INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

WACV WAIKOLOA HAWAII JAN 3-7 • 2023

GRG Graphics Research Group

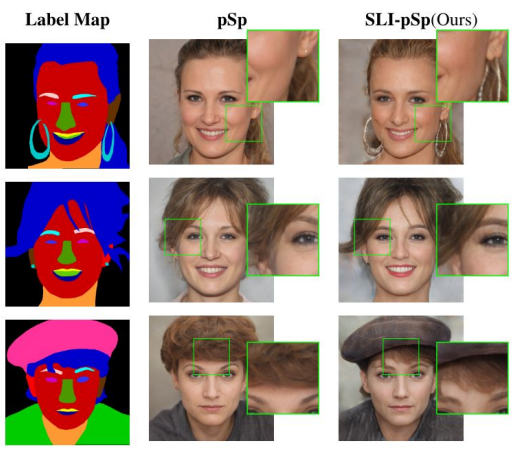
SLI-pSp: Injecting Multi-Scale Spatial Layout in pSp

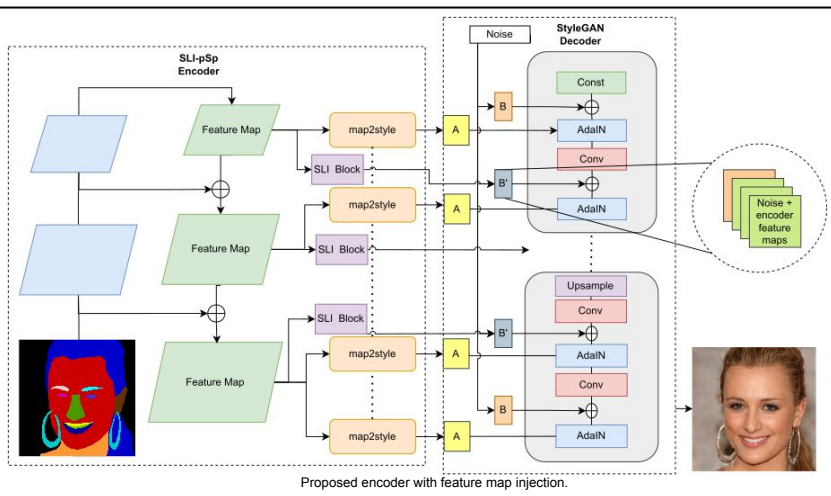
Aradhya N. Mathur — IIITD
Anish Madan — IIITD
Ojaswa Sharma — IIITD

{aradhyam, anish16223, ojaswa}@iiitd.ac.in

Abstract

We propose SLI-pSp, a general purpose Image-to-Image (I2I) translation model that encodes spatial layout information as well as style in the generator, using pSp as the base architecture. Previous methods like pSp have shown promising results by leveraging StyleGAN as a generator in various I2I tasks but they seem to miss finer or under-represented details in facial images like earrings and caps, and break down on complex datasets due to their solely global approach. We present a simple method to circumvent these issues without losing multi-modal properties of StyleGANs.





Proposed encoder with feature map injection.

Approach

- We propose a technique termed Spatial Layout Injection (SLI-pSp) that encodes spatial layout information in the input image in the StyleGAN generator along with style.
- We do so without modifying the style vector injection in the generator through pSp's *map2style* network, but rather by combining SLI with noise layers in the StyleGAN generator at multiple spatial scales.
- Such an approach helps preserve global aspects of image generation as well as enhance spatial layout details in the output. We experiment on several challenging datasets and across several I2I tasks that highlight the effectiveness of our approach over previous methods with respect to finer details in the generated image and overall visual quality.
- We extract feature maps using a feature pyramid network which are then propagated via map2style and SLI blocks and injected into the StyleGAN generator.

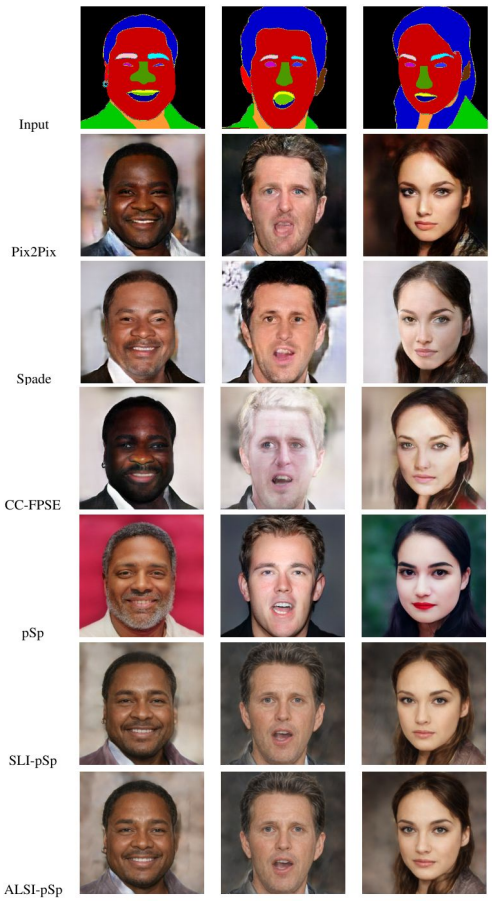
Method

- Let the encoder feature maps be denoted as E_i, where $i \in \{16,32,64\}$ corresponding to the spatial scales of the feature maps. These are generated using feature pyramid over a ResNet backbone.
- Let the noise layers in the StyleGAN generator be represented by N_j, where $j \in \{4,8,16,...,1024\}$ represents spatial sizes.
- The combined spatial layout feature maps and noise added to the generator, B' can we written as

$$B' = concat(conv(E_i), N_j) \quad i \in \{16,32,64\},$$

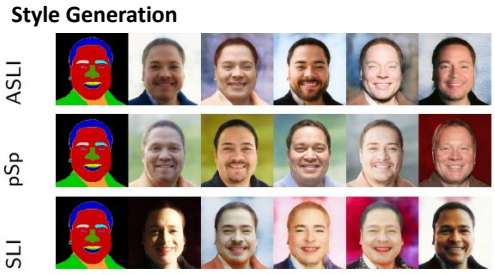
were *conv* represents a convolution layer and *concat* operation concatenates along the channel dimension as spatial sizes are the same.

Seg2Face



Method	LPIPS Loss ↓	FID ↓	mIoU ↑
pSp	0.35	53.90	0.61
SLI-pSp	**0.31**	37.32	**0.81**
ASLI-pSp	0.32	**36.89**	**0.81**

Style Generation



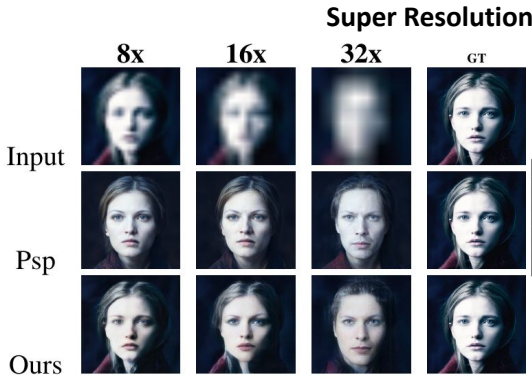
Edge2Image

Method	LPIPS Loss ↓	FID ↓
pSp	0.48	236.733
SLI-pSp	0.32	**23.71**
ASLI-pSp	**0.31**	28.71

Our method captures the finer details in structure in a profound manner.

Our method generates much reliable information in Places2 dataset due to feature injection.

Super Resolution



Method	LPIPS Loss ↓	L_2 Loss ↓	FID ↓
pSp(8x)	0.23	0.06	31.35
SLI-pSp (8x)	0.12	**0.01**	10.21
ASLI-pSp (8x)	**0.10**	**0.01**	**9.90**
pSp(16x)	0.24	0.06	32.53
SLI-pSp (16x)	0.18	**0.02**	18.33
ASLI-pSp (16x)	**0.17**	**0.02**	**18.01**

Quantitative Results for Super resolution task on CelebA. We evaluate on data with 8x and 16x downsampling. SLI variants show substantial improvements across various metrics.

Conclusion

We propose a simple fix ter–med Spatial Layout Injection (SLI) that encodes the spatial layout information from the input image and propagates it to the StyleGAN decoder with higher detail preservation.