Retrieval-based Spatially Adaptive Normalization for Semantic Image Synthesis Supplemental Materials

Yupeng Shi [1], Xiao Liu [1], Yuxiang Wei [2], Zhongqin Wu [1], Wangmeng Zuo[2,3] (✉)
[1]Tomorrow Advancing Life, [2]Harbin Institute of Technology, [3]Peng Cheng Laboratory
{csypshi, ender.liux, yuxiang.wei.cs}@gmail.com wuzhongqin@tal.com wmzuo@hit.edu.cn

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

$$I, M = \{(M_i^s, y_i^c, I_i^s)\}, \tag{1}$$

where M_i^s, y_i^c and I_i^s 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 (M_j^s, y_j^c, I_j^s) 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,

$$\sigma_{scale}\left(M_i^s, M_j^s\right) = \begin{cases} 0, & t \geq 0.5 \\ 1, & t < 0.5 \end{cases}, \tag{2}$$

$$\sigma_{shape}\left(M_i^s, M_j^s\right) = \frac{SSD\left(\hat{M}_i^s, \hat{M}_j^s\right)}{\max\left(\left\|\hat{M}_i^s\right\|_1, \left\|\hat{M}_j^s\right\|_1\right)}, \tag{3}$$

where $t = \frac{\min\left(\|M_i^s\|_1, \|M_j^s\|_1\right)}{\max\left(\|M_i^s\|_1, \|M_j^s\|_1\right)}$. $\hat{M}_i^s$ and $\hat{M}_j^s$ denote the resized versions (*i.e.*, 128×128) 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,

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

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

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

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

A.2. Distortion of Ground-truth Image.

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_t^s with the same category. Specifically, we first convert I_i^s

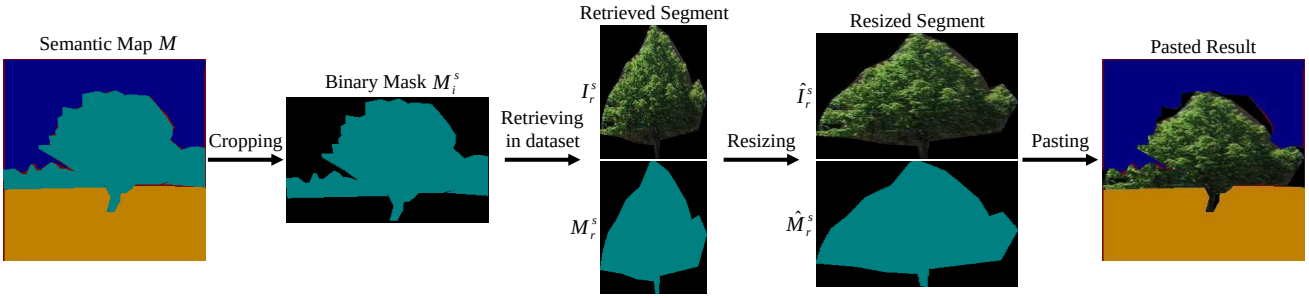


Figure A. Process to paste a retrieved segment into the semantic map. We here take "Tree" labeled in cyan as an example.

and I_t^s from RGB space into $l\alpha\beta$ space. Then the color transferred image $\tilde{I}_i^s$ in each channel of $l\alpha\beta$ space is calculated by,

$$\begin{aligned}\tilde{l}_i &= (l_i - \mu(l_i)) \cdot \frac{\sigma(l_t)}{\sigma(l_i)} + \mu(l_t) \\ \tilde{\alpha}_i &= (\alpha_i - \mu(\alpha_i)) \cdot \frac{\sigma(\alpha_t)}{\sigma(\alpha_i)} + \mu(\alpha_t) \\ \tilde{\beta}_i &= (\beta_i - \mu(\beta_i)) \cdot \frac{\sigma(\beta_t)}{\sigma(\beta_i)} + \mu(\beta_t)\end{aligned} \tag{5}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation of corresponding channel. Finally, we convert $\tilde{I}_i^s$ 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 $\tilde{I}^{gt}$ 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 $\tilde{I}^{gt}$) and the semantic map M as input and learns to modulate the activations. We here represent the input activations as $\mathbf{h}$ with a batch of N samples. H, W and C denote the height, width and the number of channels in $\mathbf{h}$, and the modulated activations at site $(n \in N, c \in C, y \in H, x \in W)$ is represented as,

$$\begin{aligned}RESAIL(\mathbf{h}, I^r, M) = \gamma_{c,y,x}(I^r, M) \frac{\mathbf{h}_{n,c,y,x} - \mu_c}{\sigma_c} \\ + \beta_{c,y,x}(I^r, M),\end{aligned} \tag{6}$$

where μ_c and σ_c denote the mean and standard deviation of the activation in channel c,

$$\begin{aligned}\mu_c &= \frac{1}{NHW} \sum_{n,y,x} \mathbf{h}_{n,y,x} \\ \sigma_c &= \sqrt{\frac{1}{NHW} \left(\sum_{n,y,x} \mathbf{h}_{n,y,x}^2 \right) - \mu_c^2}.\end{aligned} \tag{7}$$

$\gamma(\cdot)$ 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.

$$\begin{aligned}\gamma_{c,y,x}(I^r, M) &= \alpha_\gamma \gamma^s_{c,y,x}(M) + (1 - \alpha_\gamma) \gamma^r_{c,y,x}(I^r, M) \\ \beta_{c,y,x}(I^r, M) &= \alpha_\beta \beta^s_{c,y,x}(M) + (1 - \alpha_\beta) \beta^r_{c,y,x}(I^r, M)\end{aligned}, \tag{8}$$

where the $0 < \alpha_\beta, \alpha_\gamma < 1$ 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

Figure B. Distortion of ground-truth images. The top row shows the produced retrieval-based guidance images; the middle row shows the distorted ground-truth and the bottom row shows the corresponding ground-truth images.

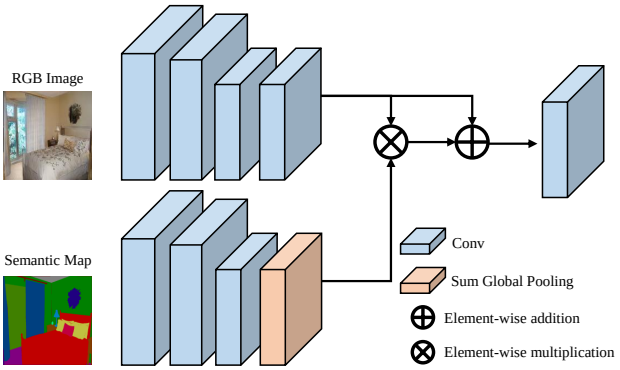


Figure C. Discriminator network.

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 β and γ. 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.

Table A. Ablation study of $\mathcal{L}_{seg}$ in Cityscapes dataset. It shows that $\mathcal{L}_{seg}$ facilitates the model learning.

$\mathcal{L}_{seg}$	FID(↓)	mIOU(↑)	AC(↑)
✗	46.8	66.3	82.7
✓	**45.5**	**69.7**	**83.2**

Table B. FID w.r.t non-similarity threshold.

Threshold	0.15	0.25	0.35	0.45	0.55	0.58
FID	45.49	46.38	48.18	48.3	50.56	51.04

Effectiveness of $\mathcal{L}_{seg}$. 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 $\mathcal{L}_{seg}$. 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 $\mathcal{L}_{seg}$ on Cityscapes [2] in Table A. From the table, we can see segmentation loss $\mathcal{L}_{seg}$ improves the learning process. Albeit $\mathcal{L}_{seg}$ 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 $\mathcal{L}_{seg}$.

Effect of Shape Non-similarity Threshold. Computed as Eq. 4, non-similarity σ is adopted to measure the shape

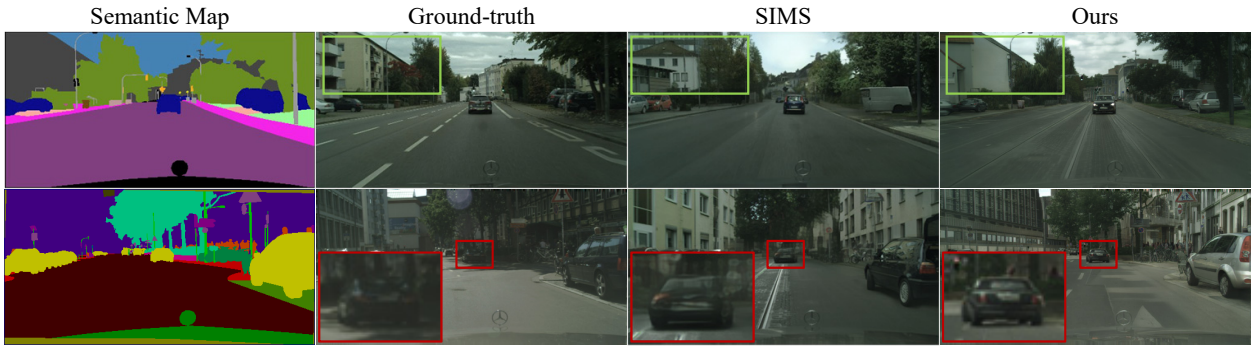
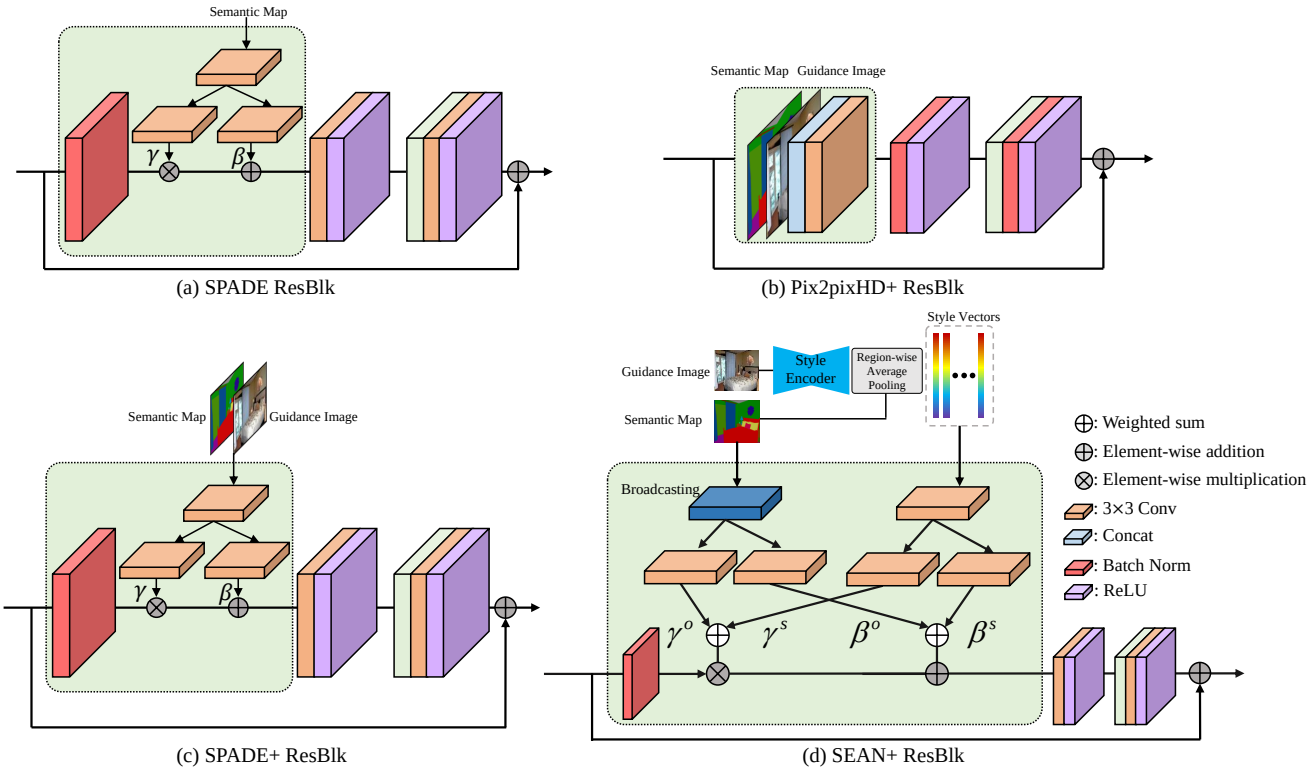


Figure D. Comparison with SIMS. SIMS suffers from low mIOU (marked in green rectangle) and blurs (marked in red rectangle) of some objects.



(a) SPADE ResBlk

(b) Pix2pixHD+ ResBlk

(c) SPADE+ ResBlk

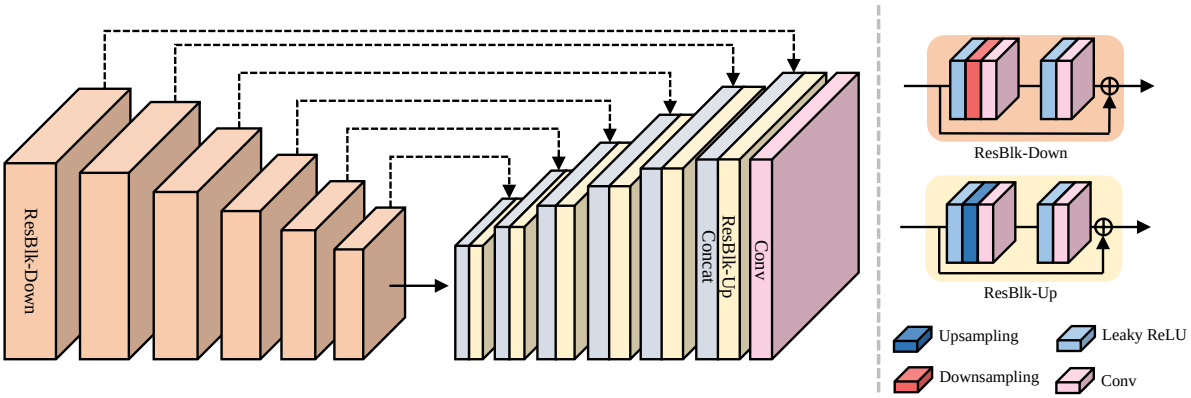
(d) SEAN+ ResBlk

Figure E. Variants of RESAIL ResBlk. (a) *SPADE* employs the SPADE module; (b) *Pix2pixHD+* denotes concatenating the guidance into the conv layer of pix2pixHD model. (c) *SPADE+* denotes using the guidance as input to the SPADE module. (d) *SEAN+* denotes using the guidance as input to the SEAN module.

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 $\sim$ 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.



(a) Segmentation Network (b) Residual Block

Figure F. Segmentation network. (a) The network is designed based on U-Net. (b) Each downsampling or upsampling operation employs a ResBlk.

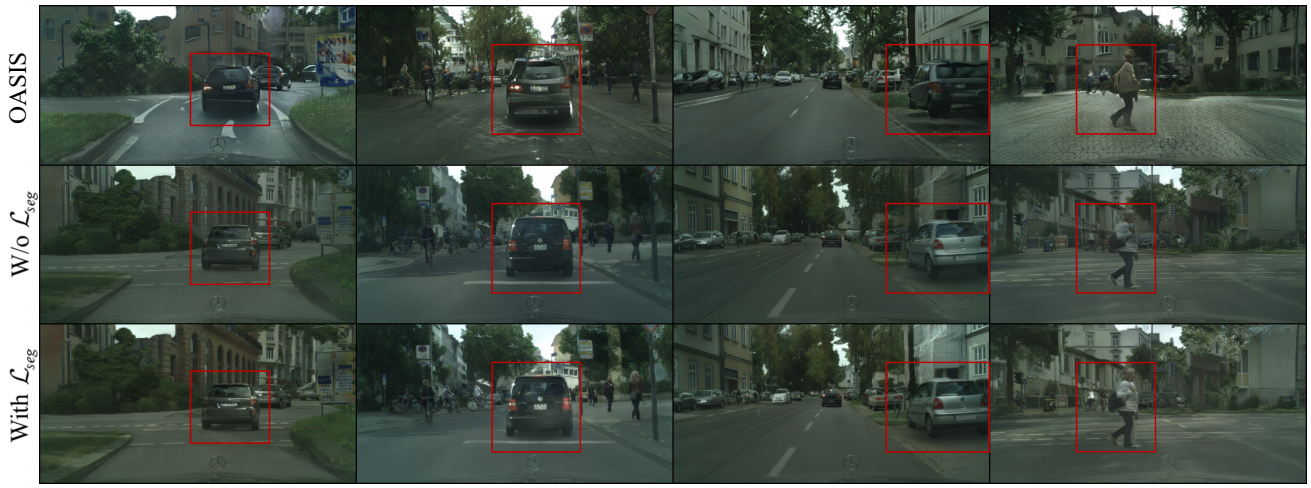


Figure G. Effect of segmentation loss $\mathcal{L}_{seg}$. Red rectangles mark the affected instances. OASIS suffers from inconsistent edge transitions whose discriminator based on a segmentation network. With the help of other losses (*e.g.*, GAN loss and perceptual loss), no obvious edge transitions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

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 4

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 3, 4

[3] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: semantic editing of scenes by adding, manipulating or erasing objects. In *Proceedings of the European Conference on Computer Vision*, pages 394–411, 2020. 2

[4] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3

[5] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8808–8816, 2018. 2

[6] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 1

[7] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations*, 2021. 3

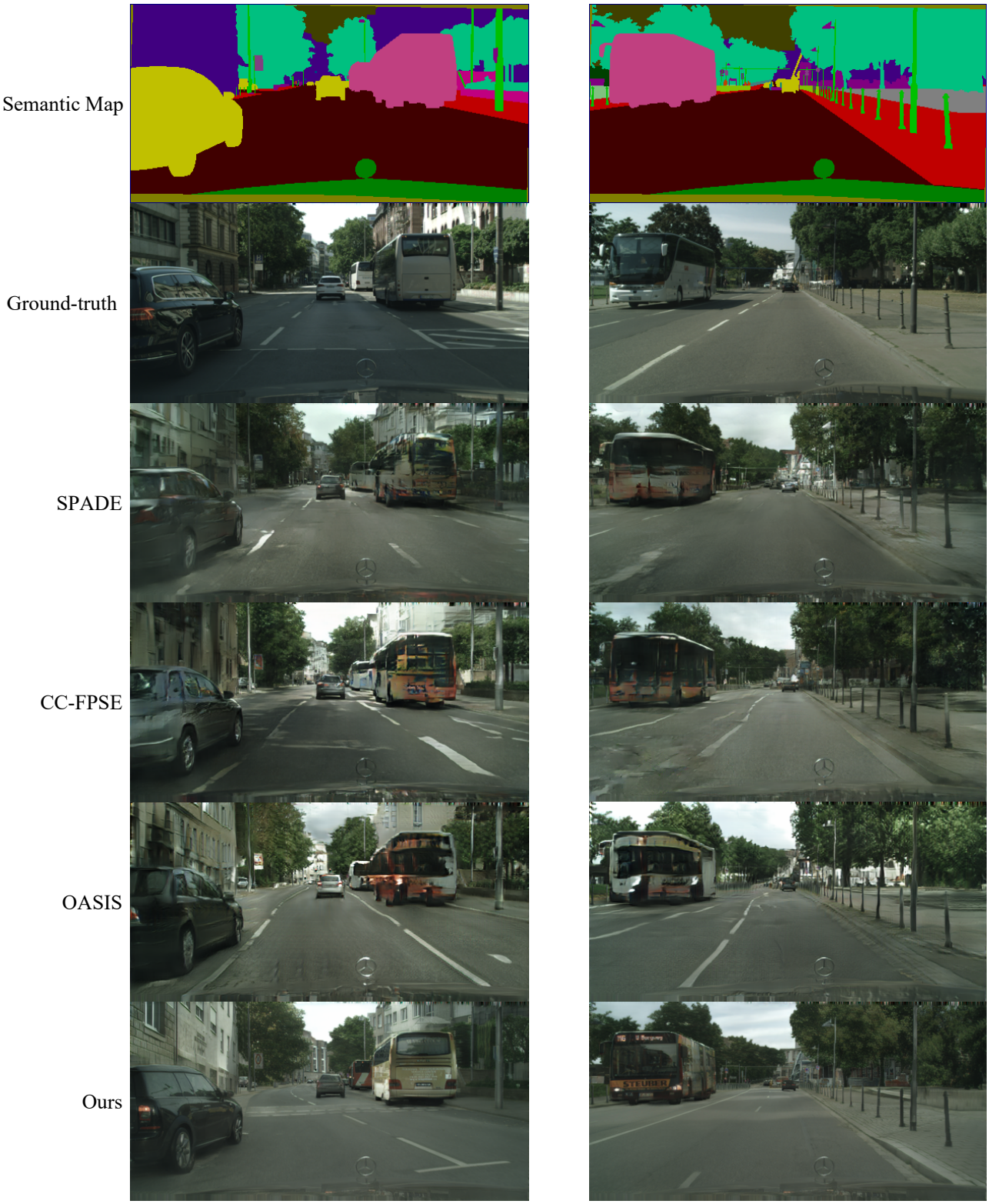


Figure H. Comparison results on Cityscapes.

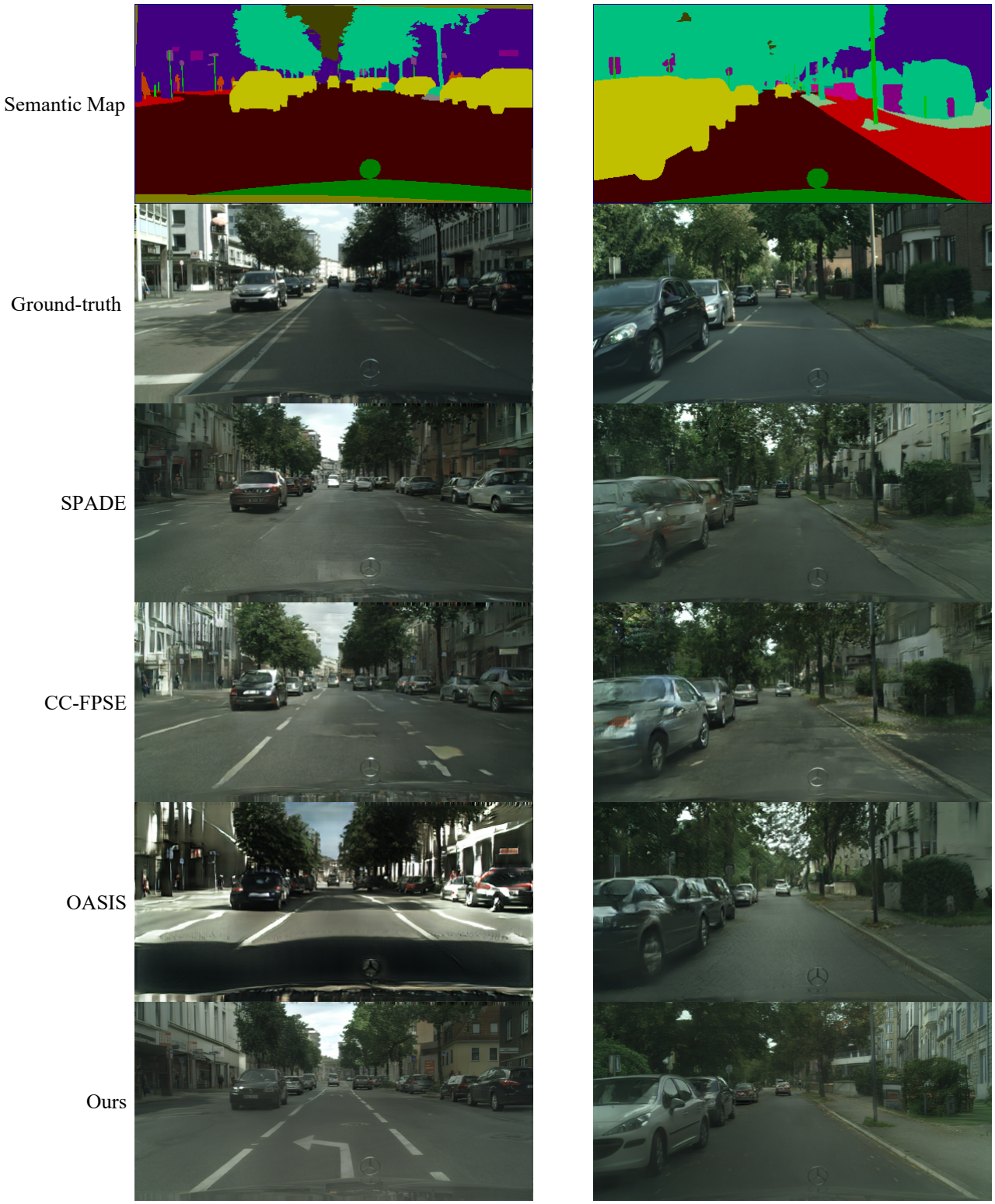


Figure I. Comparison results on Cityscapes.

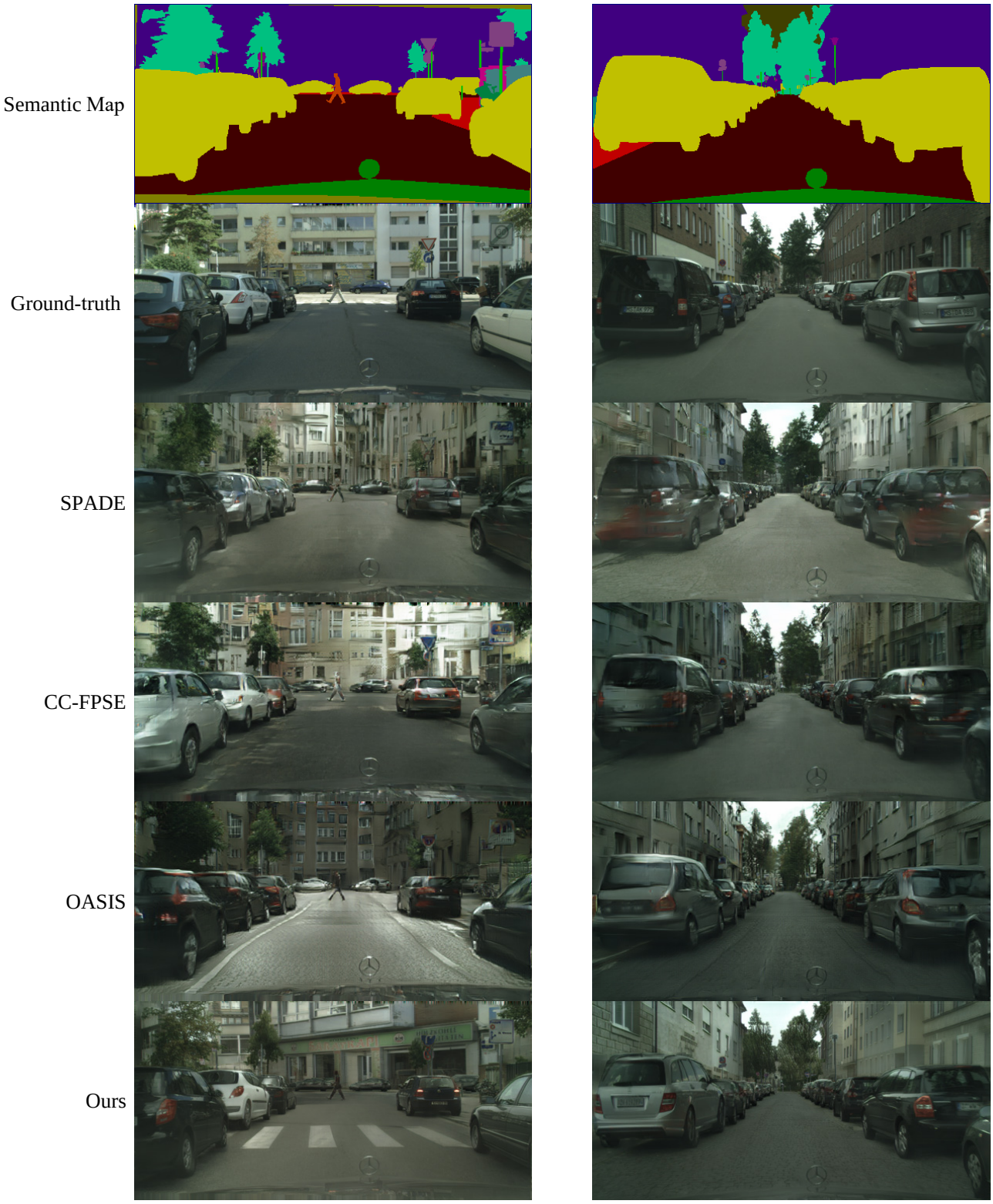


Figure J. Comparison results on Cityscapes.

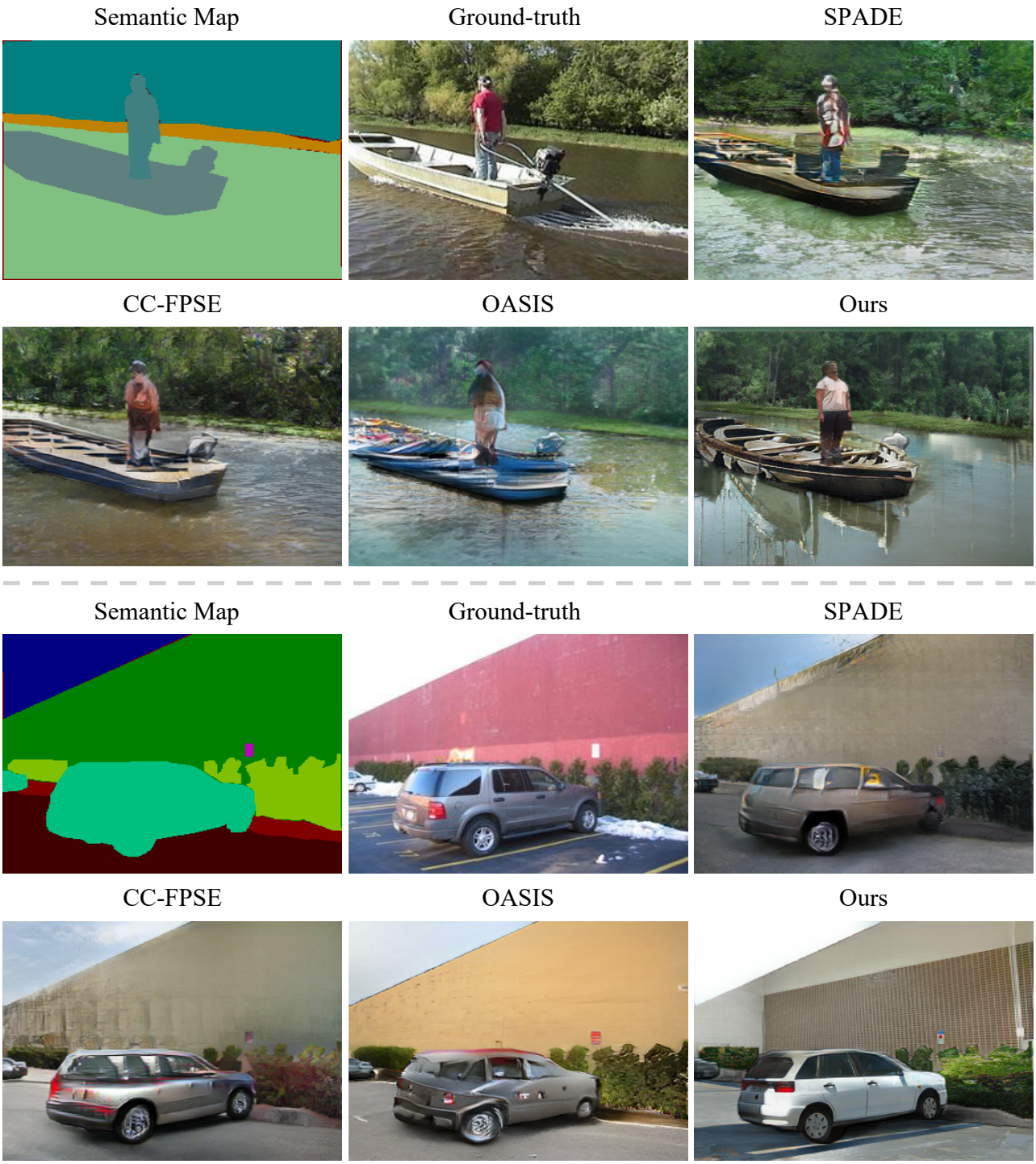


Figure K. Comparison results on ADE20K.

Figure L. Comparison results on COCO-Stuff.

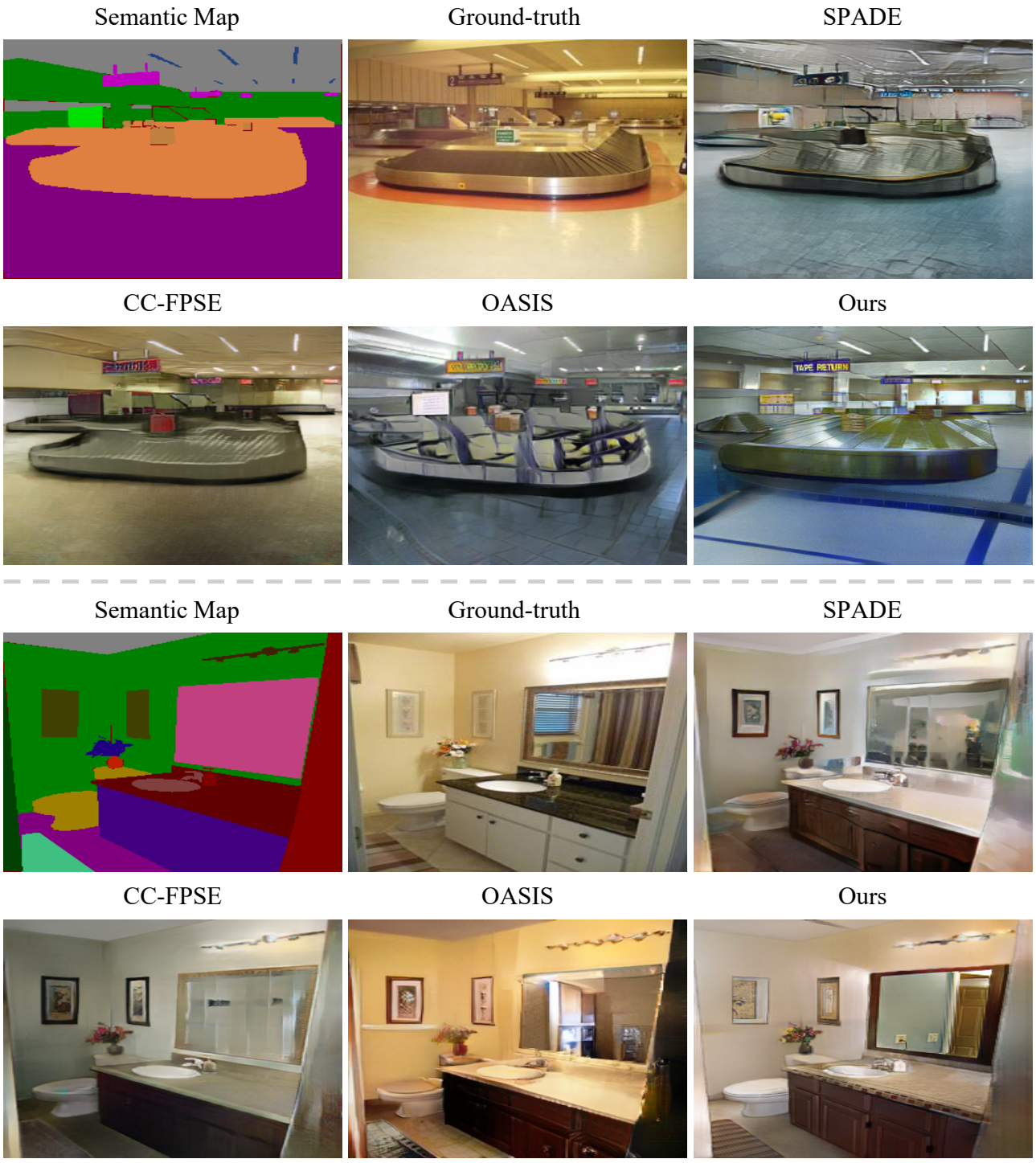


Figure M. Comparison results on ADE20K.

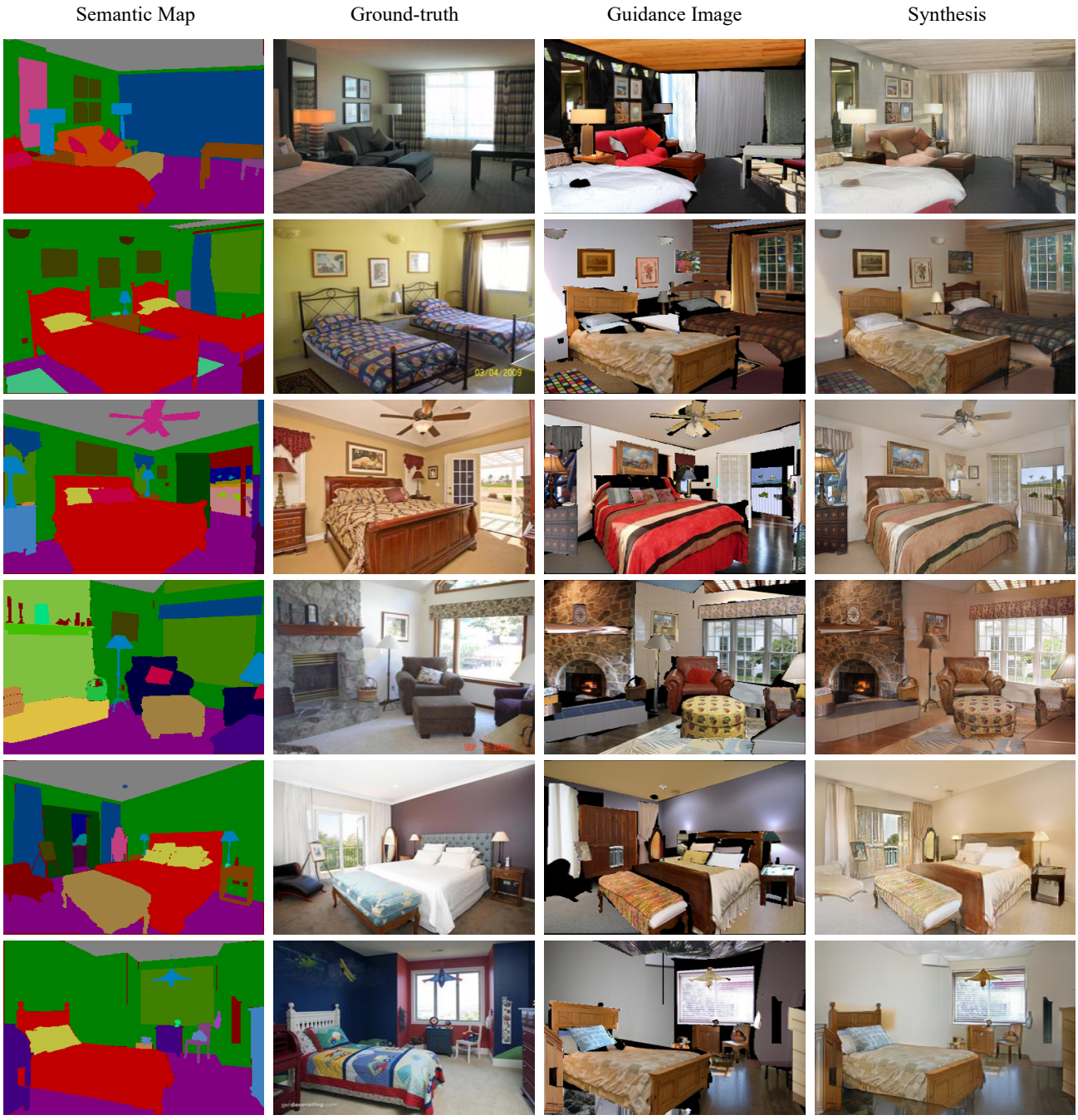


Figure N. Synthesis results on ADE20K.

Figure O. Synthesis results on Cityscapes.

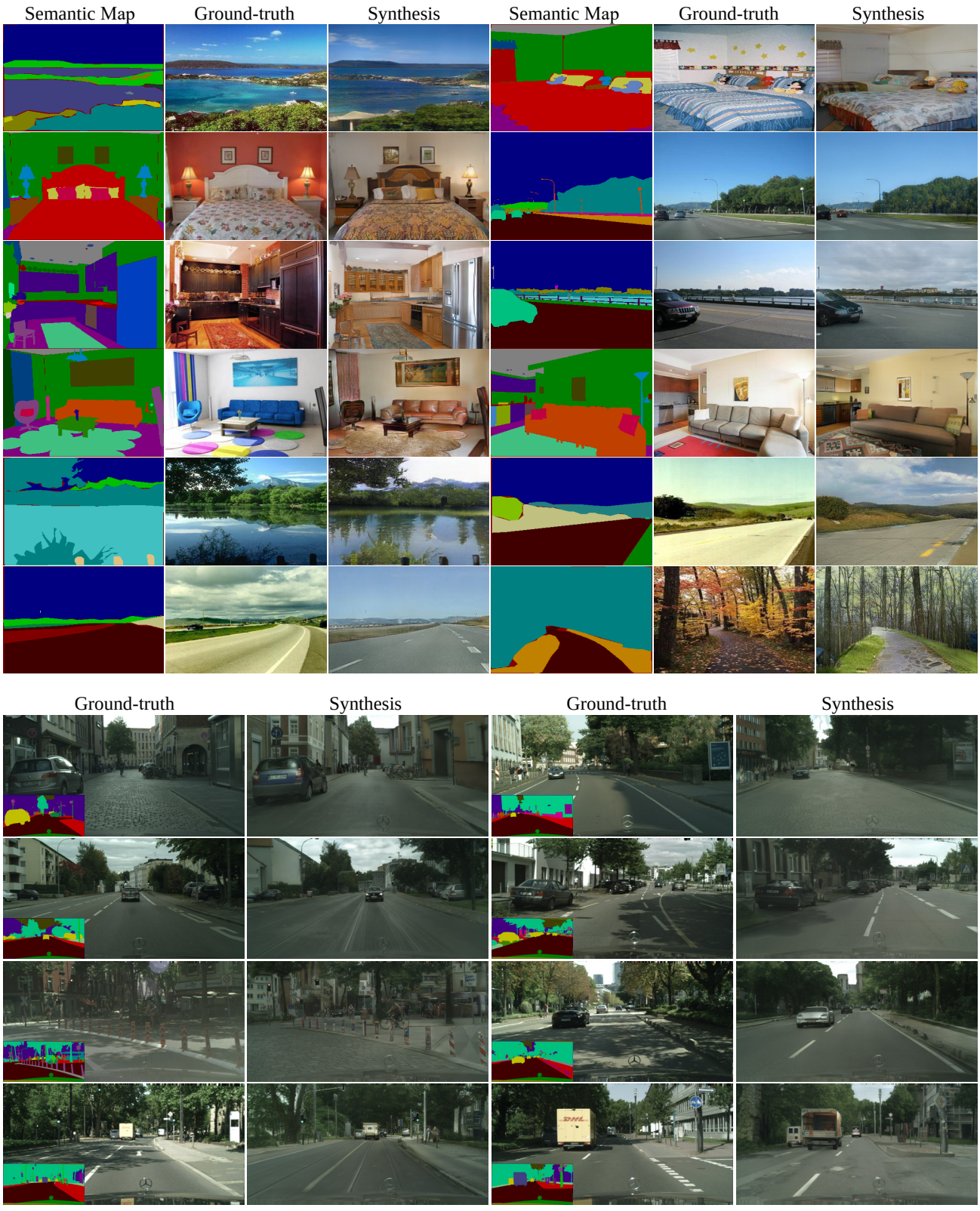


Figure P. Synthesis results on ADE20K(top) and Cityscapes(bottom).

[8] Hao Wang, Qilong Wang, Hongzhi Zhang, Jian Yang, and Wangmeng Zuo. Constrained online cut-paste for object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 1

[9] 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. 3

[10] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 4

[11] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 3