Dual Pyramid Generative Adversarial Networks for Semantic Image Synthesis

Code & Video

Shijie Li[1], Ming-Ming Cheng[2], Juergen Gall[1]

1 UNI BONN 2

BMVC 2022

❑ Motivation

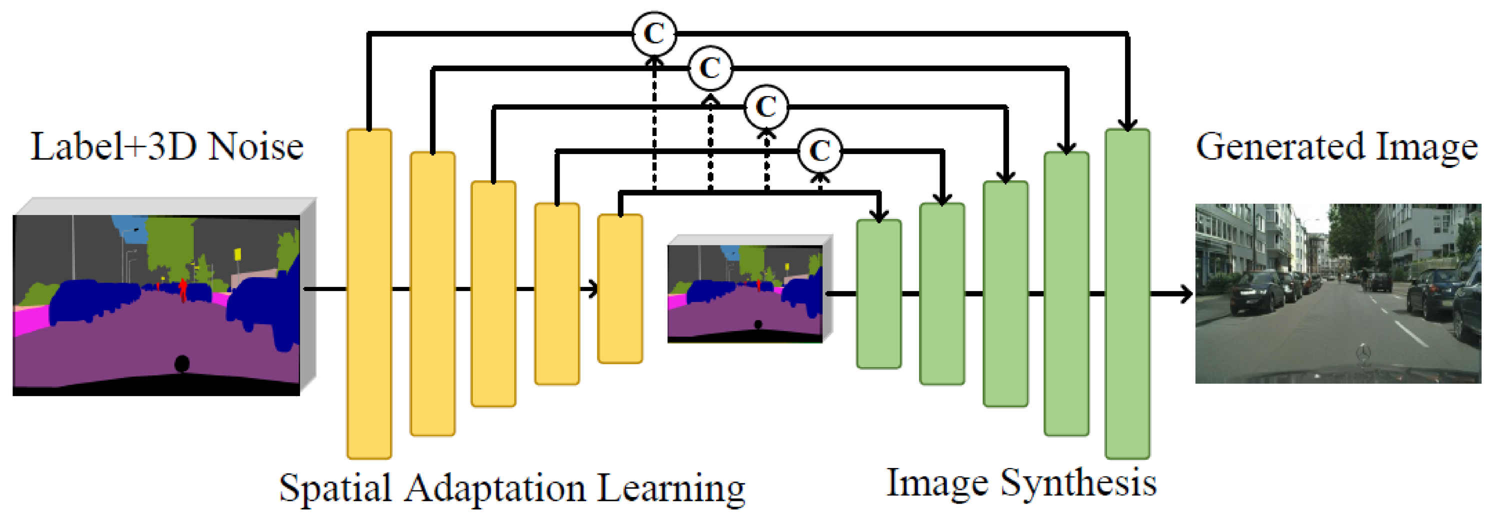
Most semantic image synthesis methods struggle to generate realistic objects as they cannot handle scale information properly. We address this issue by enhancing the multi-scale ability for both generator and discriminator. The approach thus generates more realistic

❑ Contribution

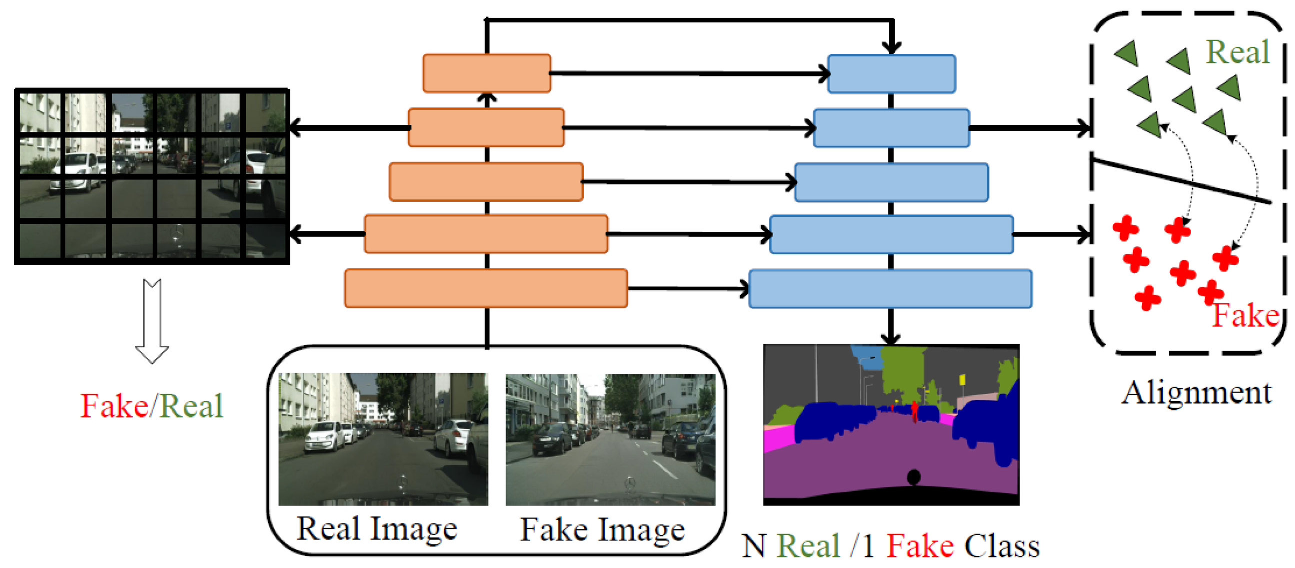
- We propose a **dual pyramid generator** for semantic image synthesis which adapts the conditioning to the size of the objects.
- We propose to unify supervision at **pixel**, **patch**, and **feature** level to enforce the generator to generate realistic objects that are well aligned with the semantic maps.
- State-of-the-art qualitative and quantitative results on **3 datasets**

❑ Architecture

- **Dual Pyramid Generator**



- **Scale-Enhancement Discriminator**



❑ Dual Pyramid Generator

- Spatially-adaptive normalization (SPADE)

$$\gamma^i_{x,y,c}(\mathbf{l}^i)\frac{h^i_{x,y,c,n}-\mu^i_c}{\sigma^i_c}+\beta^i_{x,y,c}(\mathbf{l}^i)$$

- Supervision

$$\mathcal{L}_G = -\mathbf{E}_{(\mathbf{z},\mathbf{l})}\left[\sum_{c=1}^{N}\alpha_c\sum_{x,y}^{H\times W}\mathbf{l}_{x,y,c}\log D(G(\mathbf{z},\mathbf{l}))_{x,y,c}\right]$$
$$-\frac{1}{L}\sum_{i=1}^{L}\mathbf{E}_{(\mathbf{z},\mathbf{l})}\left[\min(-1+D^i_p(\psi^i(G(\mathbf{z},\mathbf{l}))),0)\right]+\mathcal{L}_{fm}$$

❑ Scale-Enhancement Discriminator

We utilize supervisions at different levels to boost the ability of discriminator to handle multi-scale information

- Pixel-level

$$\mathcal{L}_{pixel} = -\mathbf{E}_{(\mathbf{x},\mathbf{l})}\left[\sum_{c=1}^{N}\alpha_c\sum_{x,y}^{H\times W}\mathbf{l}_{x,y,c}\log D(\mathbf{x})_{x,y,c}\right]-\mathbf{E}_{(\mathbf{z},\mathbf{l})}\left[\sum_{x,y}^{H\times W}\log D(G(\mathbf{z},\mathbf{l}))_{x,y,c=N+1}\right] \quad \alpha_c = E_{\mathbf{l}}\left[\frac{H\times W}{\sum_{x,y}^{H\times W}\mathbf{l}_{x,y,c}}\right]$$

- Patch-level

$$\mathcal{L}^i_{ms} = -\mathbf{E}_{\mathbf{x}}\left[\min(-1+D^i_p(\psi^i(\mathbf{x})),0)\right]-\mathbf{E}_{(\mathbf{z},\mathbf{l})}\left[\min(-1-D^i_p(\psi^i(G(\mathbf{z},\mathbf{l}))),0)\right]$$

- Feature-level

$$\mathcal{L}^i_{fm} = \mathbf{E}_{(\mathbf{x},\mathbf{l},\mathbf{z})}\left[\frac{\sum_{x,y}^{H^i\times W^i}\left\|\phi^i(\mathbf{x})_{x,y}-\phi^i(G(\mathbf{z},\mathbf{l}))_{x,y}\right\|_2^2}{C^i\times H^i\times W^i}\right]$$

❑Experiments

- **Qualitative Evaluation**

Methods	Cityscapes		ADE20K		ADE20K-Outdoor	
	FID ↓	mIoU ↑	FID ↓	mIoU ↑	FID ↓	mIoU ↑
CRN [3]	104.7	52.4	73.3	22.4	99.0	16.5
pix2pixHD [30]	95.0	58.3	81.8	20.3	97.8	17.4
SPADE [21]	71.8	62.3	33.9	38.5	63.3	30.8
DAGAN [27]	60.3	66.1	31.9	40.5	N/A	N/A
LGGAN [28]	57.7	68.4	31.6	41.6	N/A	N/A
CC-FPSE [17]	54.3	65.5	31.7	43.7	N/A	N/A
SIMS [22]	49.7	47.2	N/A	N/A	67.7	13.1
OASIS [25]	47.7	69.3	28.3	48.8	48.6	**40.4**
DP-GAN	**44.1**	**73.6**	**26.1**	**52.7**	**45.8**	**40.4**

Comparison to state-of-the-art methods on different datasets.

	road	swalk	build	wall	fence	pole	tlight	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	obj-mIoU
SPADE [21]	97.5	80.8	88.5	54.3	50.6	40.4	39.0	41.9	88.7	69.1	92.0	66.2	41.5	89.1	64.6	73.2	42.1	29.7	61.5	53.6
DAGAN [27]	97.4	80.0	89.0	60.1	53.7	41.2	39.4	46.5	88.9	65.9	92.5	66.8	45.8	89.9	71.2	75.4	57.0	25.8	60.9	56.4
CC-FPSE [17]	97.7	82.8	**89.8**	56.1	61.3	42.3	41.8	50.4	**89.6**	69.3	92.5	68.5	48.3	90.2	69.7	74.3	45.4	43.4	65.0	58.1
LGGAN [28]	**97.8**	**83.1**	89.7	59.8	56.0	42.5	42.8	50.5	89.5	70.0	**92.7**	**69.0**	48.6	**90.6**	72.2	80.2	52.4	38.8	64.0	59.2
OASIS [25]	96.9	79.2	85.1	70.3	64.2	41.6	50.7	49.9	85.0	**74.8**	92.0	64.9	54.0	88.4	65.6	79.9	63.4	53.9	63.7	61.5
DP-GAN	97.5	81.9	87.2	**71.4**	**72.7**	**46.9**	**55.5**	**60.3**	87.3	72.9	92.4	67.4	**55.5**	89.9	**81.5**	**83.1**	**73.9**	**55.3**	**66.9**	**66.9**

Per-class IoU for Cityscapes, obj-mIoU is mIoU only for object classes.

- **Quantitative Evaluation**

(a) Label (b) Ground Truth (c) SPADE (d) OASIS (e) DP-GAN

Generated images from ADE20k dataset

Cropped objects from generated images (Cityscape)

- **Architecture Ablation**

(a) Gen / Dis

Gen	Dis	FID	mIoU	obj-mIoU
OA	OA	47.7	69.3	61.5
OA	DP	47.9	**74.0**	**67.4**
DP	OA	45.4	69.9	62.0
DP	DP	**44.1**	73.6	66.9

(b) $\mathcal{L}_{ms}$ in (5)

Enc	Dec	FID	mIoU	obj-mIoU
		49.2	67.9	59.5
	✓	44.5	72.1	64.4
✓	✓	44.3	72.8	66.4
✓		**44.1**	**73.6**	**66.9**

(c) $\mathcal{L}_{fm}$ in (6)

Enc	Dec	FID	mIoU	obj-mIoU
		44.1	69.9	62.1
✓		44.4	69.2	60.8
✓	✓	45.0	**73.8**	66.8
	✓	**44.1**	73.6	**66.9**

(a) Label (b) OA-OA (c) DP-OA (d) OA-DP (e) DP-DP

DP or OA denote if the generator or discriminator from OASIS (OA) or our approach (DP) are used