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

Aradhya Neeraj Mathur*
IIITD
aradhyam@iiitd.ac.in

Anish Madan*
IIITD
anish16223@iiitd.ac.in

Ojaswa Sharma
IIITD
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. To address these shortcomings, 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.*

Generative Adversarial Networks(GANs) [9] have revolutionized the field of generative modelling. A variety of GAN architectures have been proposed in the recent past that generate images with excellent visual fidelity and photorealism. Converting an image in a source domain to a target domain while preserving the core content and adapting the style according to target domain is termed as Image to Image (I2I) translation [13]. Such I2I problems exist in various forms [25] and many solutions have been proposed in the following themes: semantic image synthesis [26, 49, 33, 21, 32], style transfer [34, 19], image inpainting [48, 28, 30], image super resolution [46, 41], etc.

Consider the task of semantic image synthesis. A fundamental question is to understand how well we transform a semantic map into a realistic RGB image while preserving input semantics. Many previous approaches try to exploit semantic layout information in the generator by proposing encoder-decoder architectures [13, 36]. A common pitfall in such methods seems to be that the network receives the semantic information once (in the form of input), therefore it is difficult to maintain that information throughout the generation process via the decoder. To alleviate this problem, SPADE [26] proposes spatially-adaptive normalization that helps modulate the activations at various layers of the network via spatially adaptive, learnable affine transformations. However, this approach is limited in terms of representational capacity. We address the issue of representational capacity by having a separate style vector injection as well as spatial layout injection in the generator.

Another key issue is how well does the discriminator architecture utilize the semantic layouts. Previous methods such as Pix2Pix [13] and SPADE [26] use a multi-scale PatchGAN discriminator that takes in a label map and an image as inputs and outputs a predicate (real or fake) as

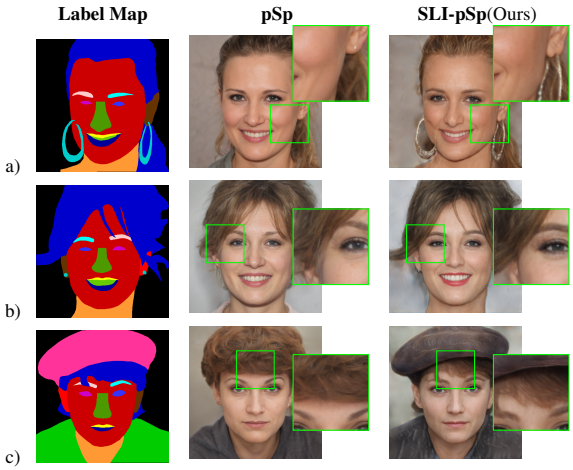

Figure 1. Our proposed **SLI-pSp** approach helps fix the lack of spatial context in the image generation process in pSp. By injecting spatial layout information in the decoder, we are able to synthesize finer and under-represented details (such as earrings, hats, and hairstyles) that are missed by pSp's global approach. We retain global properties of image generation even after our spatial layout injection (SLI), thereby making our method work much better across I2I tasks.

*These authors contributed equally to this work.

its decision. CC-FPSE [21] and OASIS [32] also propose variants of discriminators which give superior performance to previous methods. However, it is well known that GAN based architectures suffer from training instabilities due to the adversarial (min-max) nature of the learning problem. While many improvements [23, 42] have been proposed, many architectures still require lots of tuning to train properly. We address this issue by adopting and extending the pSp [29] framework that allows the use of a pretrained StyleGAN generator as a proxy for the discriminator.

One of the most popular GAN architectures in use is StyleGAN [17], due to its high quality, state-of-the-art image generation capabilities. pSp [29] proposes an I2I framework using StyleGAN as a generator. They introduce a Feature Pyramid based encoder which embeds input images into an intermediate latent space (separate from StyleGAN's latent space) and use pretraining of the StyleGAN generator in lieu of a discriminator. The key advantages of pSp include less training instability due to no discriminator, better style vectors injected into StyleGAN generator due to their encoder and intermediate latent space, along with interesting properties of StyleGAN such as style mixing and multimodal synthesis. However, due to its global approach it is not able to represent finer details in its generated images (Figure 1). While the style gets propagated by its *map2style* network, the outputs lack spatial layout consistency that is provided in the input, and is often crucial in I2I tasks. As a result, pSp struggles on datasets with more complex details and higher inter-input image variance. Following are our key contributions in this work:

1. We propose a simple solution termed **Spatial Layput Injection (SLI)** that encodes spatial layout information present in the encoder, and propagates it to the StyleGAN decoder. Since our aim is to encode style and structure, we do not replace the *map2style* network to have a spatial bottleneck (spatial size $> 1 \times 1$), but rather combine the multi-scale encoder feature maps with noise layers in the StyleGAN generator.

2. We demonstrate that our approach does not replace global consistency with local, but rather enhances spatial layout details whilst preserving global aspects of image generation. (see Figure 2).

3. With our proposed approach, we achieve state-of-the-art results on face data like CelebA and CelebAMask-HQ across a wide variety of I2I tasks. We also demonstrate the effectiveness of *SLI-pSp* on a more complex, custom dataset of building images.

4. Finally, we propose that SLI can be seen as a concept rather than a particular architecture design. We show that it is amenable to modifications by introducing Attentive SLI (ASLI-pSp) to achieve even better results. Such modifications can be made as seen fit depending on memory and cost requirements.

1. Related Work

1.1. GANs

Generative Adversarial Networks (GANs) [9] work on the principle of a min-max game between a generator model and discriminator model each trying to defeat the other. GANs have paved way for a variety of generative modelling tasks across different modalities such as RGB images, depth maps, segmentation masks, etc. When it comes to facial data, StyleGAN [17] is considered the state-of-the-art. This is due to its unique progressively growing structure which helps in generating high resolution images with great visual detail. Its mapping network helps disentangle factors of variation in the data and noise layers in the synthesis block help with stochastic variations in outputs. While StyleGAN gives impressive results, there still exist some characteristic artifacts which StyleGAN2 [18] attempts to alleviate. One of the key issues is the artifacts introduced in generated images due to AdaIN, therefore normalization is proposed on the expected statistics of the incoming feature maps. StyleGAN2-ADA [15] proposes an adaptive discriminator augmentation scheme which stabilizes and makes StyleGAN2 training work well in low data regimes. Further, Karras et al. [16] demonstrate improvement over StyleGAN2 by interpreting all signals in the network as continuous and deriving small architectural changes to further improve FID of StyleGAN2. They aim at making the synthesis network equivariant w.r.t the continuous signal in order to ensure that the finer details and coarser details transform together using Fourier features.

Bartz et al. [4] use the noise inputs of StyleGAN model to transfer content and color information in a fixed StyleGAN model. They train an encoder to reconstruct and perform denoising without re-training StyleGAN. However, instead of using stochastic noise they force the model to rely only on latent code and only train the layers responsible for predicting stochastic noise inputs. Abdal et al. [1] propose a novel algorithm to embed a given image in the latent space of StyleGAN. They show results for different tasks such as morphing, style transfer and expression transfer and also provide insights into the latent space of StyleGAN. Park et al. [27] propose Swapping Autoencoder model for performing image manipulation. Their method encodes the image into two independent components, structure and texture and then combines them to form a realistic image. The swapping autoencoder consists of an encoder E and a generator G that form a mapping between the latent image and code, dividing the latent code into structure and texture and enforcing swapping with other images. They also use a patch discriminator which learns the co-occurrence statistics of

image patches. Their model is based on Im2StyleGAN [1].

The ability to condition GANs on priors has led to exponential growth of work in conditional image synthesis, and more generally, image to image translation methods discussed further in the next section.

1.2. Conditional GANs & Image to Image Translation

Conditional GANs enable for a more controlled image generation by conditioning the generator on some prior. Previous works have demonstrated the ability of conditional GANs across various modalities such as 3DGAN [38], SegAN [40], MedGAN [3] for medical image synthesis, text-to-image synthesis such as AttnGAN [39], StackGAN [43], StackGAN++ [44], ACGAN [24], to handwritten font generation with GlyphGAN [10].

Image to Image (I2I) translation methods aim to learn the mapping of an image from a source to target domain. Isola et al. [13] first proposed the I2I method for conditioning the network on image and further introducing L_1 loss in the objective function. They demonstrated results on multiple tasks such as image $\rightarrow$ label, edges $\rightarrow$ image, day $\rightarrow$ night images, map $\rightarrow$ aerial. The works of Wang et al. [37] extend this approach for high resolution image generation and manipulation. Coming specifically to the task of semantic image synthesis, usually the semantic mask is provided to the generator via an encoder mapping [13, 37, 33]. However, it doesn't lead to great semantic coherency in the final image. Therefore, SPADE [26] method uses spatially adaptive normalization layers to modulate activations of the generator network. In CC-FPSE [21], the authors propose using a Feature Pyramid based Discriminator to work at multiple scales unlike a conventional PatchGAN discriminator. In OASIS [32], the authors propose using a semantic segmentation based discriminator to get better mIoU scores for the generated images. They also get rid of a traditional encoder and instead enable multi-modal synthesis directly by sampling a 3D noise tensor at each layer of the model.

pSp [29], on the other hand, proposes to solve various I2I tasks with a common architecture unlike previous mentioned methods. It uses an intermediate style space to embed input images and uses pretraining of a StyleGAN model as a proxy for the discriminator. It also uses a Feature Pyramid based encoder to better extract style vectors for multiple scales. We identify the shortcomings of pSp's global approach to I2I tasks, and propose Spatial Layout Injection (SLI) into the generator model to better preserve local structure information present in the input image. Such information is crucial for tasks such as semantic image synthesis, inpainting, super resolution, etc. We do so without compromising the global approach of pSp that is provided by the style vectors being injected in the generator.

Alaluf et al. [2] propose a method for iterative refinement over the baseline pSp architecture. At every time step they provide the network with current input and the output obtained at the previous time step while the inital output is initialized using the latent average.

1.3. Self-Attention Methods

Transformer architectures were initially proposed for machine translation tasks [35] in NLP, but these models have made their headway into computer vision in recent years. ViT [8] made a breakthrough with its superior performance on large scale dataset like ImageNet. Other key architectures include DETR [5] for object detection, VilBERT [22] for vision and language tasks and CCNet [12] for semantic segmentation. BoTNet [31] proposes a hybrid model which uses both convolutions and self-attention. This is different from other models like DETR and CCNet as they use self attention outside their backbone unlike BoTNet. We use BoTNet as a modification to the simple SLI-pSp model, and denote it by ASLI-pSp. Note that this modification is merely to portray how SLI can be seen as a concept and upgraded in its implementation depending on memory and cost requirements. We do not perform quantitative studies across attention models to compare performance of our technique.

2. Approach

In this section, we present SLI-pSp, a model that introduces a spatial layout injection from the encoder to StyleGAN noise layers, thereby enabling better performance in image-to-image translation tasks. We use pSp framework as the base architecture and utilize multi-scale features from the encoder to produce better local consistency in the generated images while retaining global properties. We empirically show that such feature injections help preserve semantic details better, while also enjoying the interesting properties like multi-modal synthesis that StyleGAN has to offer.

2.1. Baseline Method

We first briefly revisit the key components of StyleGAN and pSp framework that form the backbone for our method.

StyleGAN [17] proposes a novel style-based generator architecture for unconditional image generation. A key idea is to transform the latent vector $\mathbf{z} \in \mathcal{Z}$ via a mapping network $\mathbf{f}$ into a vector $\mathbf{w}$ in the intermediate latent space $\mathcal{W}$. This aids the representation to be more disentangled with respect to the factors of variation in the network. $\mathbf{w}$ is then propagated via learned affine transformations to the synthesis network as styles $\mathbf{y}$ that control the Adaptive Instance Normalization (AdaIN) after convolution operations. Finally, to encourage stochastic variation in generated images, explicit single channel Gaussian Noise is fed to each layer of the synthesis network.

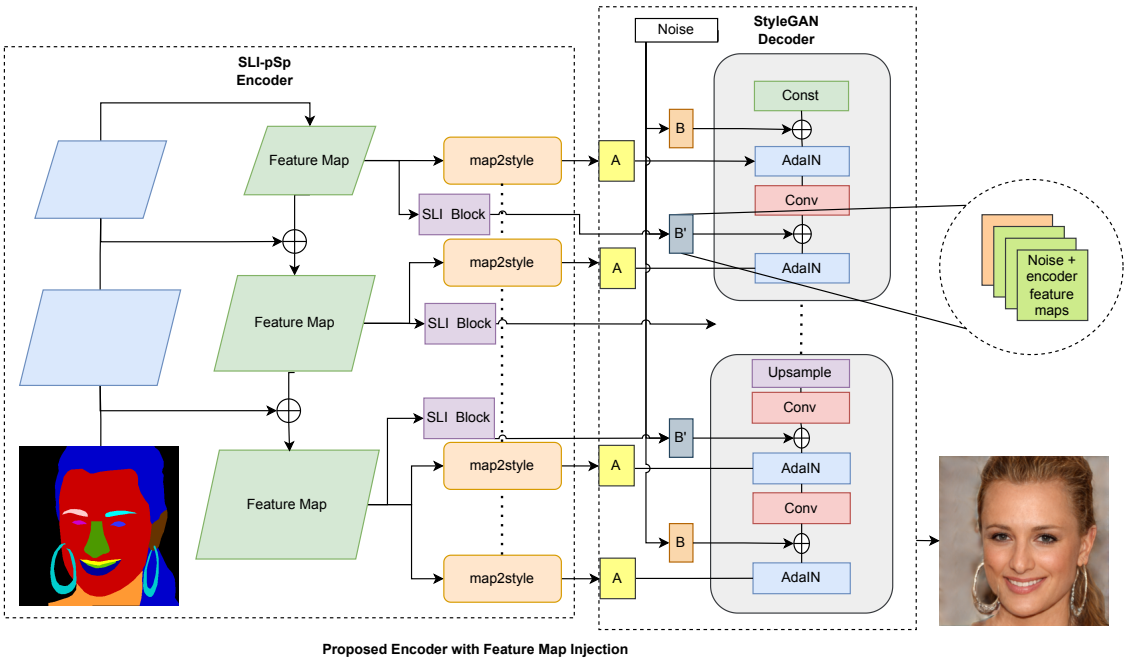

Figure 2. **SLI-pSp** framework. We extract feature maps using a feature pyramid network. These features are propagated via map2style blocks and injected into the StyleGAN generator. While this serves the global aspect of image generation, we also propagate these encoded features via Spatial Layout Injection (SLI) blocks to combine with the Noise Layers of StyleGAN. This helps in providing spatial context to the image synthesis process, as is required for various I2I tasks.

pSp [29] proposes a Feature Pyramid Network based encoder framework which embeds input images into an extended latent space $\mathcal{W}+$. The intermediate style representation offers many advantages such as the ability to resample style vectors thereby providing support for multimodal synthesis. They also propose a new way to use a pretrained StyleGAN generator model for generic image-to-image translation tasks. The style vectors are passed to the StyleGAN generator corresponding to their scales. A key advantage of this approach is that it does not require training a discriminator network.

2.2. Spatial Layout Injection

One the drawbacks of pSp's global approach is that it seems to miss finer or underrepresented details in the face images, such as earrings or caps. On probing further, this approach broke down when tried on other datasets with more variation and complexity such as Places2 [47]. This is due to a lot of high frequency information and numerous objects in a scene that are insufficient to be captured without transfer of input spatial layout information. pSp's *map2style* network downsamples the spatial resolution to 1×1, thereby hindering propagation of input layout information from the encoder.

To remedy this, we propose Spatial Layout Injection (SLI) that encodes spatial layout information and propagates it to the pretrained StyleGAN generator. One way to inject such information could be by modifying the $map2style$ network to create a spatial bottleneck and propagating style and encoded spatial layout to the generator. However, there is merit in having a global approach as is required by certain I2I tasks. Therefore, we strategically embed the feature maps from the encoder by combining it with the noise layers in the generator without changing the style injection via *map2style* in the network. Since the authors of StyleGAN state that the noise injected separately allows for local stochastic changes this makes it a suitable location for the injection of the features. Further since StyleGAN uses hierarchical generation and synthesises finer details at higher levels we restrict feature injection to lower scales to ensure higher flexibility. Let the encoder feature maps be denoted as $\hat{E}_i$, 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

$$B' = concat(conv(\hat{E}_i), N_i) \quad i \in \{16, 32, 64\}, \qquad (1)$$

were $conv$ represents a convolution layer which reduces number of channels from 512 to 256 and $concat$ operation concatenates along the channel dimension as spatial sizes are the same. Instead of using a single channel noise which is then broadcasted and added to the output of correspond-

ing convolution, we use a 256-channel noise which is then combined with a 256-channel *SLI*.

This ensures that style propagation via the $map2style$ network remains unhindered, whilst scene layout information is also injected into the generator. These feature maps are multi-scale so it is possible to choose the scale and number of feature maps to inject in the generator. An added advantage of SLI is that it is input modality agnostic, i.e the input could be a segmentation map, edge map, blurry image etc. without any change required in the architecture. Moreover, this is done without compromising the StyleGAN properties, e.g multi-modal synthesis via style-mixing as shown in supplementary.

2.3. Variant - Attentive Spatial Layout Injection

To show the versatility of the SLI concept, we employ a variant of SLI with attention, namely Attentive SLI (ASLI-pSp). We use Bottleneck Transformers as introduced in [31] for our attention layer. This modification can be understood via the following equation

$$B' = concat(botnet(conv(\hat{E}_i)), N_i) \quad i \in \{16, 32, 64\},$$

where $botnet(.)$ is the Bottleneck Transformer operation that takes in the output of the encoder feature maps after they are passed through the $conv$ layer. The botnet-transformed feature maps are then similarly concatenated with the noise layers in the StyleGAN generator. We will see in later sections how such a modification can boost scores like FID.

2.4. Loss Functions

We use similar loss functions as done in pSp as it serves for a fair comparison with our proposed approach. We also employ a weighted combination of losses that are listed below. The pixel-wise reconstruction loss (or $\mathcal{L}_2$ $loss$) is defined as

$$\mathcal{L}_2(x) = ||x - SLIpSp(x)||_2$$

We also use an LPIPS [45] loss which helps preserve image quality. Here, F(·) denotes the perceptual feature extractor.

$$\mathcal{L}_{LPIPS}(x) = ||F(x) - F(SLIpSp(x))||_2$$

The latent vector regularisation loss helps the encoder *E* to generate style vectors in the latent space closer to average latent vector $\overline{w}$.

$$\mathcal{L}_{reg} = ||E(x) - \overline{w}||_2$$

For facial images, preserving identity is crucial. Hence, an ID Loss is also employed when dealing with facial images. The loss measures the cosine similarity between input and output image.

$$\mathcal{L}_{ID}(x) = 1 - \langle R(x), R(SLIpSp(x)) \rangle,$$

where R(.) is a pretrained ArcFace Network [7]. Thus, the final loss equation can be expressed as

$$\mathcal{L}(x) = \lambda_{L_2}\mathcal{L}_2(x) + \lambda_{LPIPS}\mathcal{L}_{LPIPS}(x) + \lambda_{reg}\mathcal{L}_{reg}(x) + \lambda_{ID}\mathcal{L}_{ID}(x)$$

3. Experimental Setup

We demonstrate our approach across a wide variety of tasks and datasets to illustrate the benefits of combining style and spatial layout information in an I2I setting. Specifically, we pick the following tasks: Segmentation Map to Face (Seg2Face), Super Resolution and Edges to RGB Image. We use the default training settings as in the original method proposed in [29] and train for a max 500000 steps on NVIDIA Tesla V100 32GB GPU and a batch size of 8.

3.1. Datasets

- **CelebA-HQ** [14]: It contains about 30,000 high resolution face images from the CelebA dataset. The train set consists of about 24,000 images. It is used in our Super Resolution task.
- **CelebAMask-HQ** [20]: It is a derivative of the CelebA-HQ dataset with same train-test splits, but comes with segmentation masks. Each mask of an image is manually annotated, and the dataset contains 19 classes such as skin, nose, eyes, lip, hair, etc. We use this for Seg2Face Task.
- **AFHQ-Dog** [6]: This dataset is a collection of around 5000 high quality dog images (faces), of which about 500 images form the test set. We showcase Super Resolution results on this dataset.
- **Places2-CustomBuildings**: We extract a few building categories from the Places2 dataset [47] for pretraining of the StyleGAN generator. These include building facade, courthouse, manufactured home, office building, parking garage-outdoor and residential neighborhood and total to about 150,000 images. We select a subset of around 30,000 images from these as train set (and $\sim$ 3700 as test set) for evaluating the baseline pSp, and our methods SLI-pSp and ASLI-pSp. A custom dataset is chosen to showcase the effectiveness of SLI in the generated images. We use this dataset for Edges to RGB image task.

3.2. Training Details

Similar to the pSp framework, we train our network using the ResNet-IR architecture. Further, for each of the datasets mentioned above we train the StyleGAN separately and the use it as the decoder with the modified psp-Encoder for further training in a conditioned setting. The pSp-Encoder is trained from scratch while the StyleGAN is further fine-tuned. Input image size is 256×256

for all tasks across different datasets. The learning rate for the experiment is 0.0001 with Ranger optimizer. The loss coefficients used were the same as in pSp training, i.e $\lambda_{LPIPS} = 0.8, \lambda_{L_2} = 1.0, \lambda_{reg} = 0.005$. $\mathcal{L}_{ID}$ is not used for any task except CelebA-Super Resolution, which uses $\lambda_{ID} = 1.0$. The training times of both StyleGAN and pSp are the similar to the original works and we use the original configurations for the same.

4. Results and Discussion

To evaluate the effectiveness of SLI, we compare our approach to pSp and a few other baselines across a variety of I2I tasks. We describe the experiment, its results and quantitative evaluations to assess image diversity, quality and correctness. We then discuss the key differences of our approach with respect to the results.

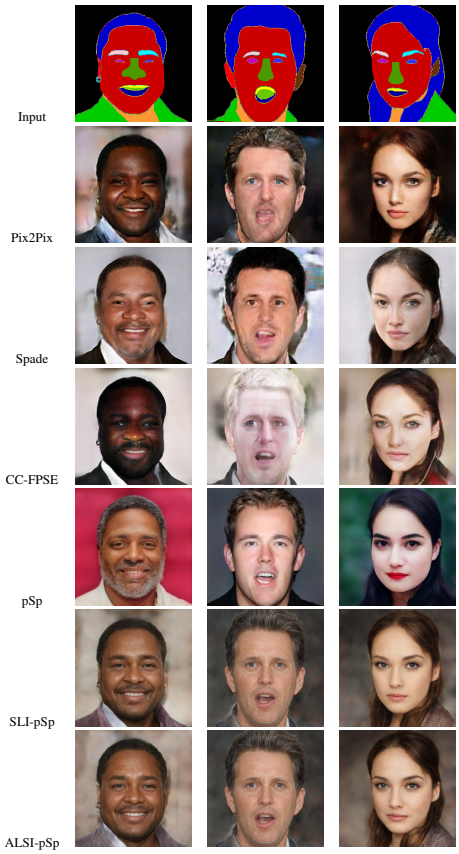

Figure 3. We show results on the Seg2Face task using the CelebAMask-HQ dataset and compare it to some previous works.

4.1. Segmentation Map to Face Image

Method: The generated face images are conditioned on corresponding segmentation maps provided during training and inference. During evaluation, we do style-mixing at higher level features, i.e we combine latent code of input image (segmentation mask) with a randomly sampled latent vector. We evaluate our method against various baselines.

We evaluate *SLI-pSp* and *ASLI-pSp* on CelebAMask-HQ dataset by synthesizing face images from segmentation maps. One of the biggest differences in the generated outputs is that SLI variants can synthesize smaller and underrepresented objects in the dataset, such as earrings while maintaining image quality as shown in Figure 1. In addition to that, we observe a tighter correspondence of the generated images with the label map. This is especially evident in the case of varying hair styles (See Figure 1 (b), (c) and Figure 3). We quantify this tighter correspondence by computing mean IoU scores across classes using the generated images for each method.

Table 1. Quantitative evaluation on the test set.

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**

Additionally, as it is shown in Figure 1(c), pSp confuses the cap class with hair, and generates a photorealistic but semantically incorrect image. In this case, both SLI variants are able to capture the correct semantic information. Note not only does our method generate the correct cap structure while displaying the right texture, it also displays realistic lighting along the contours where the shadow of the cap is visible.

To evaluate quantitatively, we employ various metrics such as LPIPS, FID and mIoU. FID [11] is used as a proxy for realism, and is sensitive to both quality and diversity. The SLI-variants perform significantly better than pSp with a difference of greater than 16 points. The mean IoU scores are computed by evaluating the label map synthesised by our generated images against ground truth label maps using a pre-trained segmentation network. We outperform pSp in this regard as well, and it strengthens our qualitative analysis about SLI variants being better with respect to under represented objects in the data.

4.2. Edges to RGB Image

We explore our method on another task which takes an edge image as input and generates an RGB image. After evaluating on face data which StyleGAN can handle relatively well, we wanted to stress test our method on a complex dataset. Therefore, we compiled building images from various categories in the Places2 dataset, termed as *Places2-CustomBuildings*. The variance among images is high as these buildings have been captured in varying lighting conditions, different poses and comprise of different types of structures like church, skyscrapers, residential homes, etc. We then create edge images for the dataset using Canny edges and some post-processing to decrease noisy edges.

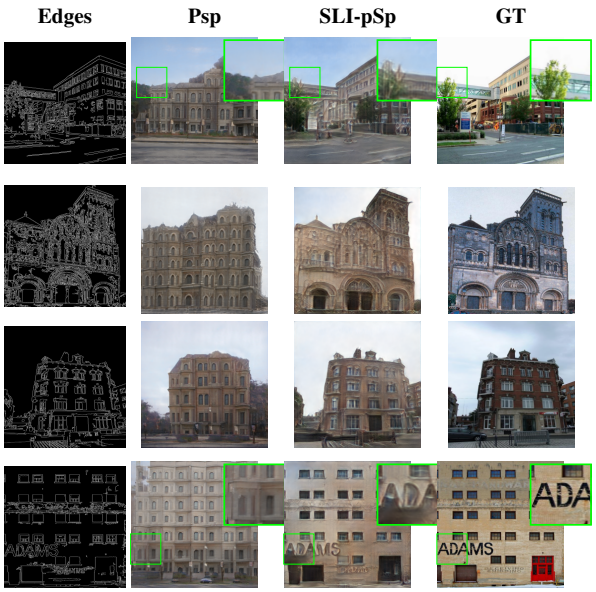

Figure 4. Results on Places2 dataset. Our method generates much reliable information due to feature injection.

Table 2. Quantitative Results for Edges to RGB image task on Places2-CustomBuildings dataset.

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

We observe that pSp doesn't perform well on this dataset at all and produces generic facade details which seem to be common across different input edge images, although it does very loosely follow the outline of the building and the sky. We attribute this to two primary factors, a) the lack of spatial information in the image synthesis process, and b) lack of semantic context in the edges. This can be verified by the very high FID score achieved by pSp on this task. The same set of edges can be used for representing a variety of different scene components unlike face data, where one can visually distinguish between the edges that create different facial structures such as nose, ears or mouth, thereby making the problem highly ill-posed. On the other hand, we notice that *SLI-pSp* is able to capture spatial layout very well and is able to map the highly concentrated edges to semantically consistent building components which make sense globally (image-level). We see in Figure 4(c), how our method is able to capture the individual window designs, and even engravings in the buildings (Figure 4(d)) unlike pSp. We also outperform pSp by quite a margin quantitatively (Table. 2). While the generated image by our method lacks the visual quality that we got on facial images (primarily due to complexity of task and data), we put forward a strong case for SLI as an idea to be explored even further, along with the idea of using style vectors in the generation process, as done in pSp.

4.3. Super-Resolution

Method: We follow the pSp setup to synthesize high resolution images from their low-resolution counterparts. We downsample the high resolution images at various scales (x2, x4, x8, x16, x32) using bi-cubic interpolation and use it as the input image.

We showcase our method's performance on the Super-resolution task. We choose 2 datasets: CelebA and AFHQ-Dog on which we evaluate our method's performance.

In the case of CelebA, we notice that our method visually performs at par with pSp, but is better at preserving spatial information present in the downsampled input. SLI variants are especially better at preserving the hairstyles and color scheme of the clothes present in the low resolution image (See Figure 5 (top)). On observing Figure 5(bottom), we notice how well SLI variants can capture objects like caps that are omitted by pSp and misrepresented as hair. Similarly, we also notice the hairstyle is kept consistent in the upsampled images by SLI variants as opposed to pSp. The above claims are well supported quantitatively (Table 3) by substantial improvements in FID scores. LPIPS and L_2 losses also show a reduction on the test set, thereby signalling better performance using SLI. We do notice a slight reduction in sharpness of the image generated by our method that can be attributed to the introduction of local information injected by SLI, and is definitely something to improve as future work.

Table 3. 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.

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**

We also evaluate the task of super-resolution on the AFHQ-Dog dataset, which is different compared to the human faces dataset on which StyleGAN works quite well. We observe how pSp misses out on upsampling colors correctly, and that *SLI-pSp* does a better job at generating more accurate colors and lighting in the scene. We also notice in Figure 6 (1st and 3rd row), how SLI variants are able to preserve the yellow part of the green grass. This can be argued to be representative of style of the input image (and not merely structural) which our method is able to capture. We do notice a little washing out of colors in our ASLI variant that would be investigated in future work.

Overall, for tasks like super-resolution, we argue for approaches which encode spatial features accurately in the generation process as it is imperative to have an accurate structure in the generated "upsampled" image. This is achieved to a great extent by our *SLI-pSp* method.

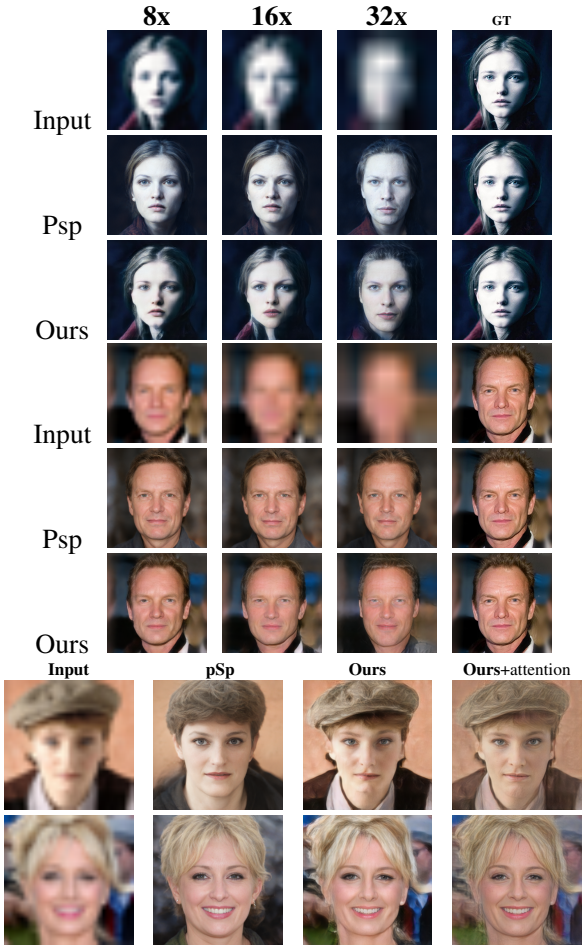

Figure 5. Superresolution Results. We observe a mild tradeoff between preservation of semantics and overall image quality in both methods due to introduction of locality bias

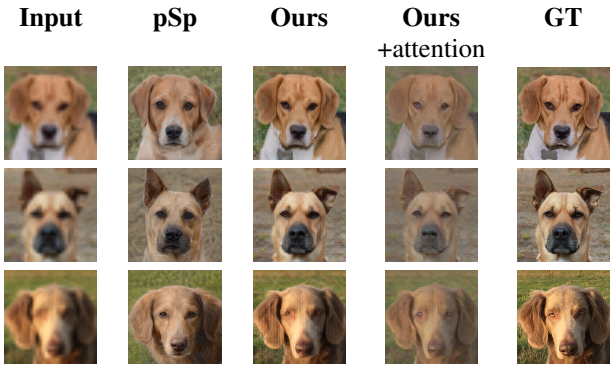

Figure 6. Results on AFHQ super-resolution. First row shows results for incorrect color, second and third row show results for incorrect structure and colors.

5. Conclusion, Limitations and Future Scope

In this work, we have identified the shortcomings of pSp, a generic Image-to-Image translation framework. We discover that due to the global approach of pSp, and the lack of spatial layout context in the image generation process, it is not able to synthesize finer and less represented details in the dataset, and fails to work in more complex data settings. We propose a simple fix termed Spatial Layout Injection (SLI) that encodes the spatial layout information from the input image and propagates it to the StyleGAN decoder. While there is advantage in having a global approach for certain problem settings, we argue that it is not sufficient, and that we need both style and spatial information in the image synthesis process. Hence, *SLI-pSp* combines the multi-scale encoder feature maps with the noise layers in the StyleGAN generator, without modifying the style propagation component. We demonstrate through our detailed experiments that we don't replace global context with local, but rather amplify the spatial layout details required for many I2I tasks. We showcase results and comparisons with pSp on various tasks such as segmentation map to face image, edges to RGB image, and super resolution. We evaluate these tasks across various datasets and show qualitatively and quantitatively, the importance of SLI as a concept in the I2I process. While SLI is done by a simple *conv* operation, we show how it can be upgraded to a different design choice by introducing *ASLI-pSp*, depending on computational budget.

While we achieve good performance across tasks, there is a sharpness reduction issue and artifacts are observed in some samples in the case of super resolution using attention mechanism despite a better performance than pSp in terms of the FID scores. This might require probing various configuration changes to the SLI Block or even modifying the encoder for better features to inject in the generator. We hypothesise that the aritifacts observed in case of super-resolution are due to strong locality bias without tight correspondence to the output. While we achieved substantially better results in the edges $\rightarrow$ buildings task where the edges have a tighter correspondence to the output, notwithstanding its ill-posedness and complexity, there is much room for improvement as the generated images lacked the finesse present in the facial images. Since our key focus is on improving synthesis quality we leave inversion for future research. This could be a potential future direction where the generator architecture might need to be modified or maybe a discriminator would aid in better visual quality. We hope that the simple yet important concept that we explored would encourage the vision and graphics community to further conduct research in this direction.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021.

[3] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized Medical Imaging and Graphics*, 79:101684, 2020.

[4] Christian Bartz, Joseph Bethge, Haojin Yang, and Christoph Meinel. One model to reconstruct them all: A novel way to use the stochastic noise in stylegan. *arXiv preprint arXiv:2010.11113*, 2020.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[10] Hideaki Hayashi, Kohtaro Abe, and Seiichi Uchida. Glyphgan: Style-consistent font generation based on generative adversarial networks. *Knowledge-Based Systems*, 186:104927, 2019.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[15] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

[16] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[19] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017.

[20] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *arXiv preprint arXiv:1910.06809*, 2019.

[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[23] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[24] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[25] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *arXiv preprint arXiv:2101.08629*, 2021.

[26] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[27] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *CoRR*, abs/2007.00653, 2020.

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[29] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[30] Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C-C Jay Kuo. Contextual-based image inpainting: Infer, match, and translate. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[31] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16519–16529, 2021.

[32] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020.

[33] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7870–7879, 2020.

[34] Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Art2real: Unfolding the reality of artworks via semantically-aware image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5849–5859, 2019.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[36] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[38] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *CoRR*, abs/1610.07584, 2016.

[39] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[40] Yuan Xue, Tao Xu, Han Zhang, L. Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l_1 loss for medical image segmentation. *CoRR*, abs/1706.01805, 2017.

[41] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018.

[42] Dan Zhang and Anna Khoreva. Progressive augmentation of gans. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6249–6259, 2019.

[43] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[44] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[46] Yongbing Zhang, Siyuan Liu, Chao Dong, Xinfeng Zhang, and Yuan Yuan. Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution. *IEEE transactions on Image Processing*, 29:1101–1112, 2019.

[47] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[48] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017.

[49] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.