Semantic Image Synthesis with Spatially-Adaptive Normalization(SPADE)

CVPR2019 Oral paper

+

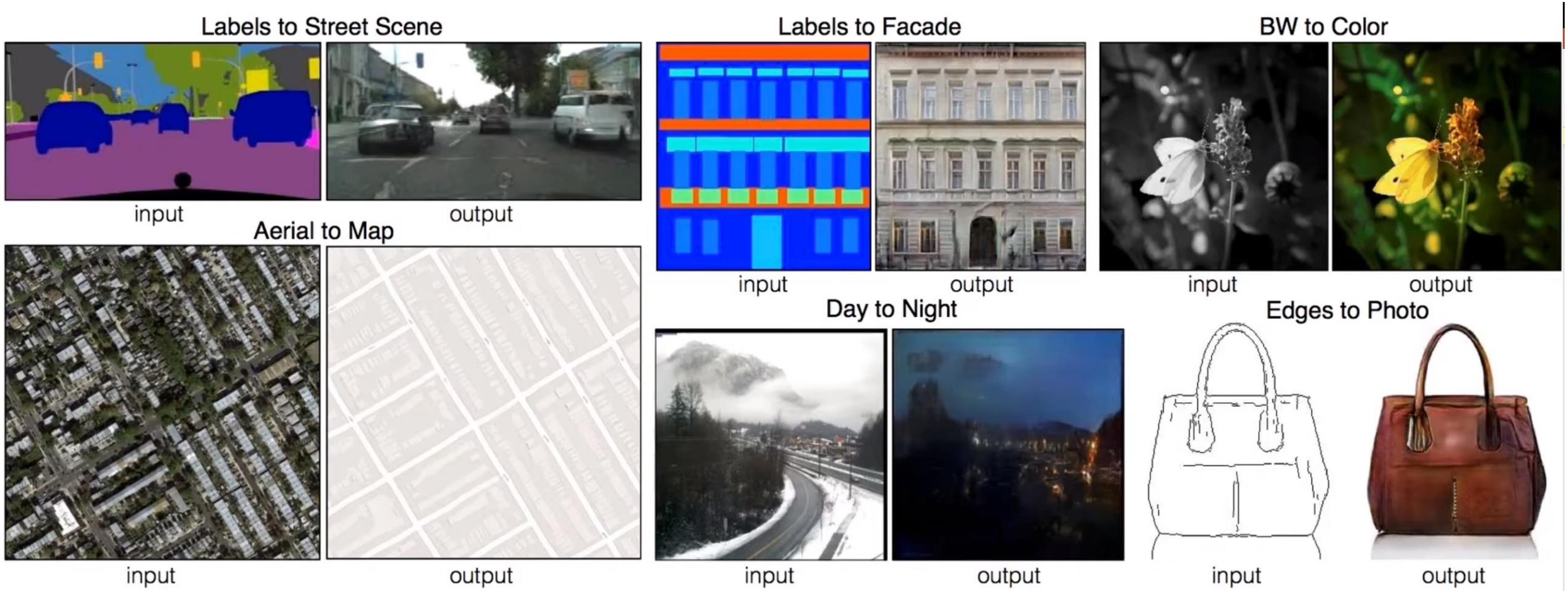
SEAN: Image Synthesis with Semantic Region-Adaptive Normalization

CVPR2020 Oral paper

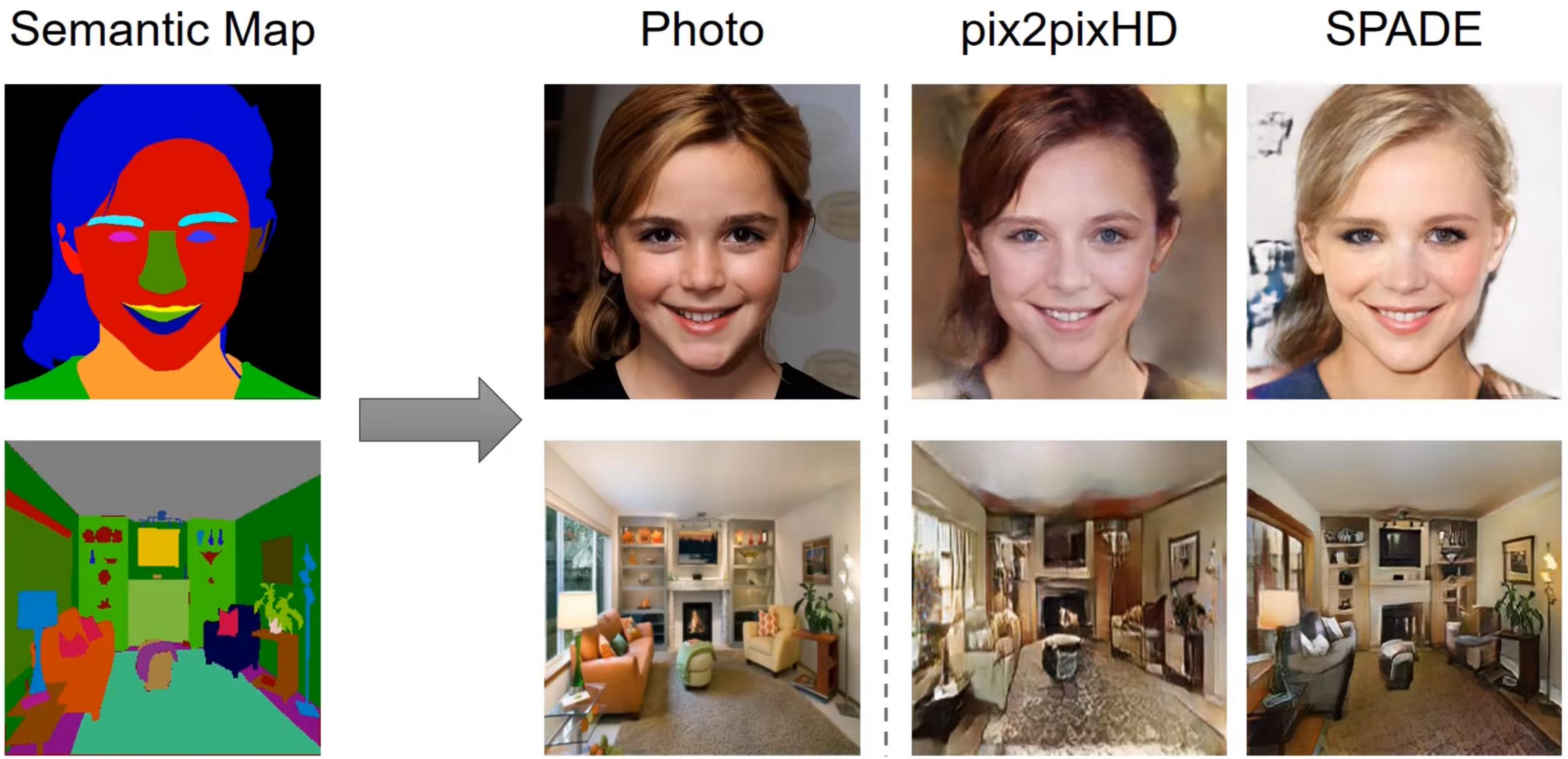
20.10.13 Leeminsoo

DAVIAN
Data and Visual Analytics Lab

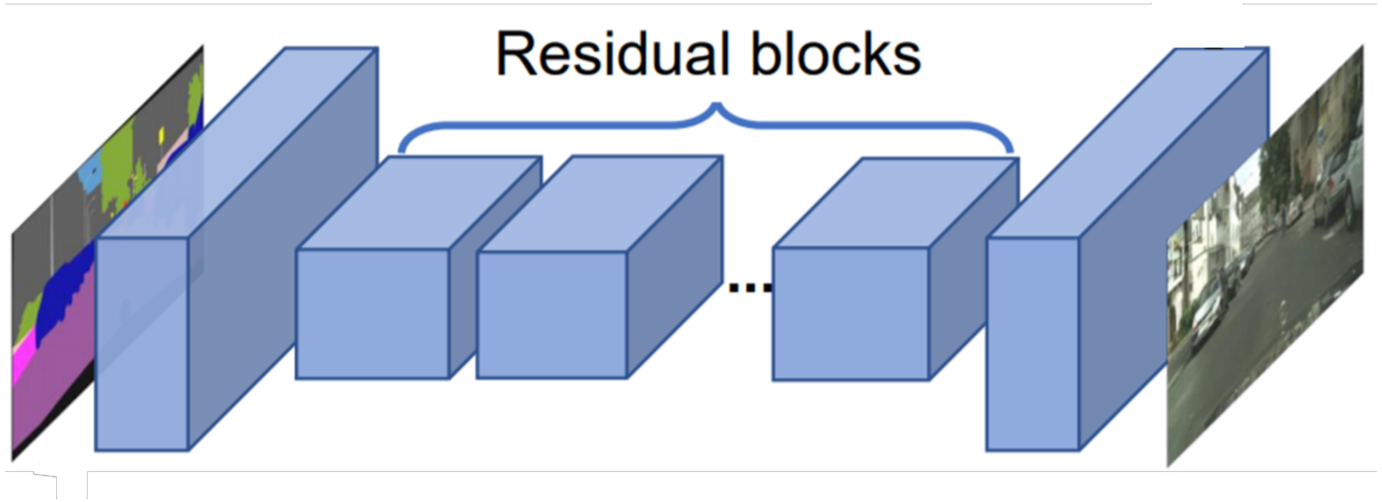
Introduction



Introduction



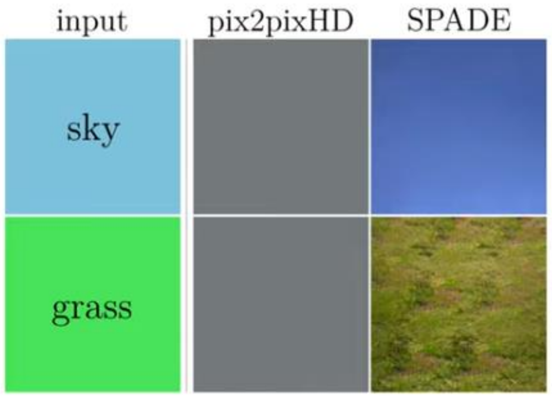
Pix2Pix



- Previous methods(pix2pix based model) directly feed the semantic layout as input to the deep network, which is then processed through stacks of convolution, normalization, and nonlinearity layers.
- The normalization layers tend to "wash away" semantic information.

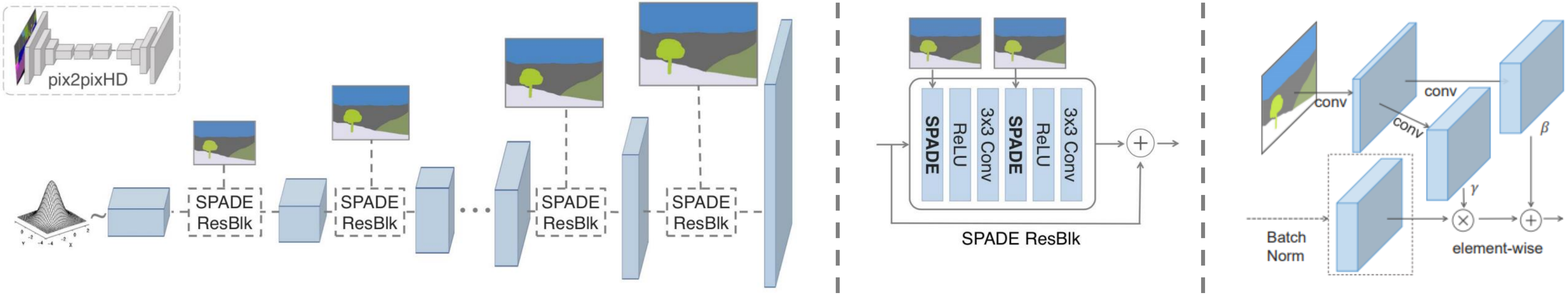
SPADE

Why does SPADE work better?



- Specifically, while normalization layers such as the InstanceNorm are essential pieces in almost all the state-of-the-art conditional image synthesis models, they tent to wash away semantic information when applied to uniform of flat segmentation masks.
- Let us consider a simple module that first applies convolution to a segmentation mask and then normalization. Furthermore, let us assume that a segmentation mask with a single label is given as input to the module.
- After applying InstanceNorm, the normalized activation will become all zeros no matter what the input semantic label is given

SPADE



- The paper proposes spatially-adaptive de-normalization(SPADE), a simple but effective layer for synthesizing photorealistic images given an input semantic layout.
- The segmentation mask is fed through spatially adaptive modulation without normalization, so SPADE can better preserve semantic information.
- SPADE allows user to control over both semantic and style, and to produce multi-modal synthesis

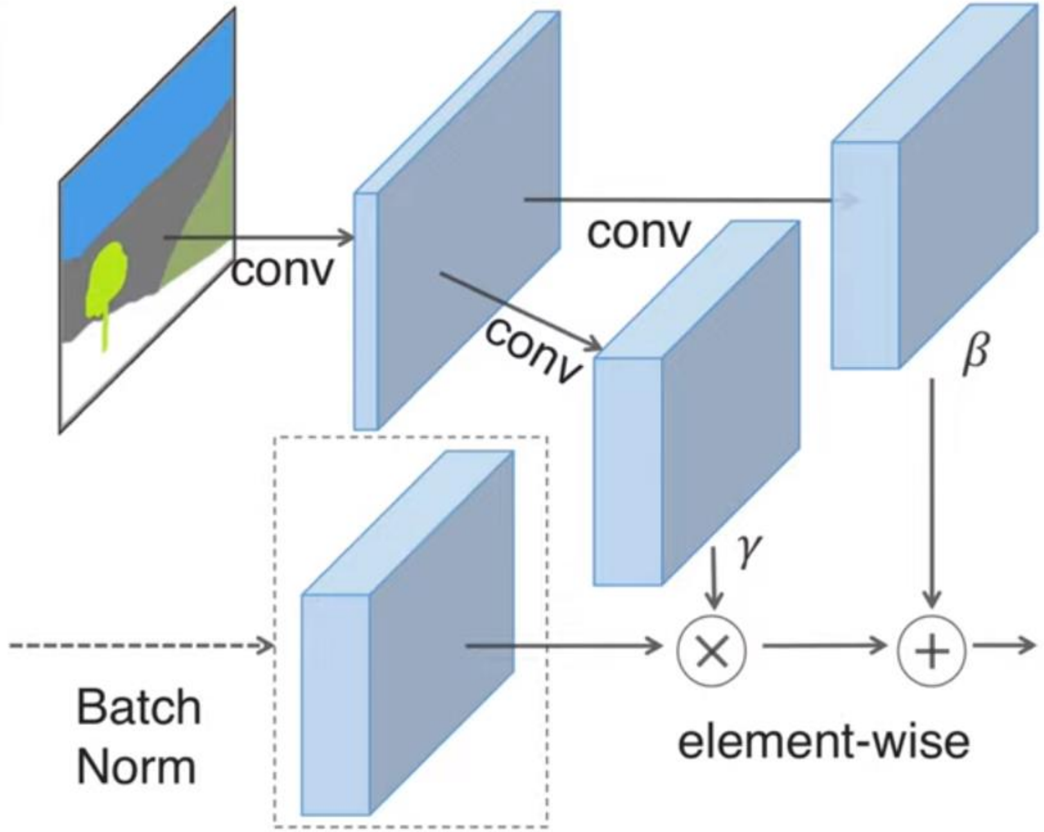
SPADE

$$\boxed{\overset{\text{external data}}{\gamma^i_{c,y,x}(\mathbf{m})}\,\overset{\text{Normalization}}{\frac{h^i_{n,c,y,x}-\mu^i_c}{\sigma^i_c}}+\overset{\text{external data}}{\beta^i_{c,y,x}(\mathbf{m})}}$$

$$\mu^i_c = \frac{1}{NH^iW^i}\sum_{n,y,x} h^i_{n,c,y,x}$$

$$\sigma^i_c = \sqrt{\frac{1}{NH^iW^i}\sum_{n,y,x}(h^i_{n,c,y,x})^2 - (\mu^i_c)^2}.$$

$$(n \in N, c \in C^i, y \in H^i, x \in W^i)$$



- SPADE enjoys the benefit of normalization without losing the semantic input information.

SPADE

Training

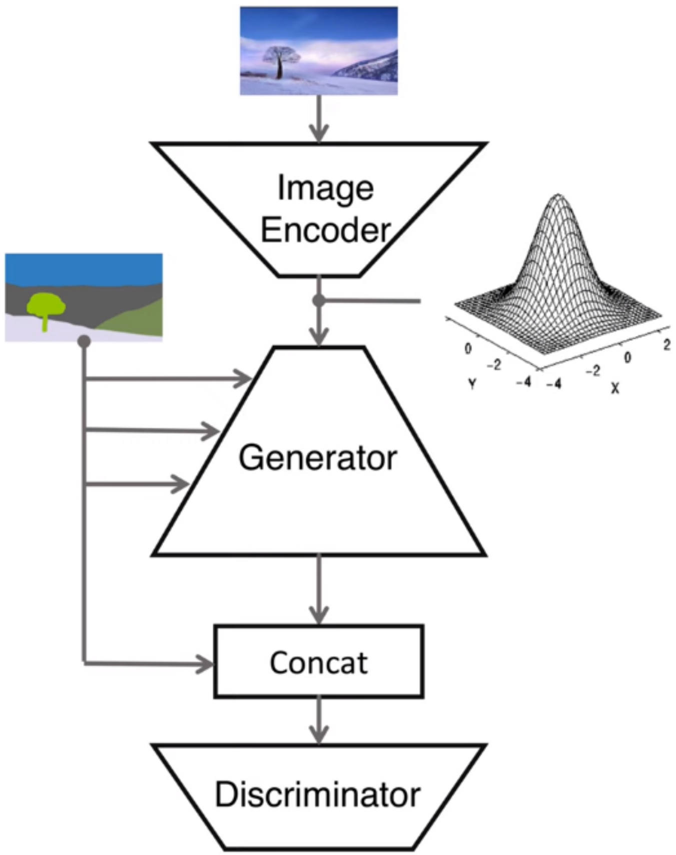
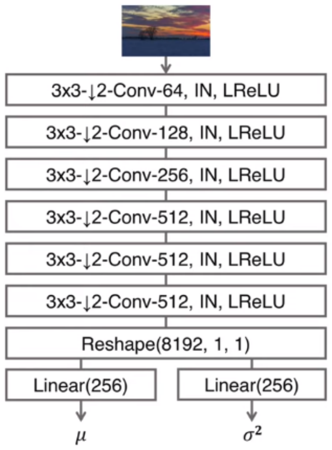
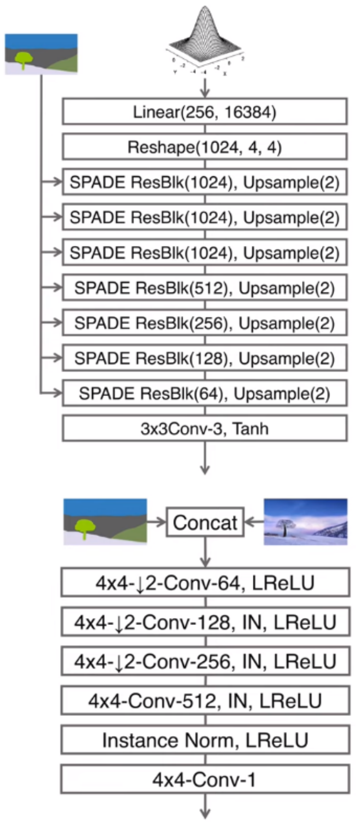


Image reconstruction



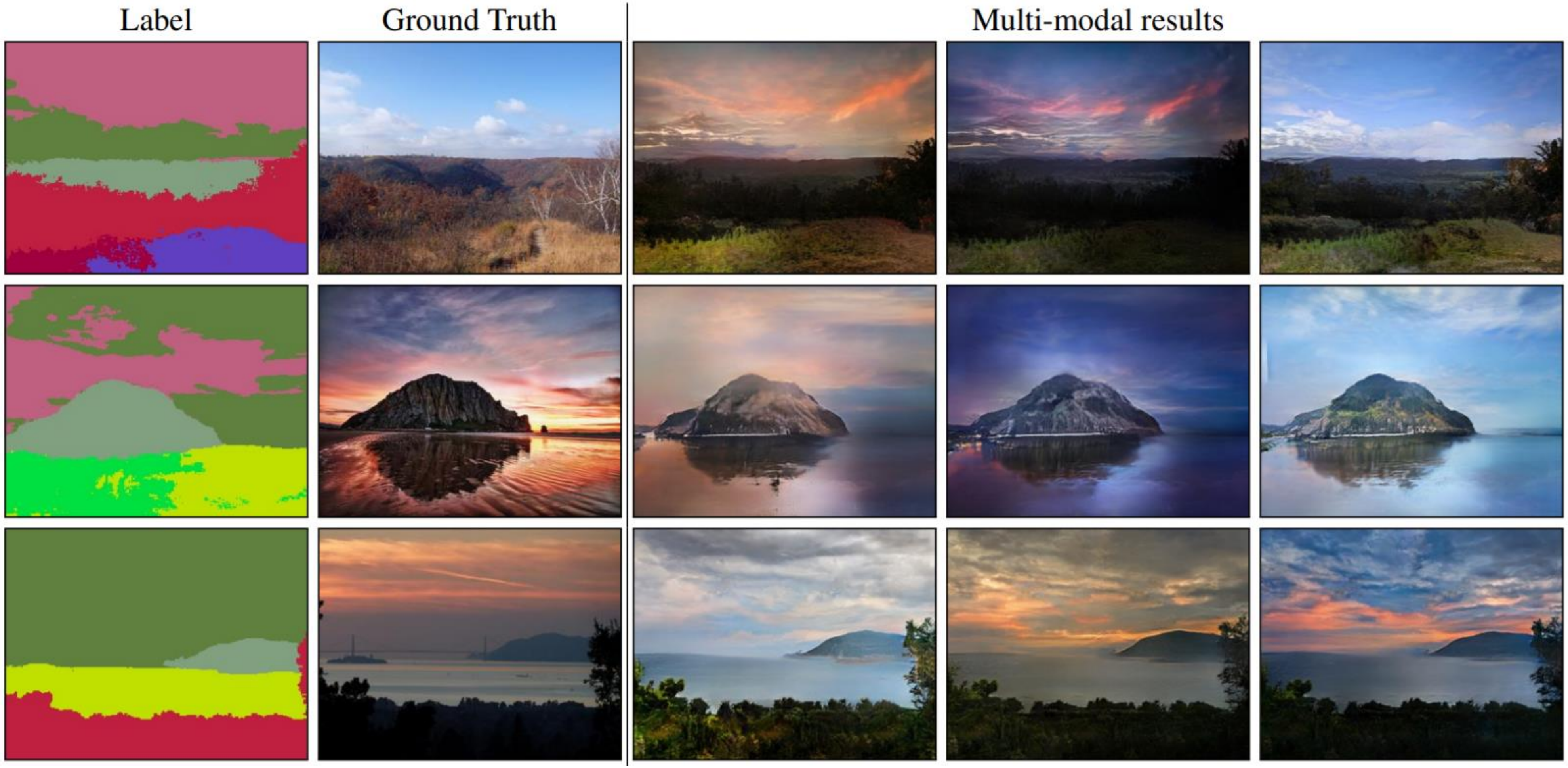


$$\min_G((\max_D(L_{GAN}) + L_{Percep} + L_{FM} + L_{KL})$$

$$L_{FM} = \mathbb{E}_{(\mathbf{s},\mathbf{x})} \sum_{i=1}^{T} \frac{1}{N_i} [||D^{(i)}(\mathbf{s},\mathbf{x}) - D^{(i)}(\mathbf{s},G(\mathbf{s}))||_1],$$

$$L_{Percep} = \lambda \sum_{i=1}^{N} \frac{1}{M_i} [||F^{(i)}(\mathbf{x}) - F^{(i)}(G(\mathbf{s}))||_1]$$

SPADE



[Multi-modal synthesis results on the Flickr Landscapes Dataset]

SPADE

Method	COCO-Stuff			ADE20K			ADE20K-outdoor			Cityscapes		
	mIoU	accu	FID	mIoU	accu	FID	mIoU	accu	FID	mIoU	accu	FID
CRN [6]	23.7	40.4	70.4	22.4	68.8	73.3	16.5	68.6	99.0	52.4	77.1	104.7
SIMS [43]	N/A	N/A	N/A	N/A	N/A	N/A	13.1	74.7	67.7	47.2	75.5	**49.7**
pix2pixHD [48]	14.6	45.8	111.5	20.3	69.2	81.8	17.4	71.6	97.8	58.3	81.4	95.0
Ours	**37.4**	**67.9**	**22.6**	**38.5**	**79.9**	**33.9**	**30.8**	**82.9**	**63.3**	**62.3**	**81.9**	71.8

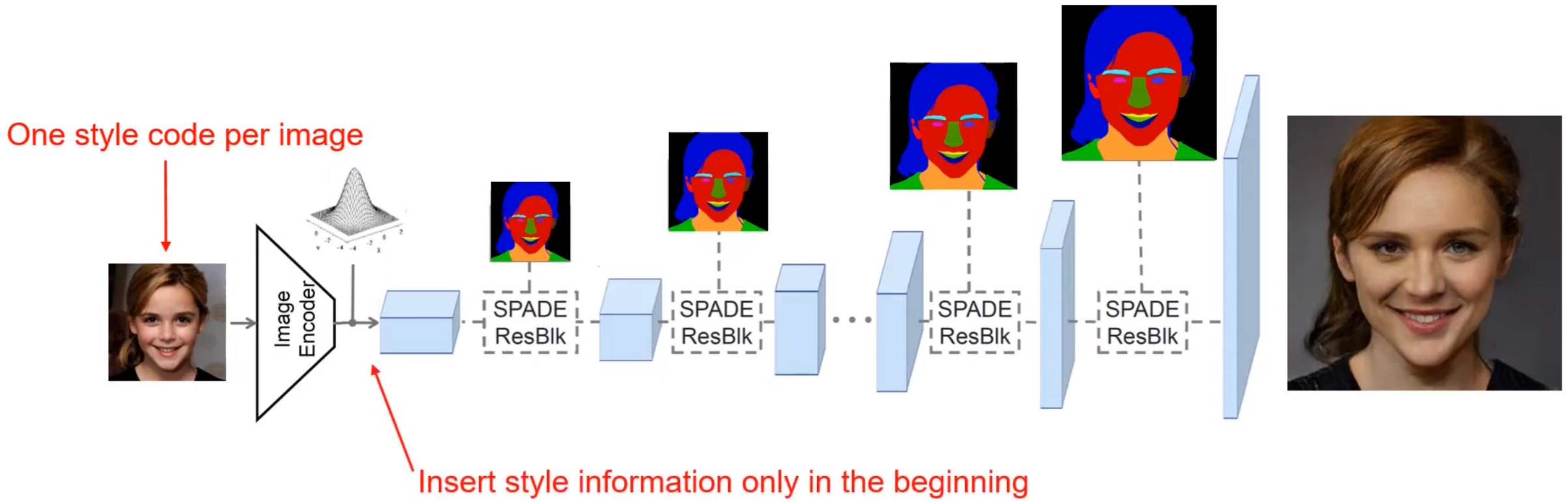
[Performance of segmentation model on the synthesized images and FID score]

Method	#param	COCO.	ADE.	City.
decoder w/ SPADE (Ours)	96M	**35.2**	38.5	62.3
compact decoder w/ SPADE	61M	**35.2**	38.0	**62.5**
decoder w/ Concat	79M	31.9	33.6	61.1
pix2pixHD++ w/ SPADE	237M	34.4	**39.0**	62.2
pix2pixHD++ w/ Concat	195M	32.9	38.9	57.1
pix2pixHD++	183M	32.7	38.3	58.8
compact pix2pixHD++	103M	31.6	37.3	57.6
pix2pixHD [48]	183M	14.6	20.3	58.3

[Ablation study with pix2pixHD]

SEAN

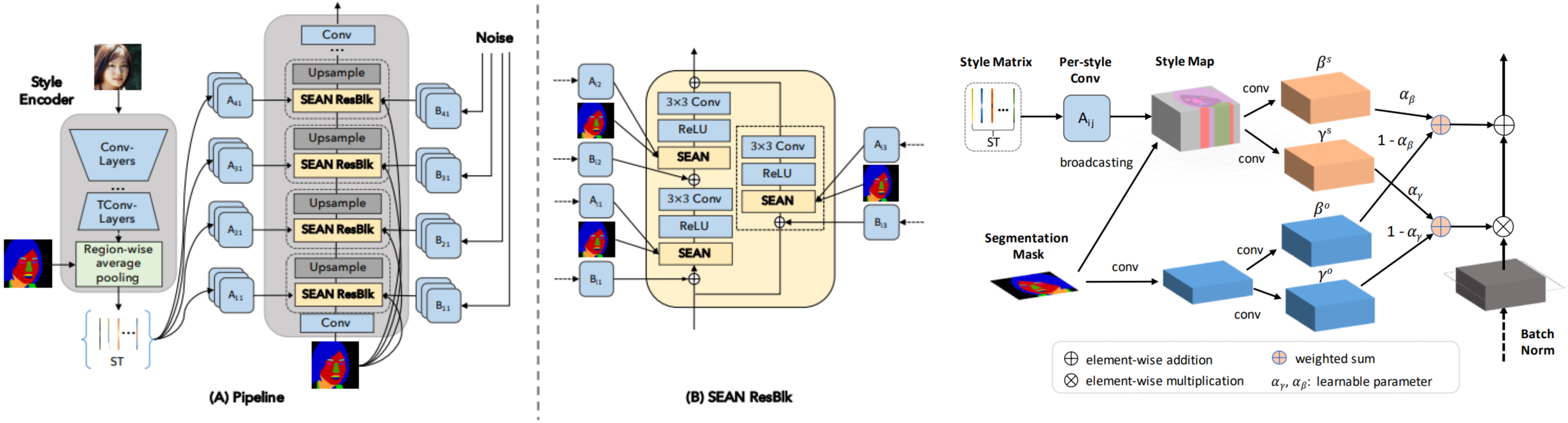
Two shortcomings



SEAN

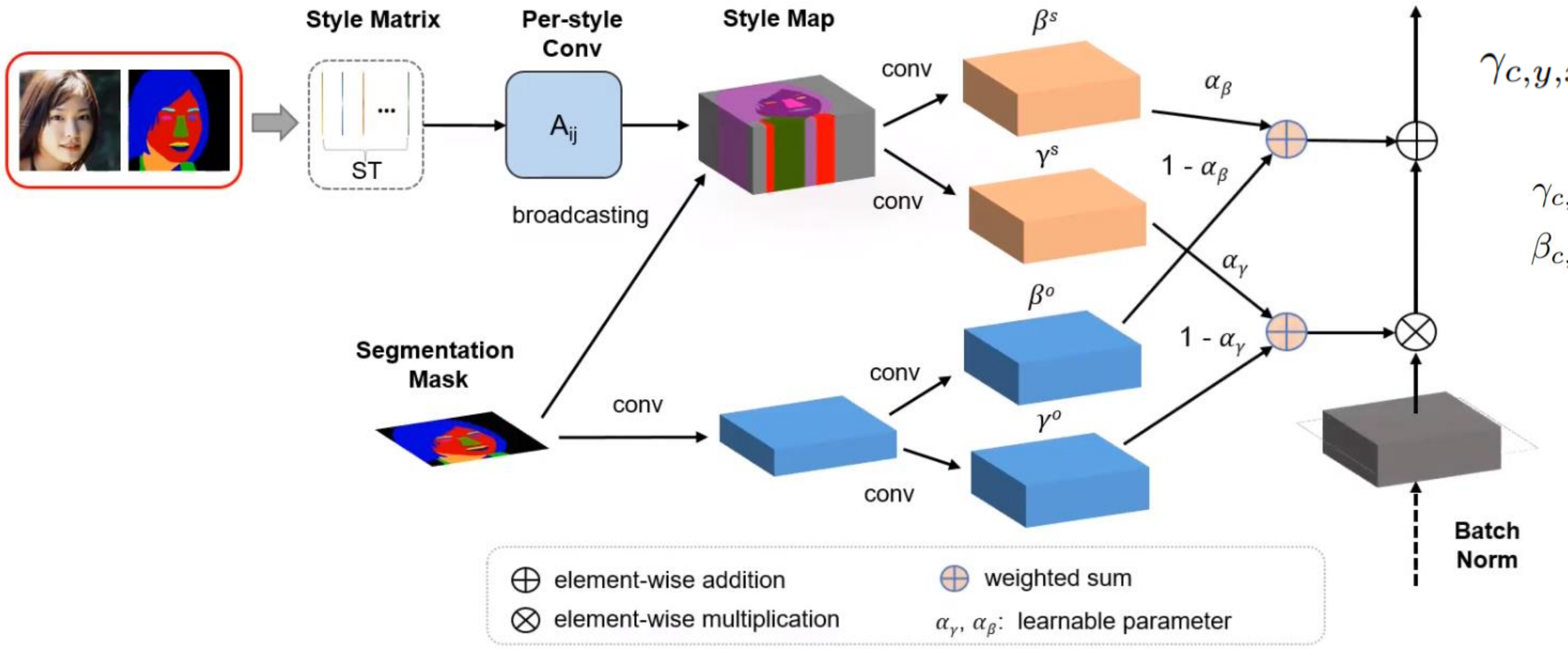
Two modifications
One style code per region
Image Encoder
SPADE ResBlk
SPADE ResBlk
SPADE ResBlk
SPADE ResBlk
Inject style information at multiple locations

SEAN



- The paper proposes semantic region-adaptive normalization(SEAN), a simple but effective building block for GAN conditioned on segmentation masks.
- Using SEAN normalization, we can build a network architecture that can control the style of each semantic region individually.
- We can interactively edit images by changing segmentation masks or the style for any given region.

SEAN



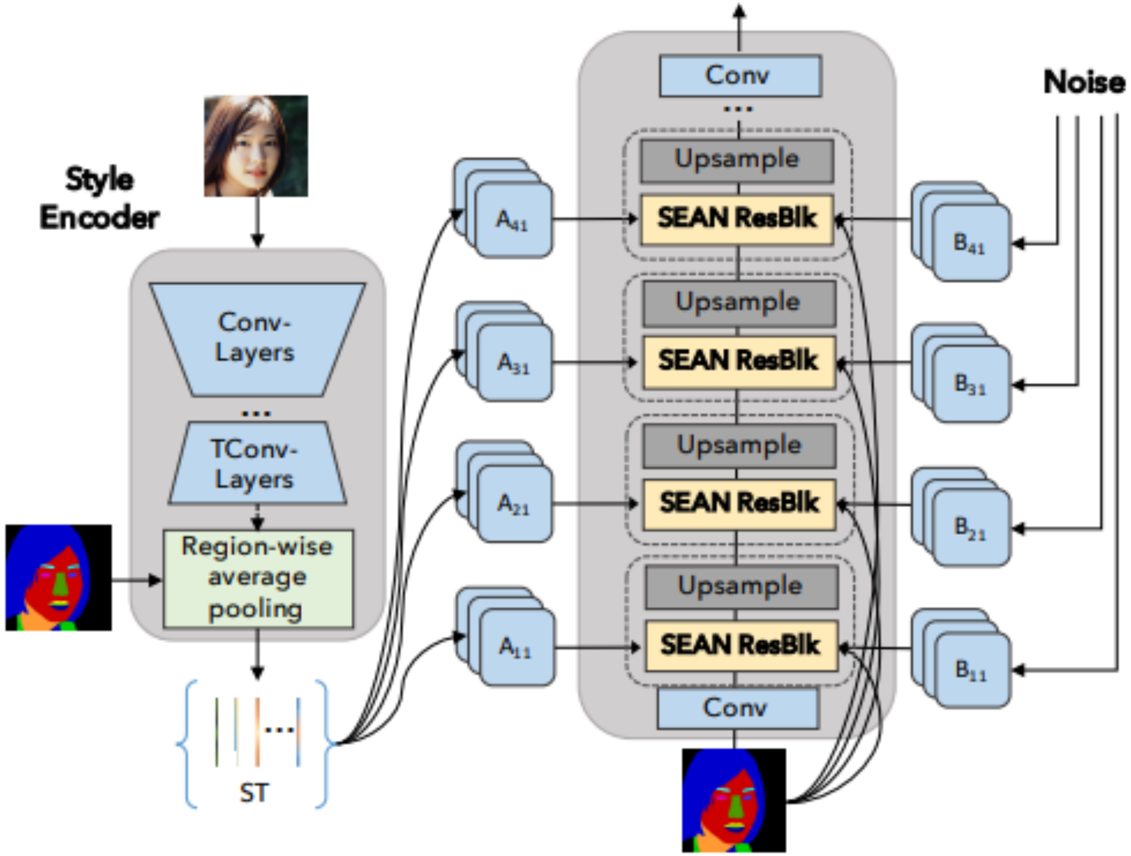
$$\gamma_{c,y,x}(\mathbf{ST},\mathbf{M})\frac{h_{n,c,y,x}-\mu_c}{\sigma_c}+\beta_{c,y,x}(\mathbf{ST},\mathbf{M})$$

$$\gamma_{c,y,x}(\mathbf{ST},\mathbf{M})=\alpha_\gamma\gamma^s_{c,y,x}(\mathbf{ST})+(1-\alpha_\gamma)\gamma^o_{c,y,x}(\mathbf{M})$$
$$\beta_{c,y,x}(\mathbf{ST},\mathbf{M})=\alpha_\beta\beta^s_{c,y,x}(\mathbf{ST})+(1-\alpha_\beta)\beta^o_{c,y,x}(\mathbf{M})$$

$$\mu_c=\frac{1}{NHW}\sum_{n,y,x}h_{n,c,y,x}$$

$$\sigma_c=\sqrt{\frac{1}{NHW}\left(\sum_{n,y,x}h^2_{n,c,y,x}\right)-\mu_c^2}$$

SEAN

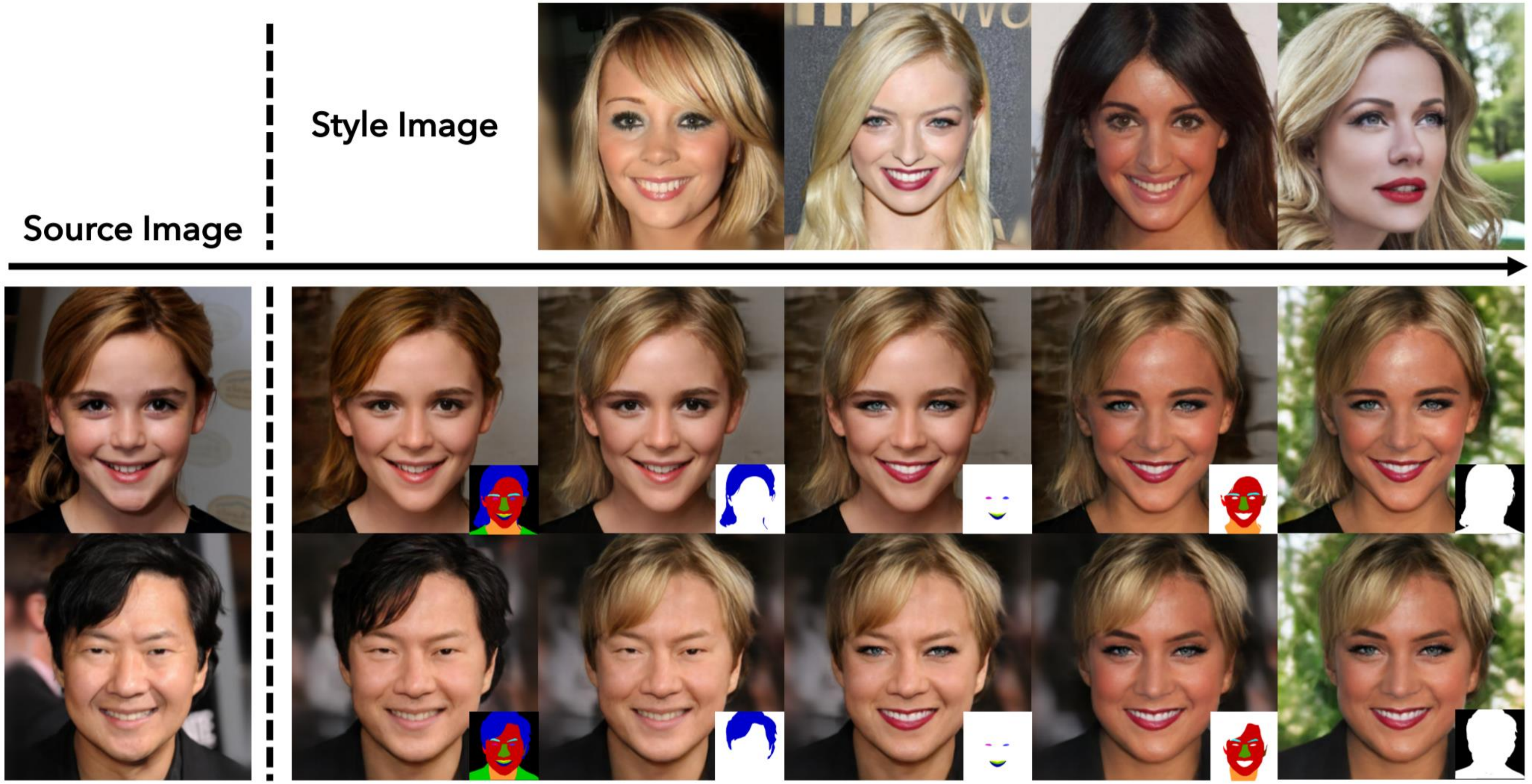


$$\min_{E,G}\left(\left(\max_{D_1,D_2}\sum_{k=1,2}\mathcal{L}_{\text{GAN}}\right)+\lambda_1\sum_{k=1,2}\mathcal{L}_{\text{FM}}+\lambda_2\mathcal{L}_{\text{percept}}\right)$$

$$\mathcal{L}_{FM}=\mathbb{E}\sum_{i=1}^{T}\frac{1}{N_i}\left[\left\|D_k^{(i)}(\mathbf{R},\mathbf{M})-D_k^{(i)}(G(\mathbf{ST},\mathbf{M}),\mathbf{M})\right\|_1\right]$$

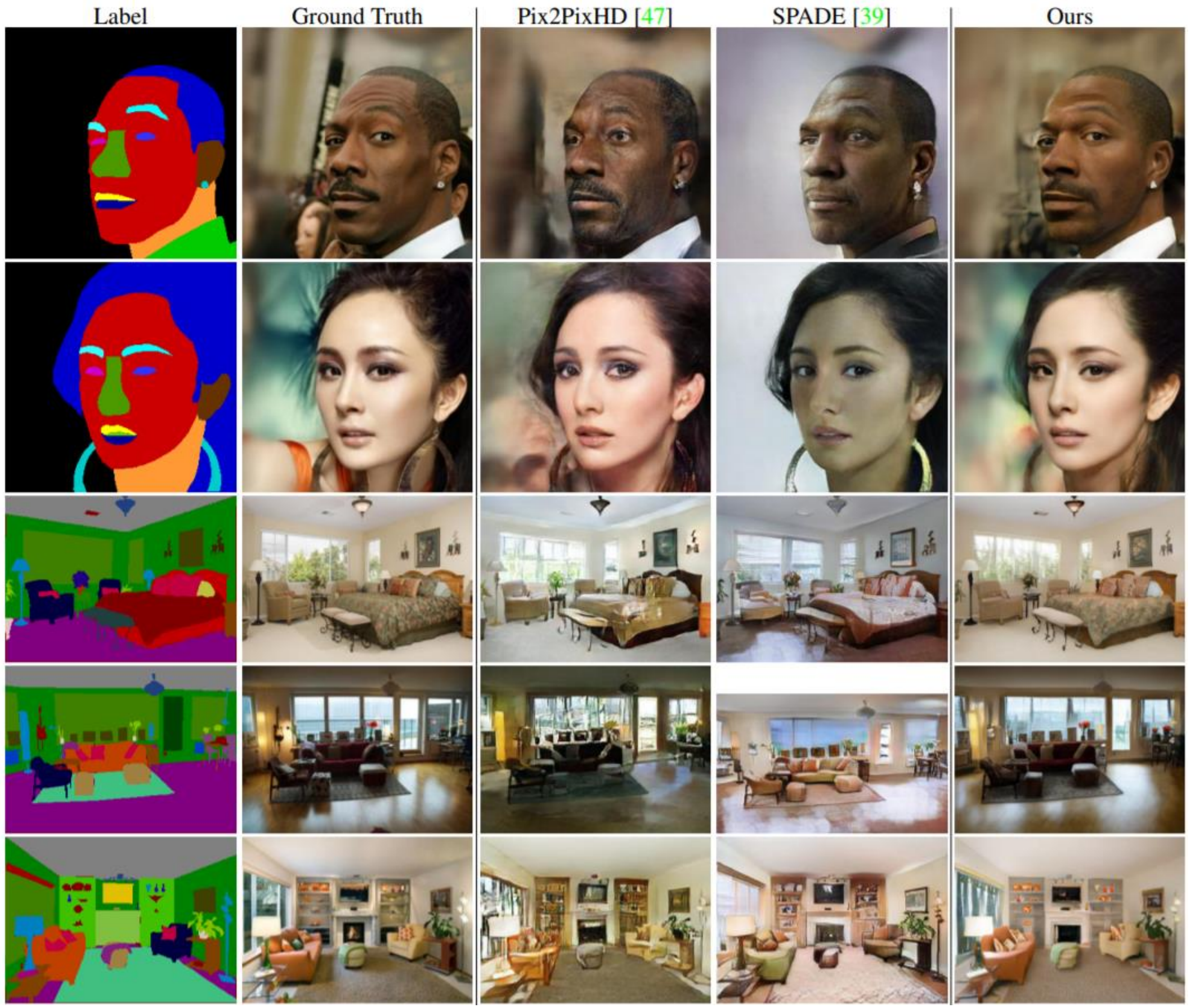
$$\mathcal{L}_{percept}=\mathbb{E}\sum_{i=1}^{N}\frac{1}{M_i}\left[\left\|F^{(i)}(\mathbf{R})-F^{(i)}(G(\mathbf{ST},\mathbf{M}))\right\|_1\right]$$

SEAN



[Face image editing controlled via style images and segmentation masks]

SEAN



[Visual comparison of semantic image synthesis results]

SEAN

Method	CelebAMask-HQ			CityScapes			ADE20K			Façades
	mIoU	accu	FID	mIoU	accu	FID	mIoU	accu	FID	FID
Ground Truth	73.14	94.38	9.41	66.21	93.69	32.34	39.38	78.76	14.51	14.40
Pix2PixHD [47]	76.12	95.76	23.69	50.35	92.09	83.24	22.78	73.32	43.0	22.34
SPADE [39]	**77.01**	**95.93**	22.43	56.01	93.13	60.51	**35.37**	**79.37**	34.65	24.04
Ours	75.69	95.69	**17.66**	**57.88**	**93.59**	**50.38**	34.59	77.16	**24.84**	**19.82**

[Performance of segmentation model on the synthesized images and FID score]

EOD