A. Additional Implementation Details

Generator. Our generator consists of several SEAN ResBlks. Each of them is followed by a nearest neighbor up-sampling layer. Note that we only inject the style codes $ST$ into the first 6 SEAN ResBlks. The other inputs are injected to all SEAN ResBlks. The architecture of our generator is shown in Figure 1.

![Figure 1: SEAN Generator.](image)

Discriminator. Following SPADE [9] and Pix2PixHD [12], we employed two multi-scale discriminators with instance normalization (IN) [11] and Leaky ReLU (LReLU). Similar to SPADE, we apply spectral normalization [8] to all the convolutional layers of the discriminator. The architecture of our discriminator is shown in Figure 2.

![Figure 2: Following SPADE and Pix2PixHD, our discrimin-](image)

Style Encoder. Our style encoder consists of a “bottle-neck” convolutional neural network and a region-wise average pooling layer (Figure 4). The inputs are the style image and the corresponding segmentation mask, while the outputs are the style codes $ST$.

Loss function. The design of our loss function is inspired by those of SPADE and Pix2PixHD which contains three components:

1) Adversarial loss. Let $E$ be the style encoder, $G$ be the SEAN generator, $D_1$ and $D_2$ be two discriminators at different scales [12], $R$ be a given style image, $M$ be the corresponding segmentation mask of $R$, we formulate the conditional adversarial learning part of our loss function as:

$$\min_{E,G} \max_{D_1,D_2} \sum_{k=1,2} \mathcal{L}_{GAN}(E, G, D_k)$$
Specifically, $\mathcal{L}_{GAN}$ is built with the Hinge loss that:

$$
\mathcal{L}_{GAN} = \mathbb{E} \left[ \max(0, 1 - D_k(R, M)) \right] + \mathbb{E} \left[ \max(0, 1 + D_k(G(ST, M), M)) \right]
$$

(2) Feature matching loss [12]. Let $T$ be the total number of layers in discriminator $D_k$, $D_k^{(i)}$ and $N_i$ be the output feature maps and the number of elements of the $i$-th layer of $D_k$ respectively, we denote the feature matching loss term $\mathcal{L}_{FM}$ as:

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

(4)

(3) Perceptual loss [4]. Let $N$ be the total number of layers used to calculate the perceptual loss, $F^{(i)}$ be the output feature maps of the $i$-th layer of the VGG network [10], $M_i$ be the number of elements of $F^{(i)}$, we denote the perceptual loss $\mathcal{L}_{percept}$ as:

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

(5)

The final loss function used in our experiment is made up of the above-mentioned three loss terms as:

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

(6)

Following SPADE and Pix2PixHD, we set $\lambda_1 = \lambda_2 = 10$ in our experiments.

**Training details.** We perform 50 epochs of training on all the datasets mentioned in the main paper. During training, all input images are resized to a resolution of $256 \times 256$, except for the CityScapes dataset [1] whose images are resized to $512 \times 256$. We use Glorot initialization [2] to initialize our network weights.

**B. Additional Experimental Details**

**Table 3 (main paper).** Supplementing row 5 and 6 in Table 3 of the main paper. Figure 3 shows how the two variants of style encoders (i.e., the SEAN-level encoder and the ResBlk-level encoder) are used in a SEAN ResBlk. Specifically, the SEAN-level encoders extract different style codes for each SEAN block while the same style codes extracted by the ResBlk-level encoder are used by all SEAN blocks within a SEAN ResBlk.

**Figure 6 (main paper).** We used the Ground Truth (second column in Figure 6 of the main paper) as the style input for all methods. For Pix2pixHD, we generate the results by: (i) encoding the style image into a style vector; (ii) broadcasting the style vector and concatenating it to the mask input of the generator.
C. Justification of Encoder Choice

Figure 5 shows that the images generated by the unified encoder are of better visual quality than those generated by the SEAN-level encoder, especially for challenging inputs (e.g., extreme poses, unlabeled regions), which justifies our choice of unified encoder.

D. Additional Analysis

ST-branch vs. Mask-branch. The contributions of ST-branch and mask-branch are determined by a linear combination (parameters $\alpha_\beta$ and $\alpha_\gamma$). The resulting parameters are typically in the range of $0.35 - 0.7$ meaning that both branches are actively contributing to the result. See Fig 6 for one example. It is possible to completely drop the mask-branch, but the results will get worse. It was our initial intuition that the mask branch provides the rough structure and the ST-branch additional details. However, in the end, the interaction is quite complicated and cannot be understood by just varying the mixing parameter.

Extreme Cases. To further demonstrate SEAN’s power in texture transfer, we show that highly complex textures from an artistic image can be transferred to a human face (Fig 7). In addition, our method is highly flexible that enables users to paint a semantic region at a spatially unreasonable location arbitrarily (Fig 8).

E. User Study

We conducted a user preference study with Amazon Mechanical Turk (AMT) to illustrate our superior reconstruction results against existing methods (Table 1). Specifically, we created 600 questions for AMT workers to answer. In the end, our questions are answered by 575 AMT workers. For each question, we show the user a set of 5 images: a ground truth image, its corresponding segmentation mask, and 3 reconstruction images obtained by our method, Pix2PixHD [12] and SPADE [9]. The user is then asked to select the reconstructed image closest to the ground truth and with fewest artifacts. To relieve the impact of image orders and make a fair comparison, we picked 100 image sets randomly and created the 600 questions by enumerating all the 6 possible orders of the 3 reconstructed images in each of them.

F. Additional Results

To demonstrate that the proposed per-region style control method builds the foundation of a highly flexible image-editing software, we designed an interactive UI for a demo. Our UI enables high quality image synthesis by transferring the per-region styles from various images to an arbi-

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Preference (%)</td>
<td>23.17</td>
<td>8.83</td>
<td>68.00</td>
</tr>
</tbody>
</table>


arbitrary segmentation mask. New styles can be created by interpolating existing styles. Please find the recorded videos of our demo in the supplementary material.

Figure 9 shows additional style transfer results on CelebAMask-HQ [6, 5, 7] dataset. Figure 10 and Figure 11 show additional style interpolation results on CelebAMask-HQ and ADE20K datasets.

Figure 12, 13, 14 and 15 show additional image reconstruction results of our method, Pix2PixHD and SPADE on the CelebAMask-HQ [6, 5, 7], ADE20K [13], CityScapes [1] and our Façades datasets respectively. It can be observed that our reconstructions are of much higher quality than those of Pix2PixHD and SPADE.

References

Figure 9: Style transfer on CelebAMask-HQ dataset
Figure 10: Style interpolation on CelebAMask-HQ dataset

Figure 11: Style interpolation on ADE20K dataset
Figure 12: Visual comparison of semantic image synthesis results on the CelebAMask-HQ dataset. We compare Pix2PixHD, SPADE, and our method.
Figure 13: Visual comparison of semantic image synthesis results on the ADE20K dataset. We compare Pix2PixHD, SPADE, and our method.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Label" /></td>
<td><img src="image2.png" alt="Ground Truth" /></td>
<td><img src="image3.png" alt="Pix2PixHD" /></td>
<td><img src="image4.png" alt="SPADE" /></td>
<td><img src="image5.png" alt="Ours" /></td>
</tr>
<tr>
<td><img src="image6.png" alt="Label" /></td>
<td><img src="image7.png" alt="Ground Truth" /></td>
<td><img src="image8.png" alt="Pix2PixHD" /></td>
<td><img src="image9.png" alt="SPADE" /></td>
<td><img src="image10.png" alt="Ours" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Label" /></td>
<td><img src="image12.png" alt="Ground Truth" /></td>
<td><img src="image13.png" alt="Pix2PixHD" /></td>
<td><img src="image14.png" alt="SPADE" /></td>
<td><img src="image15.png" alt="Ours" /></td>
</tr>
<tr>
<td><img src="image16.png" alt="Label" /></td>
<td><img src="image17.png" alt="Ground Truth" /></td>
<td><img src="image18.png" alt="Pix2PixHD" /></td>
<td><img src="image19.png" alt="SPADE" /></td>
<td><img src="image20.png" alt="Ours" /></td>
</tr>
<tr>
<td><img src="image21.png" alt="Label" /></td>
<td><img src="image22.png" alt="Ground Truth" /></td>
<td><img src="image23.png" alt="Pix2PixHD" /></td>
<td><img src="image24.png" alt="SPADE" /></td>
<td><img src="image25.png" alt="Ours" /></td>
</tr>
<tr>
<td><img src="image26.png" alt="Label" /></td>
<td><img src="image27.png" alt="Ground Truth" /></td>
<td><img src="image28.png" alt="Pix2PixHD" /></td>
<td><img src="image29.png" alt="SPADE" /></td>
<td><img src="image30.png" alt="Ours" /></td>
</tr>
</tbody>
</table>

Figure 14: Visual comparison of semantic image synthesis results on the ADE20K dataset. We compare Pix2PixHD, SPADE, and our method.
Figure 15: Visual comparison of semantic image synthesis results on the CityScapes and Façades dataset. We compare Pix2PixHD, SPADE, and our method.