Semantically Multi-modal Image Synthesis

Zhen Zhu[1*], Zhiliang Xu[1*], Ansheng You[2], Xiang Bai[1†]
[1]*Huazhong University of Science and Technology*, [2]*Peking University*
{zzhu, zhiliangxu1, xbai}@hust.edu.cn, youansheng@pku.edu.cn

CVPR 2020

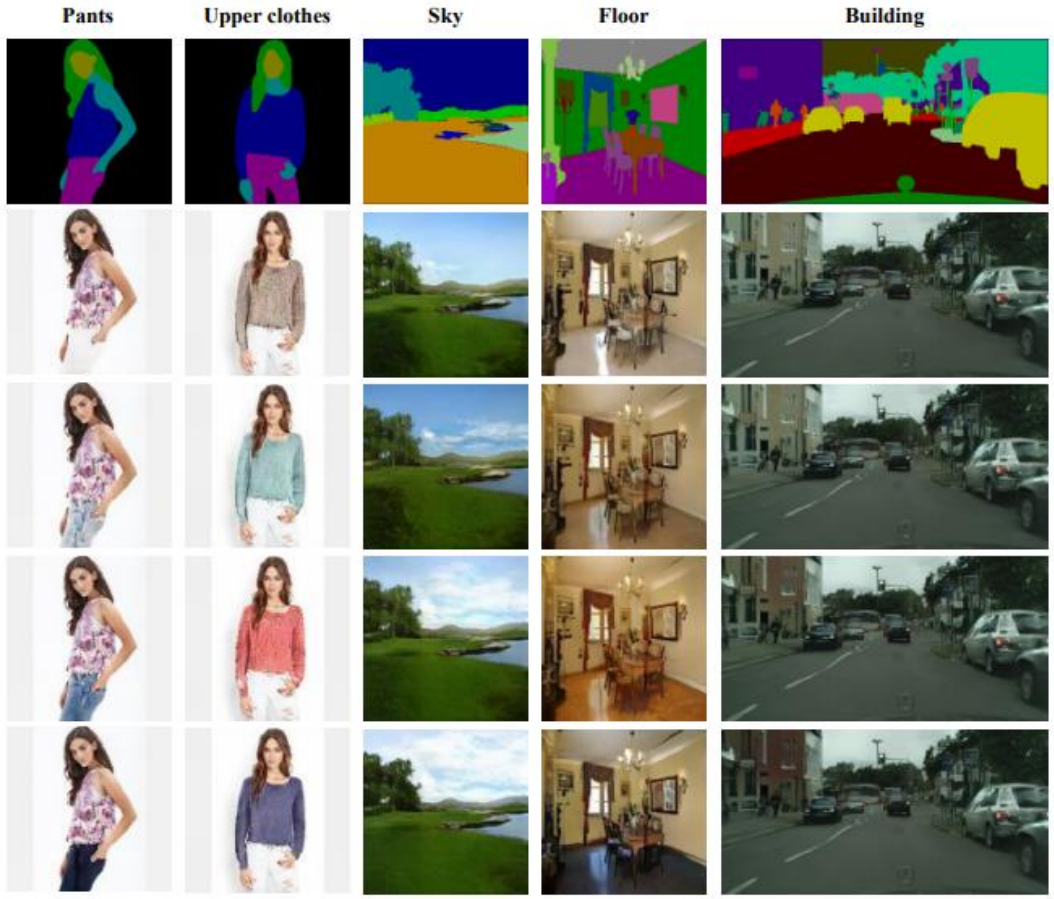
2020.06.11

Presented by Yonggyu Kim

Introduction

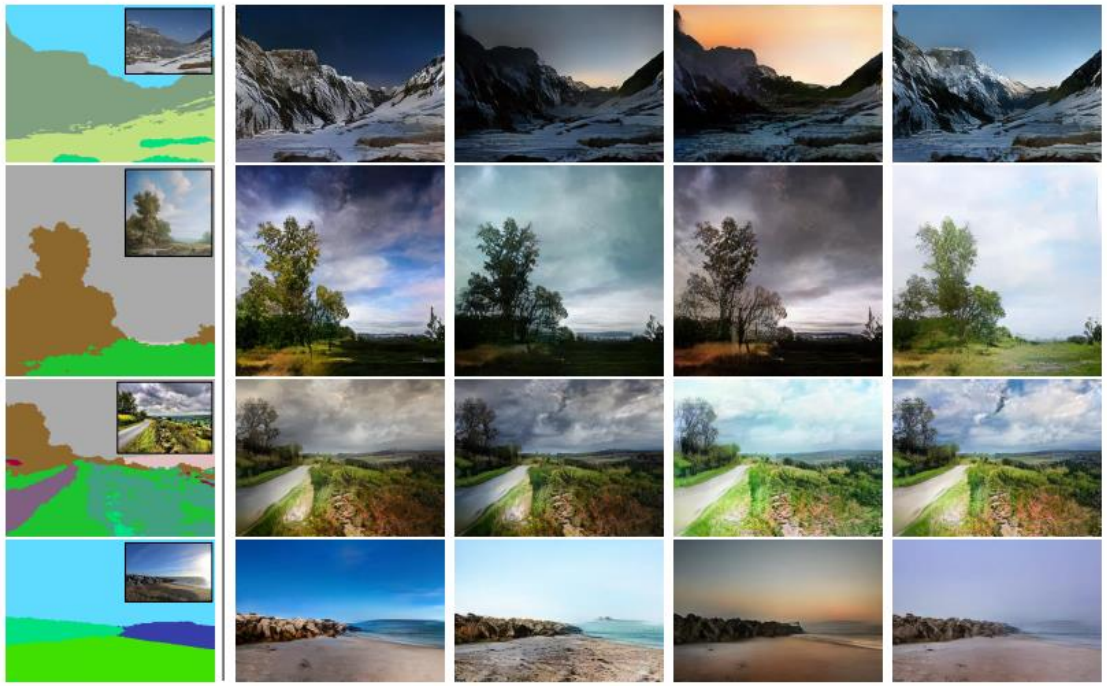
- **Semantically multi-modal image synthesis (SMIS) task**

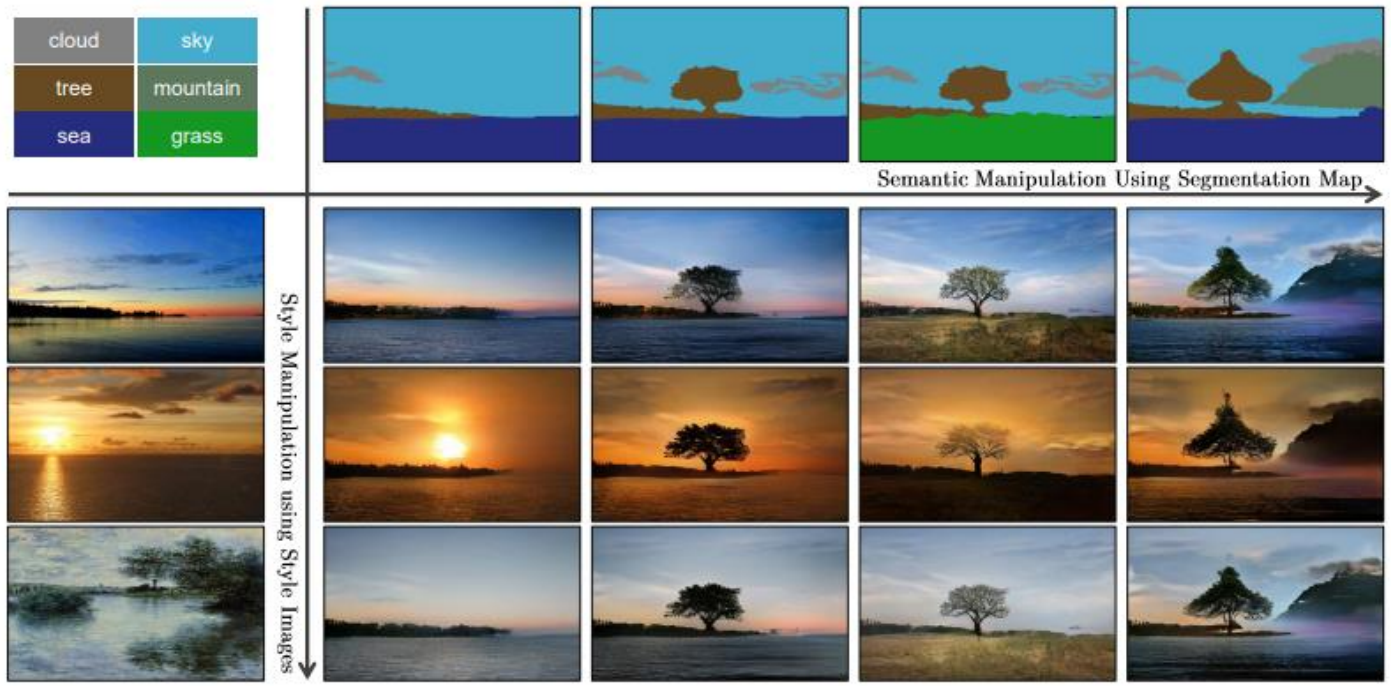
  Generating multi-modal images at the semantic level
  (= Translating semantic labels to natural images)



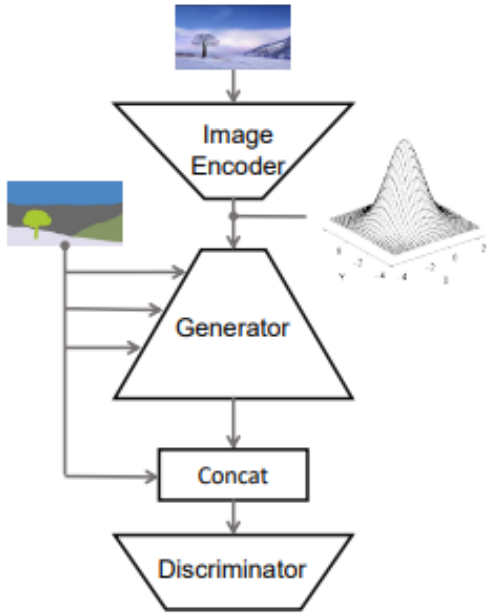
- **Contribution**
  - Group Decreasing Network(GroupDNet) introduces more controllability
  - Two new metrics mCSD & mOCD
  - A variety of interesting application such as appearance mixture, semantic manipulation, and style morphing

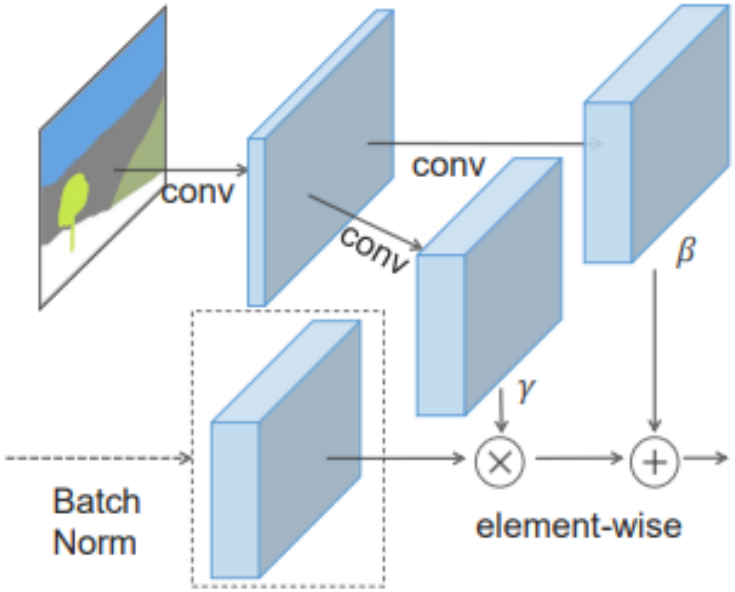
Preliminary : Spatially-adaptive denormalization

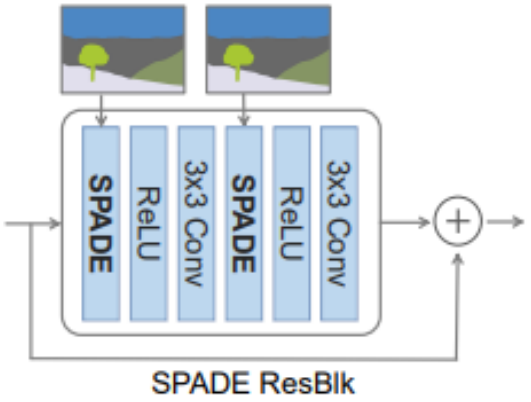


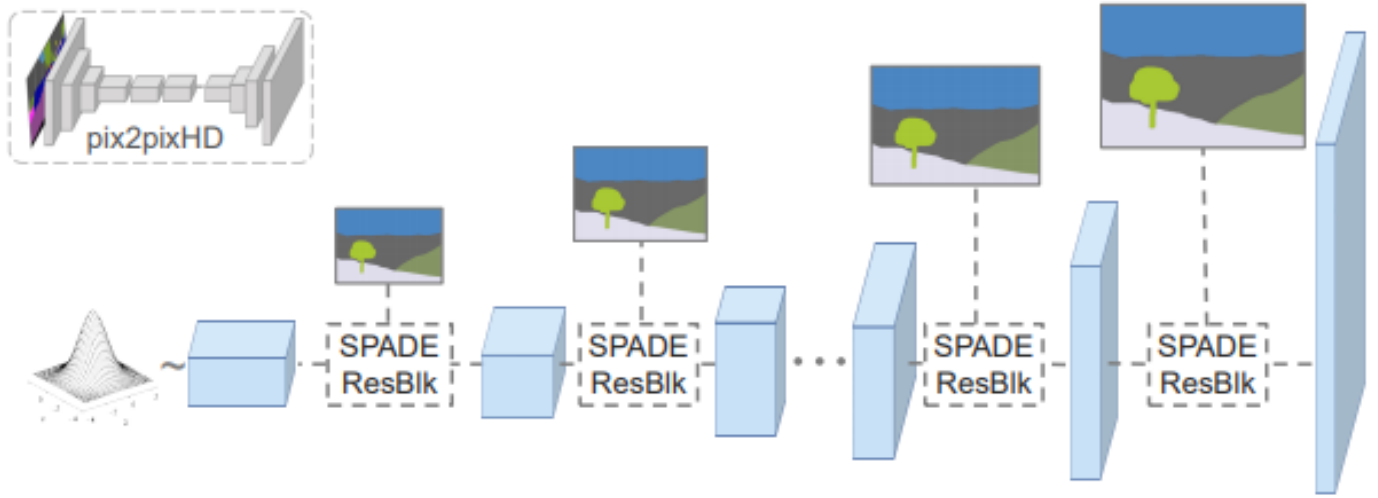


Preliminary : Spatially-adaptive denormalization



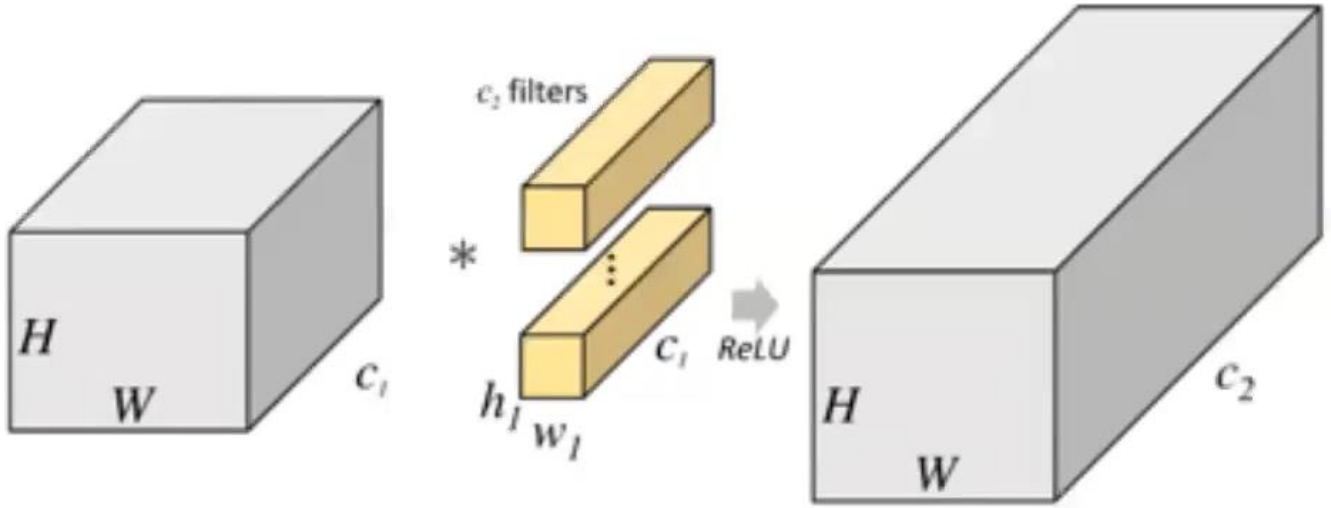






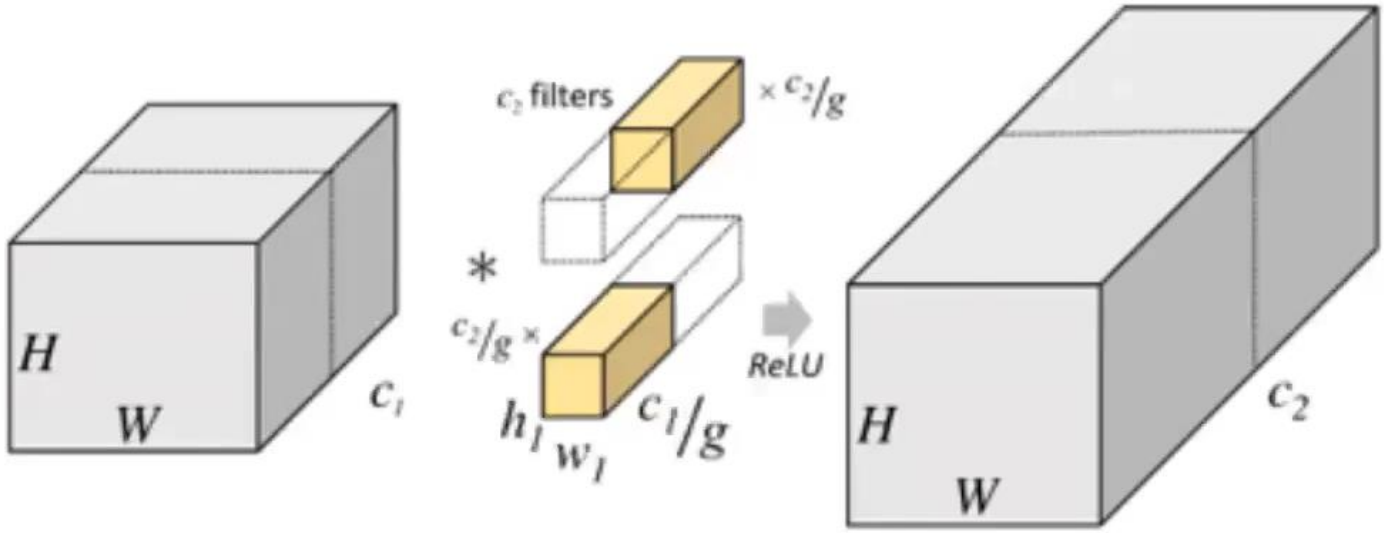
Preliminary : Group Convolution

standard convolutions

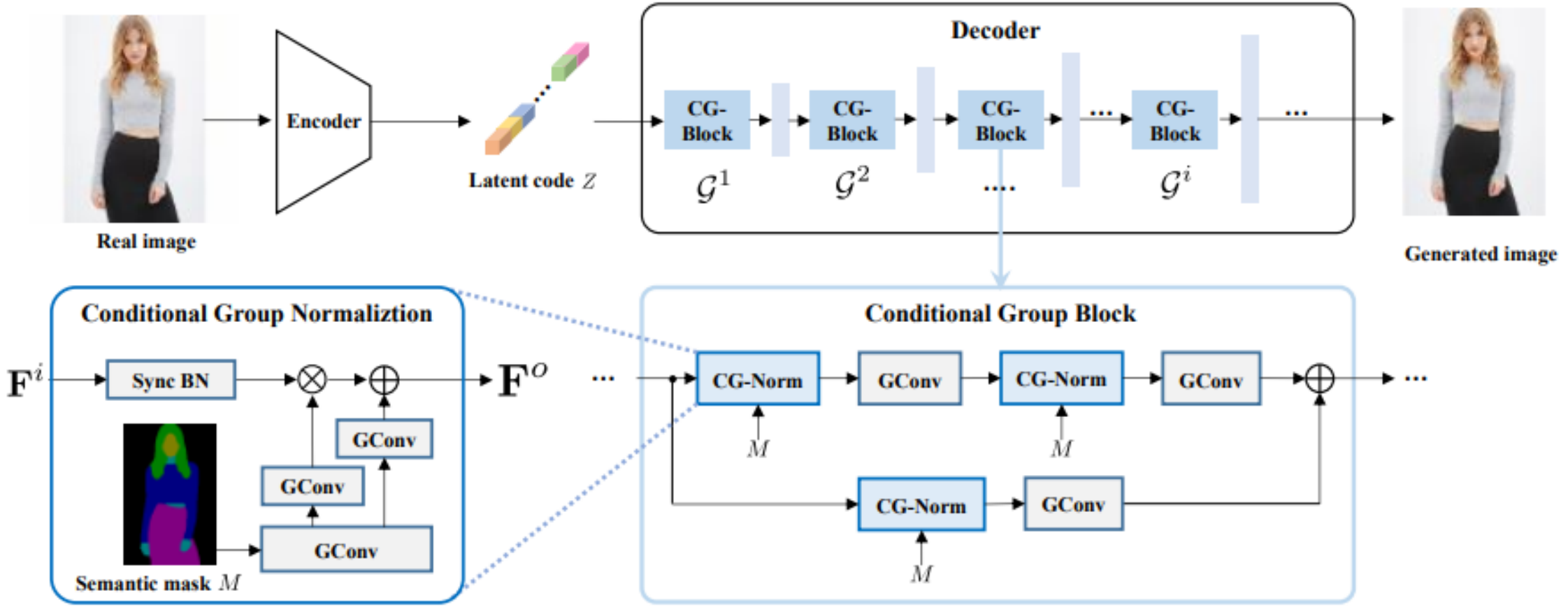


vs.

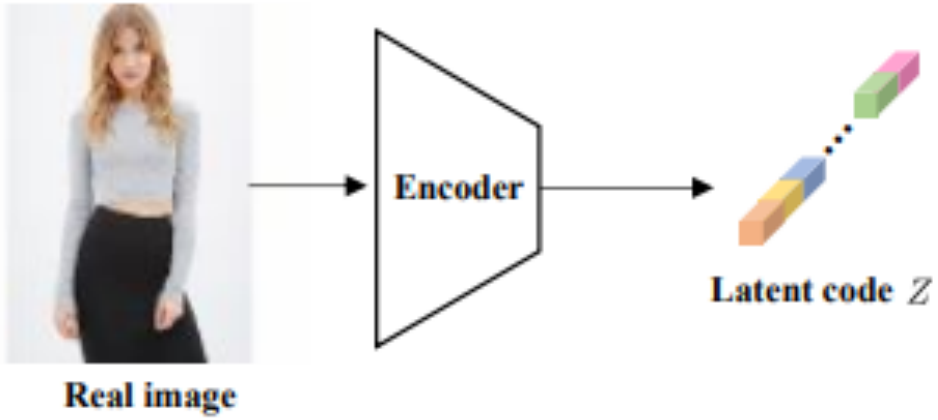
grouped convolutions

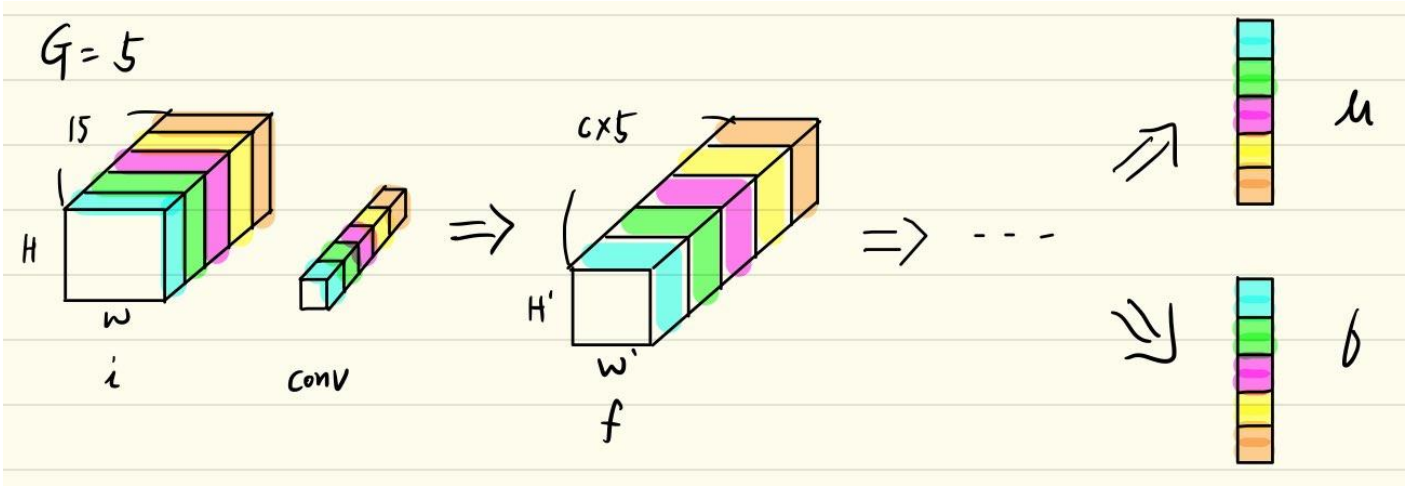


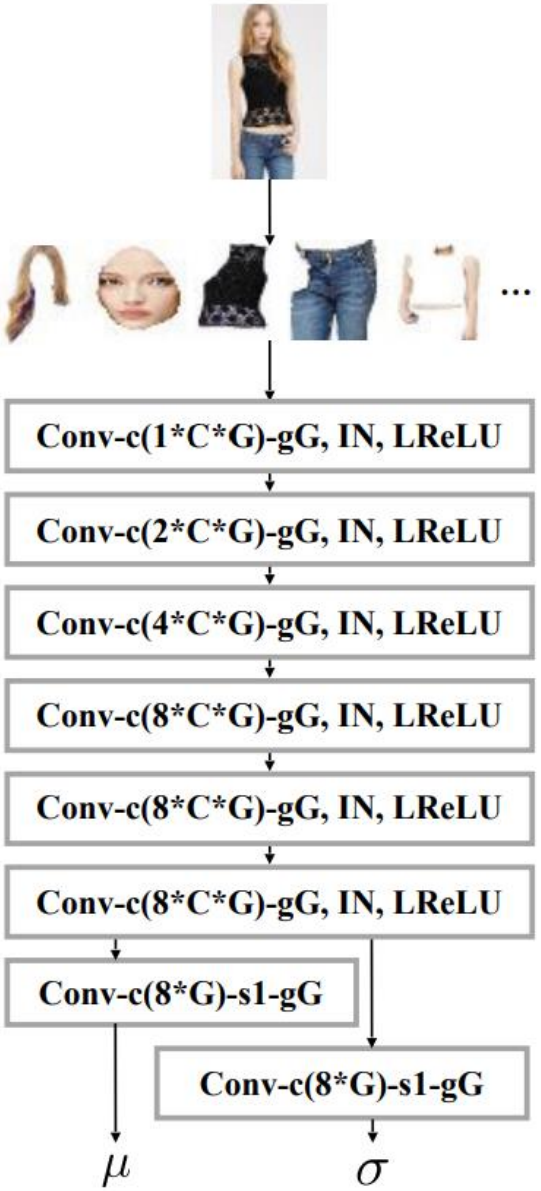
Method


Real image
Encoder
Latent code z
Decoder
CG-Block
CG-Block
CG-Block
CG-Block
$\mathcal{G}^1$
$\mathcal{G}^2$
$\mathcal{G}^i$
Generated image
Conditional Group Normaliztion
$\mathbf{F}^i$
Sync BN
GConv
GConv
GConv
Semantic mask M
$\mathbf{F}^o$
Conditional Group Block
CG-Norm
GConv
CG-Norm
GConv
CG-Norm
GConv
M
M
M

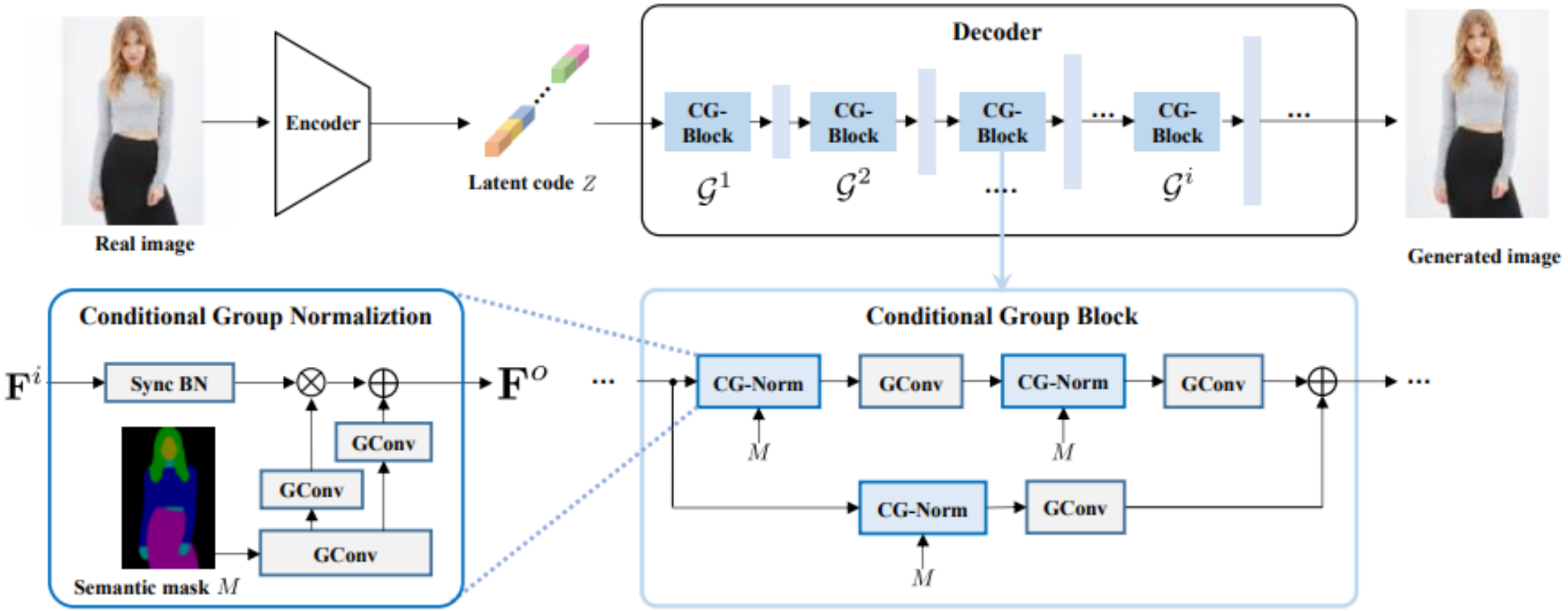
Method


Encoder
Latent code z
Real image


G = 5
15
H
w
i
conv
c×5
H'
w'
f
μ
σ

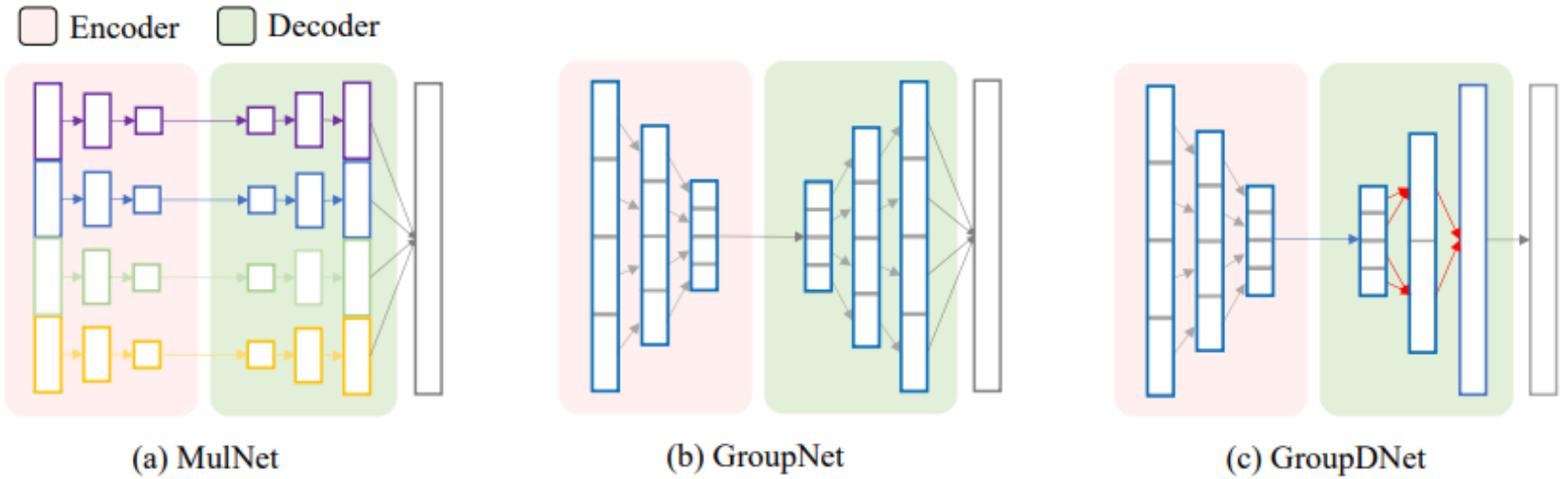

Conv-c(1*C*G)-gG, IN, LReLU
Conv-c(2*C*G)-gG, IN, LReLU
Conv-c(4*C*G)-gG, IN, LReLU
Conv-c(8*C*G)-gG, IN, LReLU
Conv-c(8*C*G)-gG, IN, LReLU
Conv-c(8*C*G)-gG, IN, LReLU
Conv-c(8*G)-s1-gG
Conv-c(8*G)-s1-gG
μ
σ

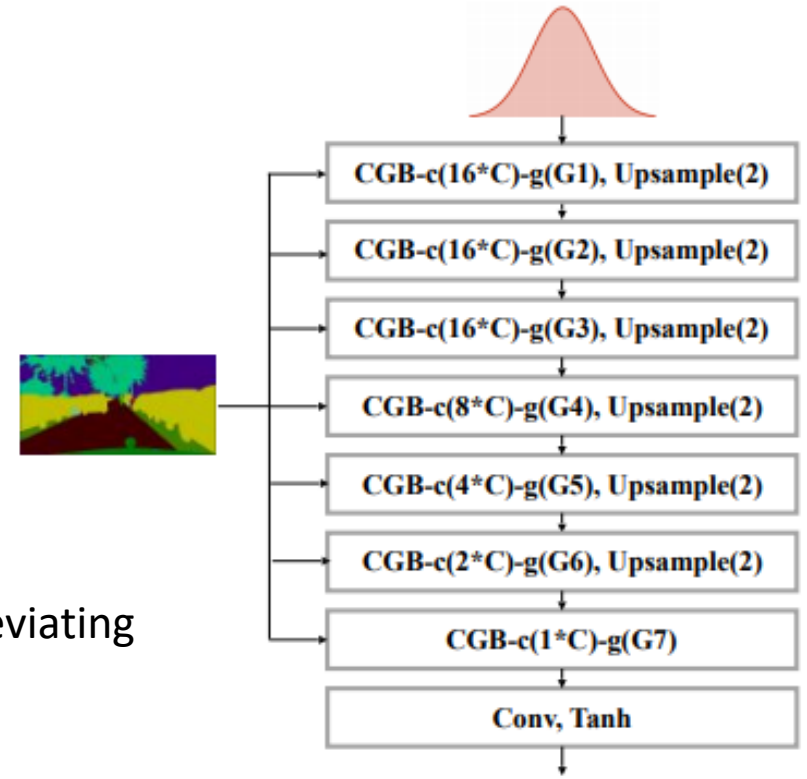
Method



$$\mathcal{L}_{\text{full}} = \arg\min_{G}\max_{D}\mathcal{L}_{\text{GAN}} + \lambda_1\mathcal{L}_{\text{FM}} + \lambda_2\mathcal{L}_{\text{P}} + \lambda_3\mathcal{L}_{\text{KL}}$$

Method





- **Class balance**
  - Unbalanced classes share parameters with their neighbor classes, hugely alleviating the class imbalance problem.
- **Class correlation**
  - GroupDNet carves these relationships throughout the decoder; hence, it exploits the correlations more accurately and thoroughly.
- **GPU memory**
  - ADE20K dataset, Tesla V100 graphics card

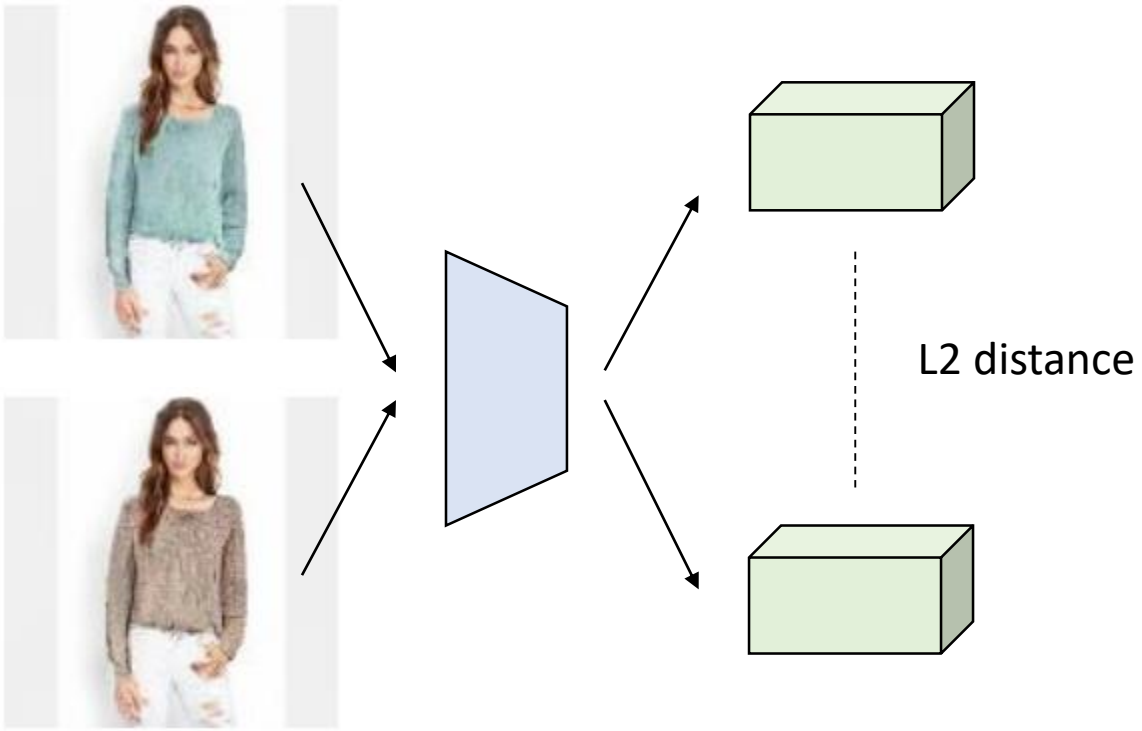
Experiments

- **Mean Class-Specific Diversity(mCSD) & mean Other-Classes Diversity(mOCD)**

  - two new metric based on LPIPS metric

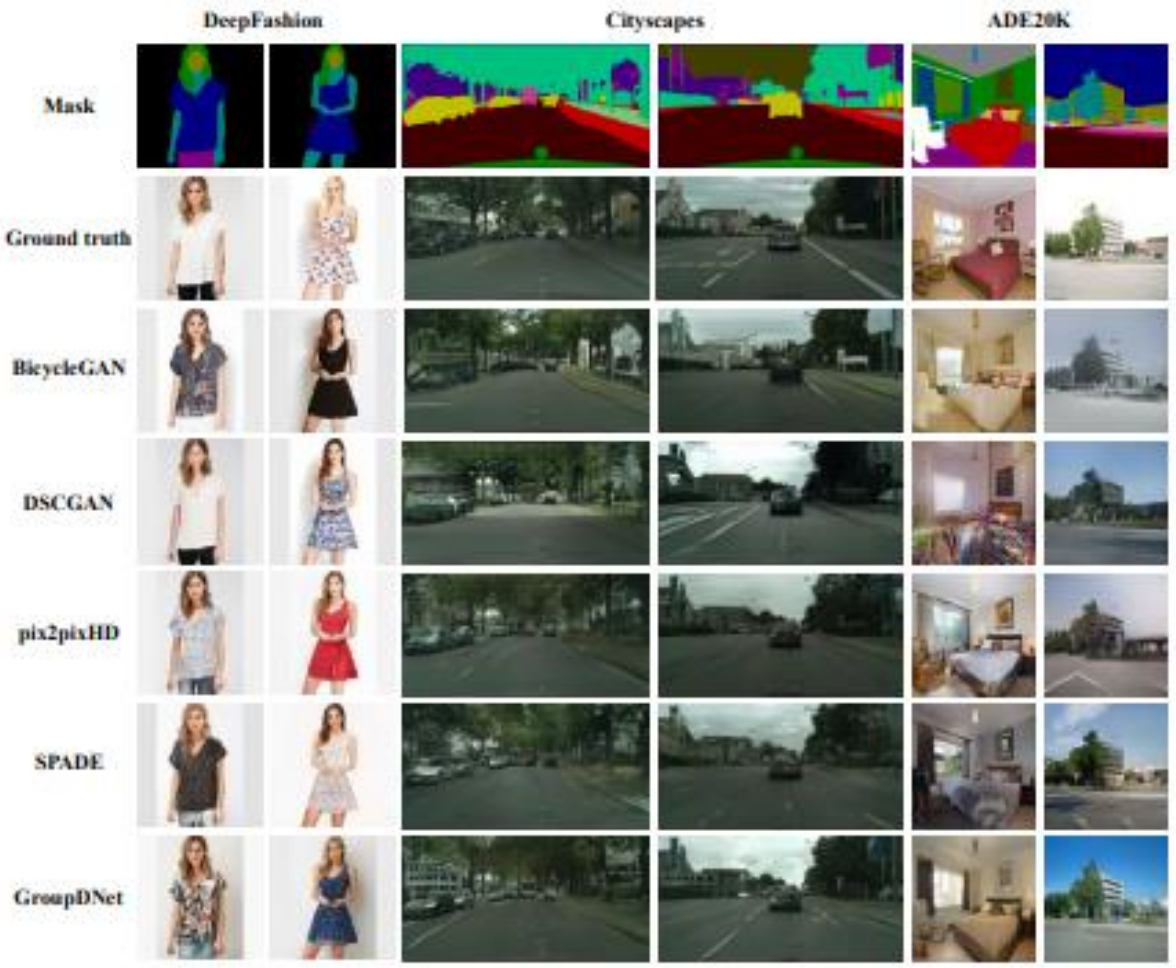
$$S = \{I_1^1, ..., I_1^n, ..., I_C^1, ..., I_C^n\}$$

- **LPIPS**



$$\text{mCSD} = \frac{1}{C}\sum_{c=1}^{C} L_c, \ \text{mOCD} = \frac{1}{C}\sum_{c=1}^{C} L_{\neq c}.$$

Experiments



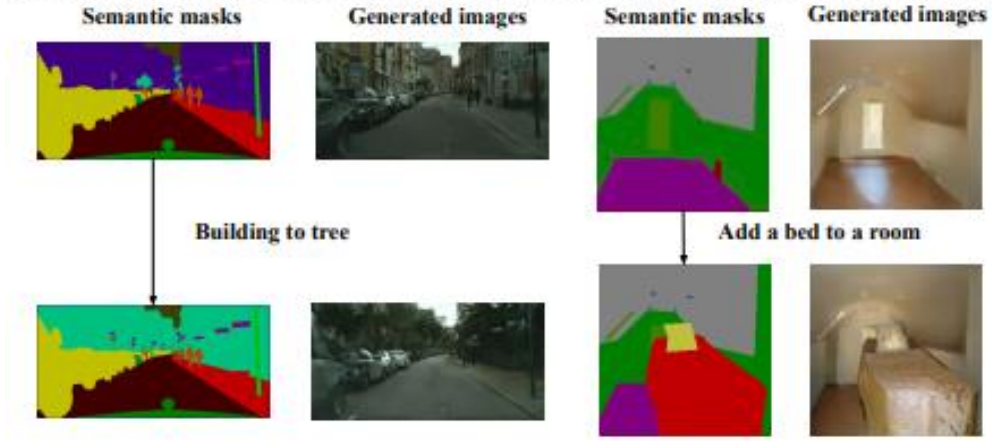
Models	FID↓	mCSD↑	mOCD↓	LPIPS↑	SHE↑	Speed↑	# Param↓
MulNet	12.07	0.0244	0.0019	0.202	79.2	6.3	105.1
GroupNet	12.58	0.0276	**0.0017**	0.203	**83.7**	8.2	97.7
Group Enc	10.83	0.0232	0.0065	0.217	69.3	19.6	105.5
Group Dec	9.84	0.0003	0.0257	0.206	26.4	12.1	111.3
VSPADE [38]	10.02	0.0304	0.1843	0.207	23.6	20.4	106.8
BicycleGAN [58]	40.07	**0.0316**	0.2147	**0.228**	24.8	66.9	**58.4**
DSCGAN [47]	38.40	0.0245	0.1560	0.163	27.6	**67.2**	**58.4**
GroupDNet	**9.50**	0.0264	0.0033	**0.228**	81.2	12.2	109.1

Method	DeepFashion			Cityscapes			ADE20K		
	mIoU↑	Acc↑	FID↓	mIoU↑	Acc↑	FID↓	mIoU↑	Acc↑	FID↓
BicycleGAN [58]	76.8	97.8	40.07	23.3	75.4	87.74	4.78	29.6	87.85
DSCGAN [47]	81.0	98.3	38.40	37.8	86.7	67.77	10.2	58.8	83.98
pix2pixHD [43]	85.2	98.8	17.76	58.3	92.5	78.24	27.6	75.7	55.9
SPADE [38]	87.1	**98.9**	10.02	**62.3**	93.5	58.10	**42.0**	**81.4**	**33.49**
GroupDNet	**87.3**	**98.9**	**9.50**	**62.3**	**93.7**	**49.81**	30.4	77.1	42.17

Experiments

(a) Semantically multi-modal image synthesis

(b) Appearance mixture



(c) Semantic manipulation

(d) Style morphing