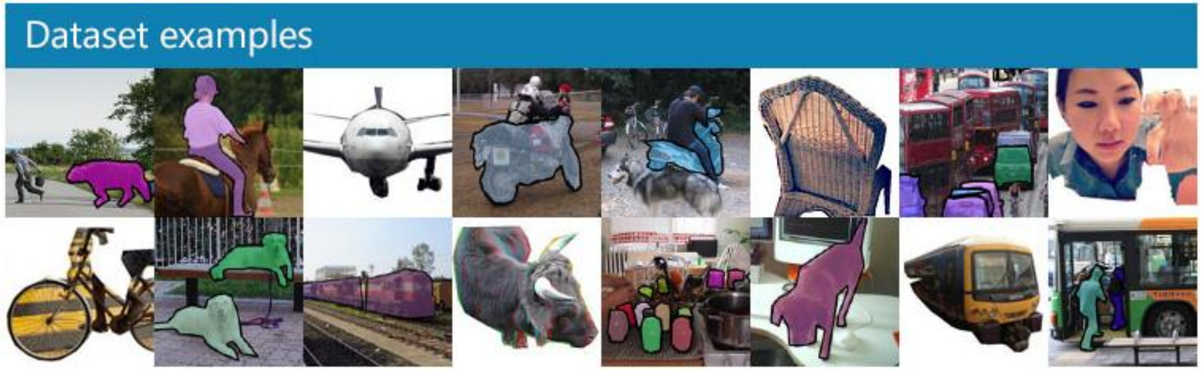
Therefore, the final loss function of our model is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{KL}} + \lambda_2 \mathcal{L}_1^{\mathrm{img}} + \lambda_3 \mathcal{L}_1^{\mathrm{latent}} + \lambda_4 \mathcal{L}_{\mathrm{adv}}^{\mathrm{img}} + \lambda_5 \mathcal{L}_{\mathrm{adv}}^{\mathrm{obj}} + \lambda_6 \mathcal{L}_{\mathrm{AC}}^{\mathrm{obj}},$$

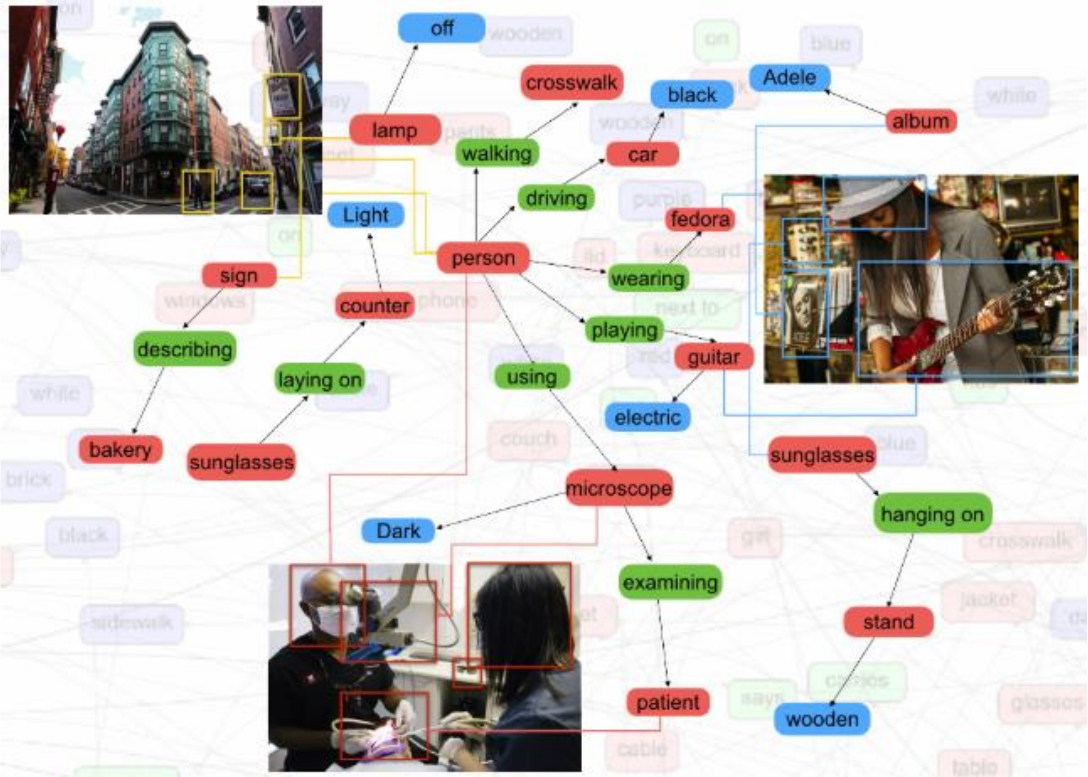
where, λ_i are the parameters balancing different losses.

Experiments

- COCO-Stuff & Visual Genome datasets

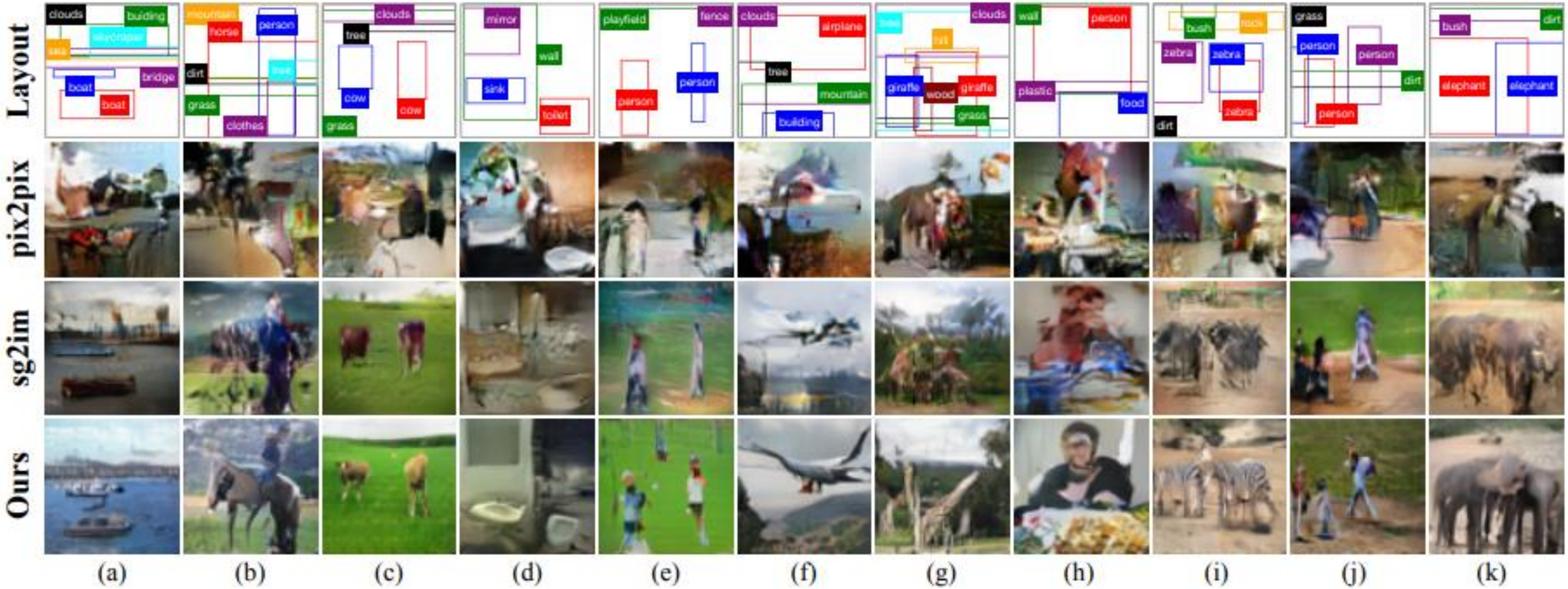



```
"annotations": [
    {
        "segmentation": [[510.66,423.01,511.72,420.03,...,510.45,423.01]],
        "area": 702.1057499999998,
        "iscrowd": 0,
        "image_id": 289343,
        "bbox": [473.07,395.93,38.65,28.67],
        "category_id": 18,
        "id": 1768
    },
```

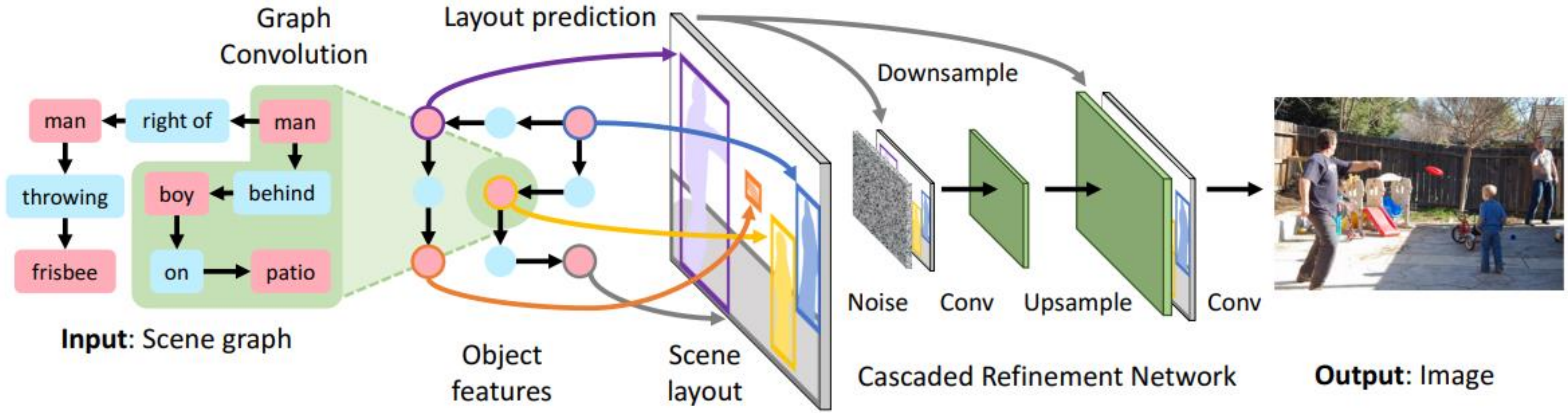



Dataset	Train	Val.	Test	# Obj.	# Obj. in Image
COCO [1]	24,972	1,024	2,048	171	3 ~ 8
VG [18]	62,565	5,506	5,088	178	3 ~ 30

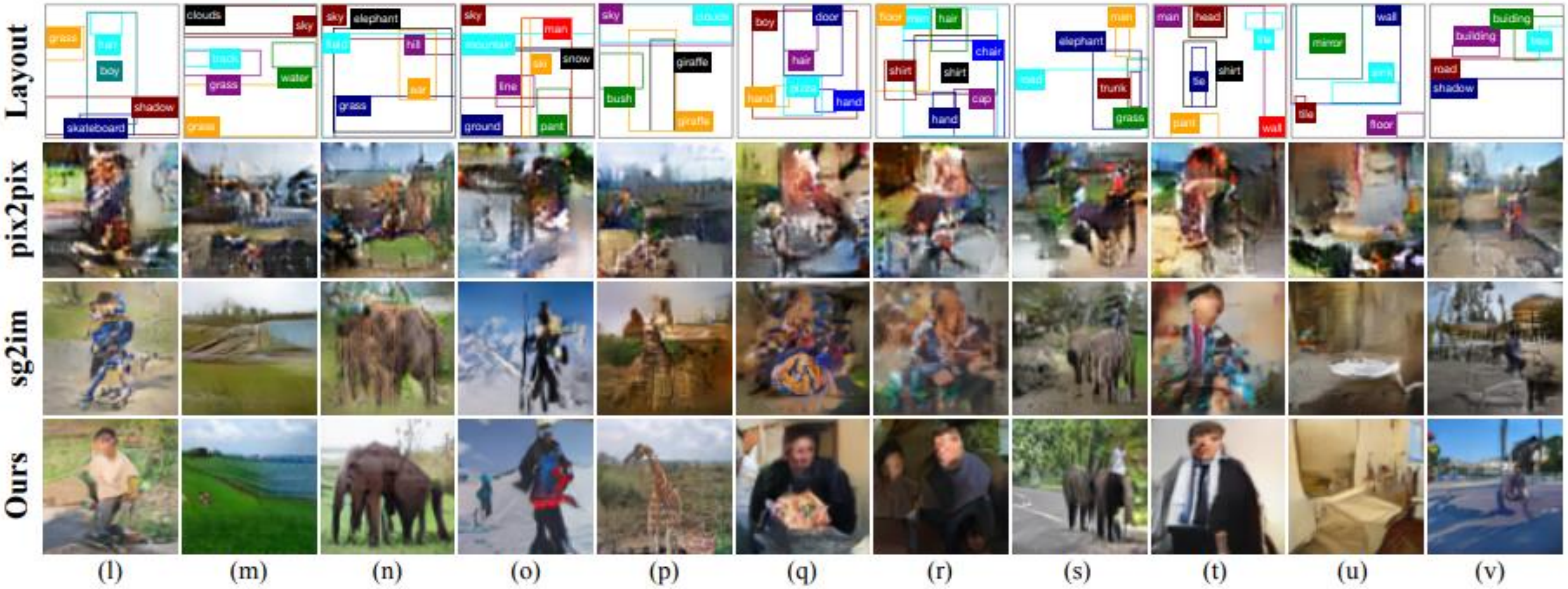
Experiments



Experiments



Experiments



Experiments

Experiments

Experiments

Method	Inception Score		Accuracy		Diversity Score	
	COCO	VG	COCO	VG	COCO	VG
Real Images (64 × 64)	16.3 ± 0.4	13.9 ± 0.5	55.16	49.13	-	-
pix2pix [12]	3.5 ± 0.1	2.7 ± 0.02	12.06	9.20	0	0
sg2im (GT Layout) [13]	7.3 ± 0.1	6.3 ± 0.2	30.04	40.29	0.02 ± 0.01	0.15 ± 0.12
Ours	**9.1 ± 0.1**	**8.1 ± 0.1**	**50.84**	**48.09**	**0.15 ± 0.06**	**0.17 ± 0.09**

Method	IS	Accu.	DS
w/o $\mathcal{L}_1^{\text{img}}$	7.6 ± 0.2	49.03	0.17 ± 0.09
w/o $\mathcal{L}_1^{\text{latent}}$	7.5 ± 0.1	48.90	0.16 ± 0.09
w/o $\mathcal{L}_{\text{AC}}^{\text{obj}}$	6.5 ± 0.1	10.06	**0.37 ± 0.11**
w/o $\mathcal{L}_{\text{adv}}^{\text{img}}$	7.1 ± 0.1	56.17	0.13 ± 0.09
w/o $\mathcal{L}_{\text{adv}}^{\text{obj}}$	7.3 ± 0.1	**57.74**	0.14 ± 0.09
full model	**8.1 ± 0.1**	48.09	0.17 ± 0.09

Conclusion

- High resolution
- More controllable image generation