Semantic-Guided Inpainting Network for Complex Urban Scenes Manipulation

Pierfrancesco Ardino, Yahui Liu, Elisa Ricci, Bruno Lepri, Marco De Nadai

https://github.com/PierfrancescoArdino/SGINet

What is a Image Manipulation?

The art of transforming an image to convey what you want

In this work we focus on image manipulation of mixed scenes and in particular on two sub-tasks:

- **Object Insertion**
- **Object removal or image inpainting**

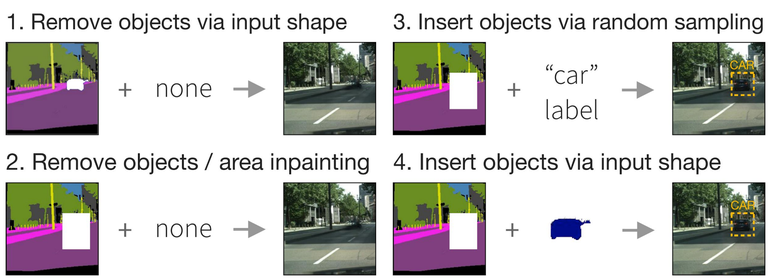
What are Mixed Scenes?

Mixed scenes contain objects and background of different classes

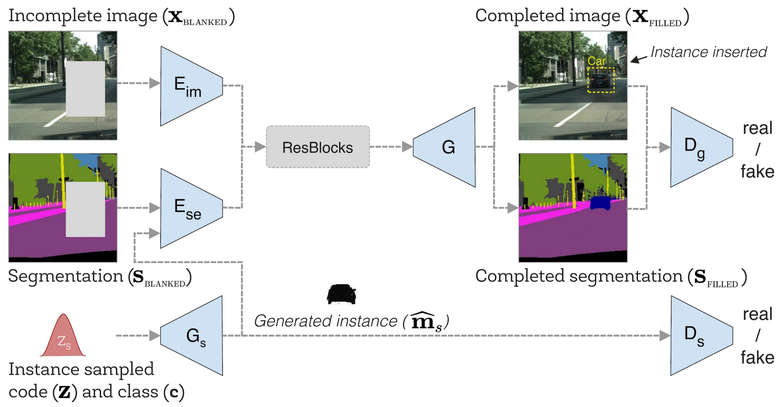
Why traditional methods struggle?

- Missing part with different semantics
- Focused either on object insertion or image inpainting

Use cases



Architecture

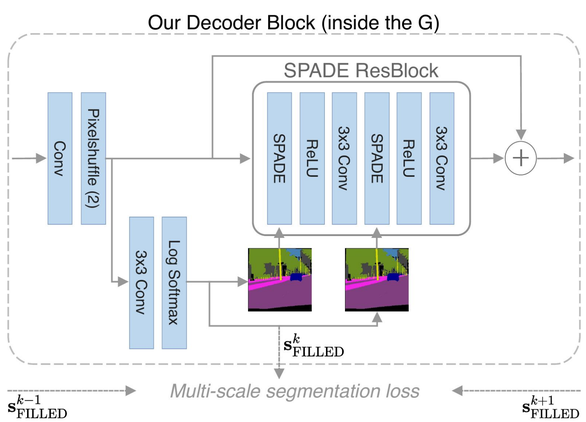


Instances are generated using a **variational autoencoder** or via **input shape** and then used in the one stage inpainting network as additional conditioning.

A unified framework for multiple use case scenarios

Decoder built on top of SPADE [1], For this reason we have developed a novel Decoder block that uses **predicted** semantic segmentation instead of **ground truth**.

Decoder Block



Metrics

We measure image quality through the Peak Signal-to-Noise Ratio (***PSNR***) and the Fréchet Inception Distance (***FID***). Higher PSNR and Lower FID values indicate higher image quality. To measure the presence of inserted objects we use the ***F1*** metric through the use of a pre-trained detection network. Higher scores indicate that the network is inserting synthetic objects that resemble the real ones.

In the ***Restore*** setting we perform an image inpainting task, in the ***Place*** setting we test the models for the task of object insertion and image inpainting together.

	Model	Cityscapes			Indian Driving		
		PSNR↑	FID↓	F1↑	PSNR↑	FID↓	F1↑
Restore	Hong *et al.* [12]	31.07	7.26	0.00	30.31	6.34	0.00
	SPG-Net [24]	31.36	7.97	0.00	29.95	6.39	0.02
	RN [8]	32.16	9.64	0.00	29.83	11.14	0.02
	Our proposal	**32.95**	**5.08**	**0.06**	**30.97**	**5.45**	**0.05**
Place	Hong *et al.* [12]	31.08	7.26	0.10	30.32	6.32	0.91
	SPG-Net*	31.37	7.96	0.60	29.94	6.38	0.87
	RN*	31.74	9.79	0.54	29.62	10.76	0.71
	Our proposal	**32.96**	**5.05**	**0.91**	**30.98**	**5.43**	**0.97**

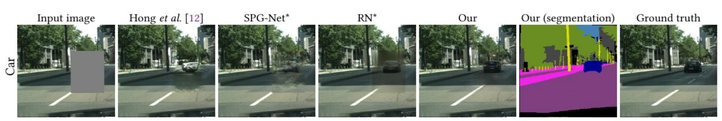
Ablation

Model	PSNR↑	FID↓
Our proposal (A)	**32.96**	**5.05**
(A) w/o $\mathcal{L}_{\text{style}}$	32.68	5.38
(A) w/o $\mathcal{L}_{\text{FM}}$	32.66	5.30
(A) w/o E_{se} and s_{BLANKED}	32.35	5.42
(A) w/o SPADE	32.57	5.56
(A) using DeepLabv3 [48] segmentation	32.85	5.11

Qualitative results

- **Restore**

- **Place**



[1]Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. "Semantic Image Synthesis with Spatially-Adaptive Normalization", in CVPR, 2019.