A. Additional Implementation Details

A.1. Retrieval-based Guidance Image

Given a semantic map $M$, we use it to retrieve and composite a guidance image $I^r$ for image synthesis.

Preprocess of Dataset. The training dataset $D^{tr}$ is firstly used to create a retrieval database consisting of a set of segments. Specifically, for each image $I \in D^{tr}$ and its corresponding semantic map $M$, we use the available instance-level annotation to decompose $I$ and $M$ as a number of segments,

$$ I, M = \{(M_i^s, y_i^c, I_i^r)\},$$

where $M_i^s, y_i^c$ and $I_i^r$ denote the cropped binary mask of the $i$-th object, its category and its corresponding RGB segment image, respectively. Besides, for background region without instance-level annotation, we take the maximal connected component as a single background object. Decomposing all the images in training dataset, we create a retrieval database, which is used in both training and testing stage.

Retrieval Strategy. Given a semantic map $M$, we first decompose it into a number of segment masks $\{(M_i^s, y_i^c)\}$. Then, we retrieve the most compatible segment from the retrieval database for each segment mask. Specifically, for segment mask $M_i^s$ with category $y_i^c$, we retrieve a segment $(\hat{M}_j^s, y_j^c, I_j^r)$ which has the same category ($y_j^c = y_i^c$) and similar shape with $M_i^s$. To measure the similarity between two segment masks ($M_i^s$ and $M_j^s$), we adopt the geometric score [8] to measure both scale and shape consistency,

$$\sigma_{scale}(M_i^s, M_j^s) = \begin{cases} 0, & t \geq 0.5 \\ 1, & t < 0.5 \end{cases},$$

$$\sigma_{shape}(M_i^s, M_j^s) = \frac{SSD(\hat{M}_i^s, \hat{M}_j^s)}{\max(\|\hat{M}_i^s\|_1, \|\hat{M}_j^s\|_1)},$$

where $t = \frac{\min(\|M_i^s\|_1, \|\hat{M}_i^s\|_1)}{\max(\|M_i^s\|_1, \|\hat{M}_i^s\|_1)}$. $\hat{M}_i^s$ and $\hat{M}_j^s$ denote the resized versions (i.e., $128 \times 128$) of $M_i^s$ and $M_j^s$ using nearest neighbor interpolation, respectively. $SSD(\cdot)$ denotes the sum square difference. The final consistency is calculated as,

$$\sigma(M_i^s, M_j^s) = \sigma_{scale}(M_i^s, M_j^s) + \gamma \sigma_{shape}(M_i^s, M_j^s),(4)$$

where $\gamma$ is the balance coefficient and we set $\gamma = 1$ in practice. Lower $\sigma(M_i^s, M_j^s)$ indicates more similarity between two segment masks.

Composition of Guidance Image. Finally, we recompose the retrieved segments as the guidance image. Let $(\hat{M}_i^s, y_i^c, I_i^r)$ denotes the retrieved segment for the given segment mask $M_i^s$. As illustrated in Fig. A, $I_i^r$ and the corresponding mask $\hat{M}_i^s$ are first resized to the size of $M_i^s$. The resized mask and image are denoted as $M_i^r$ and $I_i^r$. Then, the resized image is pasted into the guidance image according to the original position of $M_i^r$. To maintain integrity of instance, we paste the segment image following the below rules:

- Pixels of $I_i^r$ in both $\hat{M}_i^s$ and $M_i^s$ are preserved.
- If $y_i^c$ belongs to background things categories, pixels of $I_i^r$ in $\hat{M}_i^s$ but not in $M_i^s$ are zeroed out.
- If $y_i^c$ belongs to foreground (i.e., instance object) and pixels of $I_i^r$ in $\hat{M}_i^s$ but not in $M_i^s$ are located in the background categories in $M$, they are preserved.
- If $y_i^c$ belongs to foreground and pixels of $I_i^r$ in $\hat{M}_i^s$ but not in $M_i^s$ are located in the foreground categories in $M$, they are zeroed out.

We finally obtain the retrieval-based guidance image $I^r$ to guide the image synthesis.


To distort the ground-truth image $I^{gt}$, we first decompose it into a set of segment images $I^{gt} = \{I_i^s\}$. Then we apply the distortion (i.e., color, shape and resolution) on each segment image $I_i^s$.

Color. We employ the method proposed by [6] to transfer the color of segment image $I_i^s$ to a random segment image $I_i^r$ with the same category. Specifically, we first convert $I_i^s$
and \( I^s \) from RGB space into \( l_{\alpha\beta} \) space. Then the color transferred image \( \hat{I}^s \) in each channel of \( l_{\alpha\beta} \) space is calculated by,

\[
\hat{I}_i = (l_i - \mu(l_i)) \cdot \frac{\sigma(l_i)}{\sigma(l_i)} + \mu(l_i),
\]

\[
\hat{\alpha}_i = (\alpha_i - \mu(\alpha_i)) \cdot \frac{\sigma(\alpha_i)}{\sigma(\alpha_i)} + \mu(\alpha_i),
\]

\[
\hat{\beta}_i = (\beta_i - \mu(\beta_i)) \cdot \frac{\sigma(\beta_i)}{\sigma(\beta_i)} + \mu(\beta_i),
\]

where \( \mu(\cdot) \) and \( \sigma(\cdot) \) denote the mean and standard deviation of corresponding channel. Finally, we convert \( \hat{I}^s \) from \( l_{\alpha\beta} \) into RGB space to obtain the color distorted image.

**Shape.** To distort the shape of a segment image, we first sample 10 points uniformly on the edge of the segment image as source points, and shift three of them randomly to produce the target points. The source points and target points are used to produce a dense flow utilizing thin plate spline algorithm. Then we use the produced flow to warp the segment image to obtain the shape distorted image.

**Resolution.** To distort the resolution of a segment image, we downsample it with a random scale \( \tau \) \((0.5 < \tau < 1)\), and upsample it to the original size.

After distortion, distorted segment images from ground-truth \( I^{gt} \) recompose the distorted ground-truth \( \hat{I}^{gt} \) to facilitate model training. The distortion results are shown in Fig. B.

### B. Additional Details of Training Architecture

#### Details of RESAIL module.** The RESAIL module takes both the guidance image (i.e., retrieval-based guidance \( I^r \) or distorted ground-truth \( \hat{I}^{gt} \)) and the semantic map \( M \) as input and learns to modulate the activations. We here represent the input activations as \( h \) with a batch of \( N \) samples. \( H, W \) and \( C \) denote the height, width and the number of channels in \( h \), and the modulated activations at site \( (n, c, y, x) \) is represented as,

\[
RESAIL(h, I^r, M) = \gamma_{c,y,x} (I^r, M) \frac{h_{n,c,y,x} - \mu_c}{\sigma_c} + \beta_{c,y,x} (I^r, M),
\]

where \( \mu_c \) and \( \sigma_c \) denote the mean and standard deviation of the activation in channel \( c \),

\[
\mu_c = \frac{1}{NHW} \sum_{n,y,x} h_{n,y,x},
\]

\[
\sigma_c = \sqrt{\frac{1}{NHW} \left( \sum_{n,y,x} h_{n,y,x}^2 - \mu_c^2 \right)},
\]

\( \gamma(\cdot) \) and \( \beta(\cdot) \) have the same architectures and learn the parameters for modulating the scales and biases, respectively. We here take \( \gamma(\cdot) \) as an example, which consists of two separated convolutional neural networks to produce coarse and fine-grained guidance for modulation. The one network \( \gamma^s(\cdot) \) takes the semantic map \( M \) to learn the coarse modulation parameters. The other network \( \gamma^r(\cdot) \) takes the retrieved image \( I^r \) to learn the pixel-level fine-grained modulation parameters, and we also take the semantic map \( M \) to modulate the intermediate features with AdaIN blocks.

\[
\gamma_{c,y,x} (I^r, M) = \alpha_{\gamma} \gamma_{c,y,x}^s (M) + (1 - \alpha_{\gamma}) \gamma_{c,y,x}^r (I^r, M),
\]

\( \beta_{c,y,x} (I^r, M) = \alpha_{\beta} \beta_{c,y,x}^s (M) + (1 - \alpha_{\beta}) \beta_{c,y,x}^r (I^r, M),\)

where the \( 0 < \alpha_{\gamma}, \alpha_{\beta} < 1 \) are learnable scalars.

**Discriminator.** In practice, we adopt two multi-scale discriminators proposed by [3] to facilitate our model training. As shown in Fig. C, the discriminator consists of two pathways and processes the RGB image and the semantic labels respectively; then the final features are merged by element-wise addition and element-wise multiplication.

### C. Additional Ablation Studies

#### Comparison with SIMS.** Also introducing an image synthesis mechanism based on reference, SIMS [5] simply takes the retrieved image as network input, resulting in low mIOU and blurs shown as Fig. D and Table 1. While our
method leverages the retrieved images to provide pixel level fine-grained guidance via spatially adaptive normalization, making it more effective in synthesizing photo-realistic images.

**Variants of RESAIL.** We compare our RESAIL module with 4 variants and in each comparison experiment we employ the same generator architecture while only replacing the RESAIL ResBlk with other variants. We show the different ResBlks in Fig. E. In SPADE, we just employ the module proposed by [4]. In SPADE+, semantic map concatenating with the guidance image is convolved to produce the modulation parameters $\beta$ and $\gamma$. In Pix2pixHD+, we concatenate the feature with the semantic map and the guidance image following with convolution layer, and we discard the encoder part of Pix2pixHD [9]. In SEAN+, we extract per region style vectors from the guidance image with a style encoder network and input the style vector and semantic map into the SEAN [11] module. Limited by GPU memory, dimension of style vector is set to 128.

**Effectiveness of $L_{seg}$.** To prompt the model to synthesize images aligning well with the semantic layout, we introduce a pretrained segmentation network $C$ to classify each pixel of the generated image and optimize the segmentation loss $L_{seg}$. The designed segmentation network $C$ follows [7], which consists of 12 ResBlks based on a U-Net architecture as shown in Fig. F. We report the results of training our model with and without $L_{seg}$ on Cityscapes [2] in Table A. From the table, we can see segmentation loss $L_{seg}$ improves the learning process. Albeit $L_{seg}$ helps segmentation based metrics, it may introduce inconsistent edge transitions among instances, occurring in [7] which introduces a discriminator based on a segmentation network shown as Fig. G. However, with other losses (e.g., GAN loss and perceptual loss) prompting model training, this kind of artifacts are suppressed and no obvious transitions are found in our results with $L_{seg}$.

**Effect of Shape Non-similarity Threshold.** Computed as Eq. 4, non-similarity $\sigma$ is adopted to measure the shape
consistency between two segment masks. We have tested the FID results by adopting different non-similarity thresholds. From Table B, higher threshold (i.e., using more non-similar guidance) leads to worse guidance, resulting in worse FID.

**D. Additional Visual Results**

To demonstrate the effectiveness of our method on synthesizing the photo-realistic images, we show more visual results in this section. Fig. H ~ J show the comparisons on Cityscapes [2] and as shown in figures, our synthesized images are more photo-realistic with fine details. Fig. K and Fig. M show more results on ADE20K [10]. Comparisons on COCO-Stuff [1] can be found in Fig. L. The guidance image and its corresponding generated image are shown as Fig. N and Fig. O.
Figure F. Segmentation network. (a) The network is designed based on U-Net. (b) Each downsampling or upsampling operation employs a ResBlk.

Figure G. Effect of segmentation loss $L_{\text{seg}}$. Red rectangles mark the affected instances. OASIS suffers from inconsistent edge transitions whose discriminator based on a segmentation network. With the help of other losses (e.g., GAN loss and perceptual loss), no obvious edge transitions are found in our results with $L_{\text{seg}}$.

References


Figure H. Comparison results on Cityscapes.
Figure I. Comparison results on Cityscapes.
Figure J. Comparison results on Cityscapes.
Figure K. Comparison results on ADE20K.
<table>
<thead>
<tr>
<th>Semantic Map</th>
<th>Ground-truth</th>
<th>SPADE</th>
<th>CC-FPSE</th>
<th>OASIS</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Semantic Map" /></td>
<td><img src="image2" alt="Ground-truth" /></td>
<td><img src="image3" alt="SPADE" /></td>
<td><img src="image4" alt="CC-FPSE" /></td>
<td><img src="image5" alt="OASIS" /></td>
<td><img src="image6" alt="Ours" /></td>
</tr>
<tr>
<td><img src="image7" alt="Semantic Map" /></td>
<td><img src="image8" alt="Ground-truth" /></td>
<td><img src="image9" alt="SPADE" /></td>
<td><img src="image10" alt="CC-FPSE" /></td>
<td><img src="image11" alt="OASIS" /></td>
<td><img src="image12" alt="Ours" /></td>
</tr>
<tr>
<td><img src="image13" alt="Semantic Map" /></td>
<td><img src="image14" alt="Ground-truth" /></td>
<td><img src="image15" alt="SPADE" /></td>
<td><img src="image16" alt="CC-FPSE" /></td>
<td><img src="image17" alt="OASIS" /></td>
<td><img src="image18" alt="Ours" /></td>
</tr>
<tr>
<td><img src="image19" alt="Semantic Map" /></td>
<td><img src="image20" alt="Ground-truth" /></td>
<td><img src="image21" alt="SPADE" /></td>
<td><img src="image22" alt="CC-FPSE" /></td>
<td><img src="image23" alt="OASIS" /></td>
<td><img src="image24" alt="Ours" /></td>
</tr>
</tbody>
</table>

Figure L. Comparison results on COCO-Stuff.
Figure M. Comparison results on ADE20K.
Figure N. Synthesis results on ADE20K.
Figure O. Synthesis results on Cityscapes.
<table>
<thead>
<tr>
<th>Semantic Map</th>
<th>Ground-truth</th>
<th>Synthesis</th>
<th>Semantic Map</th>
<th>Ground-truth</th>
<th>Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Semantic Map" /></td>
<td><img src="image2.png" alt="Ground-truth" /></td>
<td><img src="image3.png" alt="Synthesis" /></td>
<td><img src="image4.png" alt="Semantic Map" /></td>
<td><img src="image5.png" alt="Ground-truth" /></td>
<td><img src="image6.png" alt="Synthesis" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Semantic Map" /></td>
<td><img src="image8.png" alt="Ground-truth" /></td>
<td><img src="image9.png" alt="Synthesis" /></td>
<td><img src="image10.png" alt="Semantic Map" /></td>
<td><img src="image11.png" alt="Ground-truth" /></td>
<td><img src="image12.png" alt="Synthesis" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Semantic Map" /></td>
<td><img src="image14.png" alt="Ground-truth" /></td>
<td><img src="image15.png" alt="Synthesis" /></td>
<td><img src="image16.png" alt="Semantic Map" /></td>
<td><img src="image17.png" alt="Ground-truth" /></td>
<td><img src="image18.png" alt="Synthesis" /></td>
</tr>
</tbody>
</table>

Figure P. Synthesis results on ADE20K(top) and Cityscapes(bottom).

