Supplementary Materials: Dual Pyramid Generative Adversarial Networks for Semantic Image Synthesis

Shijie Li[1]
lsj@uni-bonn.de
Ming-Ming Cheng[2]
cmm@nankai.edu.cn
Juergen Gall[1]
gall@iai.uni-bonn.de

[1] Computer Vision Group
University of Bonn
Bonn, Germany

[2] Media Computing Lab.
Nankai University,
Tianjin, China

1 Additional Ablation Studies

Quality of the generated objects In order to analyze the quality of the generated objects, we crop the instances of the objects from the generated images. The bounding boxes are obtained by the label maps. Fig. 1 shows examples of such crops. From these examples, we can see that our method generates more realistic objects at all sizes. We evaluate the quality of the objects by computing the FID score for the cropped objects instead of the images. The results are shown in Table 1. The FID score for SPADE is slightly lower than for OASIS although the generated objects do not look better. Nevertheless, the FID score of our approach is much lower. We furthermore split the objects into large, medium, and small objects based on the size of the label mask. There are 69 large objects ($>$ 10000 pixels), 378 medium objects, and 3190 small objects ($<$ 2500 pixels). Our approach performs well for objects of all sizes. Note that the FID scores between the columns are not comparable since each set (all, large, medium, small) contains a different number of objects. We furthermore report the FID score on Cityscapes if we increase or decrease the image resolution by factor 2 in Table 2. Our approach also performs better for different image resolutions.

Impact of feature map concatenation and LabelMix Finally, Table 3 shows the results when we do not concatenate (w/o cat) the feature map α_0 in (2). We can see that both FID and mIoU are worse. For completeness, we also report the impact of $\mathcal{L}_{LM}$. Without $\mathcal{L}_{LM}$ (w/o $\mathcal{L}_{LM}$) FID and mIoU are worse, which confirms the effectiveness of the LabelMix regularization.

User study We further evaluated the quality of the generated images by a user study. To this end, we showed the participants three images in randomized order that have been generated for the same label map by SPADE, OASIS, and DP-GAN, respectively. The label map itself was not shown. The participants needed then to select the image among the three

	all	large	medium	small
SPADE	33.4	84.3	63.3	36.8
OASIS	35.0	98.6	79.9	37.6
DP-GAN	24.5	77.6	57.5	27.1

Table 1: FID for cropped objects from the Cityscapes dataset.

	0.5×	1.0×	2.0×
SPADE	99.9	71.8	145.9
OASIS	91.8	47.7	115.5
DP-GAN	85.0	44.1	96.1

Table 2: FID for different image resolutions of the Cityscapes dataset.

	w/o $\mathcal{L}_{LM}$	w/o cat	Ours
FID	46.8	44.8	**44.1**
mIoU	70.9	71.2	**73.6**

Table 3: Ablation study on Cityscapes. "w/o *LM*" means without LabelMix regularization $\mathcal{L}_{LM}$ in (9). "w/o cat" means without concatenating feature map α_0 in (2).

images which looked most realistic. For the user study, we used all 500 label maps of the validation set of Cityscapes. The label maps were equally assigned to 10 participants such that each participant rated 50 label maps. The results of the user study are shown in Table 4. The participants selected in more than 60% of the cases, the image that has been generated by our method.

Efficiency Analysis We report the model size and time for inference in Table 5 and compare it to other approaches. The inference time has been measured for a single TITAN RTX. Compared to OASIS, the inference time is reduced by more than 14%. This shows that our method outperforms the state-of-the-art not only in terms of image generation quality, but it is also more efficient.

Multi-Modal Image Synthesis Fig. 2 shows examples where we generate four images from the same label map with different styles. This is achieved by randomly sampling a 3D noise tensor. Our method can produce diverse high quality images for both indoor and outdoor scenarios. The color, texture, and illumination vary, but the semantic structure is maintained which is desired.

2 Details of Architecture and Training

The details of the two pyramids of the generator are shown in Table 6 and Table 8, respectively. The discriminator is shown in Table 7. It is a U-Net architecture built from ResNet blocks. For training, we use the Adam optimizer [2] with momenta $\beta = (0, 0.999)$. The learning rates for the generator and discriminator are set to 0.0001 and 0.0004, respectively. Our method also uses an exponential moving average [1] for the generator weights with 0.9999 decay. As in [3], we use $\lambda_{LM} = 5$. All experiments have been conducted on a single TITAN RTX with a fix random seed. The source code is available at https://github.com/sj-li/DP_GAN.

	SPADE	OASIS	DP-GAN
Preferred image (%)	13.6	24.2	62.2

Table 4: User study. In 62.2% of the cases, the users considered the image generated by DP-GAN more realistic than the images generated by SPADE or OASIS.

	#Parameters (M)	Inference time (ms)
LGGAN	111.1	222.7
DAGAN	93.1	63.8
CC-FPSE	128.1	130.1
SPADE	93.0	70.8
OASIS	71.1	60.5
Ours	**69.5**	**51.8**

Table 5: Comparison of number of parameters and inference time.

Operation	Input	Size	Output	Size
Concatenate	z y	(64,256,256) (N,256,256)	z_y	(64+N,256,256)
ConvBlock	z_y	(64+N,256,256)	s	(32, 256, 256)
ConvBlock	s	(32, 256, 256)	s_5	(64, 256, 256)
ConvBlock	s_5	(64, 256, 256)	s_4	(64, 128, 128)
ConvBlock	s_4	(64, 128, 128)	s_3	(64, 64, 64)
ConvBlock	s_3	(64, 64, 64)	s_2	(64, 32, 32)
ConvBlock	s_2	(64, 32, 32)	s_1	(64, 16, 16)
ConvBlock	s_1	(64, 16, 16)	s_0	(64, 8, 8)

Table 6: Spatial adaptation learning pyramid of the generator. N refers to the number of semantic classes, z is noise sampled from a unit Gaussian, y is the label map, and ConvBlock denotes Conv2d-BatchNorm2d-ReLU block.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

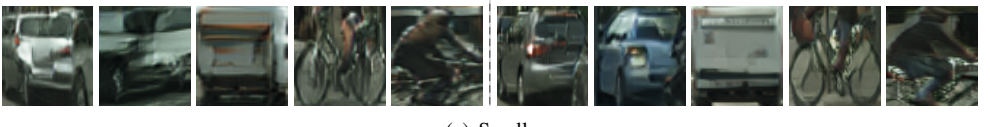
[3] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *ICLR*, 2021.

Operation	Input	Size	Output	Size	Sup
ResBlock-Down	image	(3,256,256)	$down_1$	(128,128,128)	
ResBlock-Down	$down_1$	(128,128,128)	$down_2$	(128,64,64)	
ResBlock-Down	$down_2$	(128,64,64)	$down_3$	(256,32,32)	
ResBlock-Down	$down_3$	(256,32,32)	$down_4$	(256,16,16)	$\mathcal{L}_{patch}$
ResBlock-Down	$down_4$	(256,16,16)	$down_5$	(512,8,8)	
ResBlock-Down	$down_5$	(512,8,8)	$down_6$	(512,4,4)	$\mathcal{L}_{patch}$
ResBlock-Up	$down_6$	(512,4,4)	up_1	(512,8,8)	
ResBlock-Up	cat(up_1, $down_5$)	(1024,8,8)	up_2	(256,16,16)	$\mathcal{L}_{feature}$
ResBlock-Up	cat(up_2, $down_4$)	(512,16,16)	up_3	(256,32,32)	$\mathcal{L}_{feature}$
ResBlock-Up	cat(up_3, $down_3$)	(512,32,32)	up_4	(128,64,64)	$\mathcal{L}_{feature}$
ResBlock-Up	cat(up_4, $down_2$)	(256,64,64)	up_5	(128,128,128)	$\mathcal{L}_{feature}$
ResBlock-Up	cat(up_5, $down_1$)	(256,128,128)	up_6	(64,256,256)	$\mathcal{L}_{feature}$
Conv2D	up_6	(64,256,256)	out	(N+1,256,256)	$\mathcal{L}_{pixel}$

Table 7: Discriminator. N refers to the number of semantic classes. $\mathcal{L}_{pixel}$, $\mathcal{L}_{patch}$, $\mathcal{L}_{feature}$ correspond to pixel, patch and feature supervision, respectively.

(a) Large

(b) Medium



(c) Small

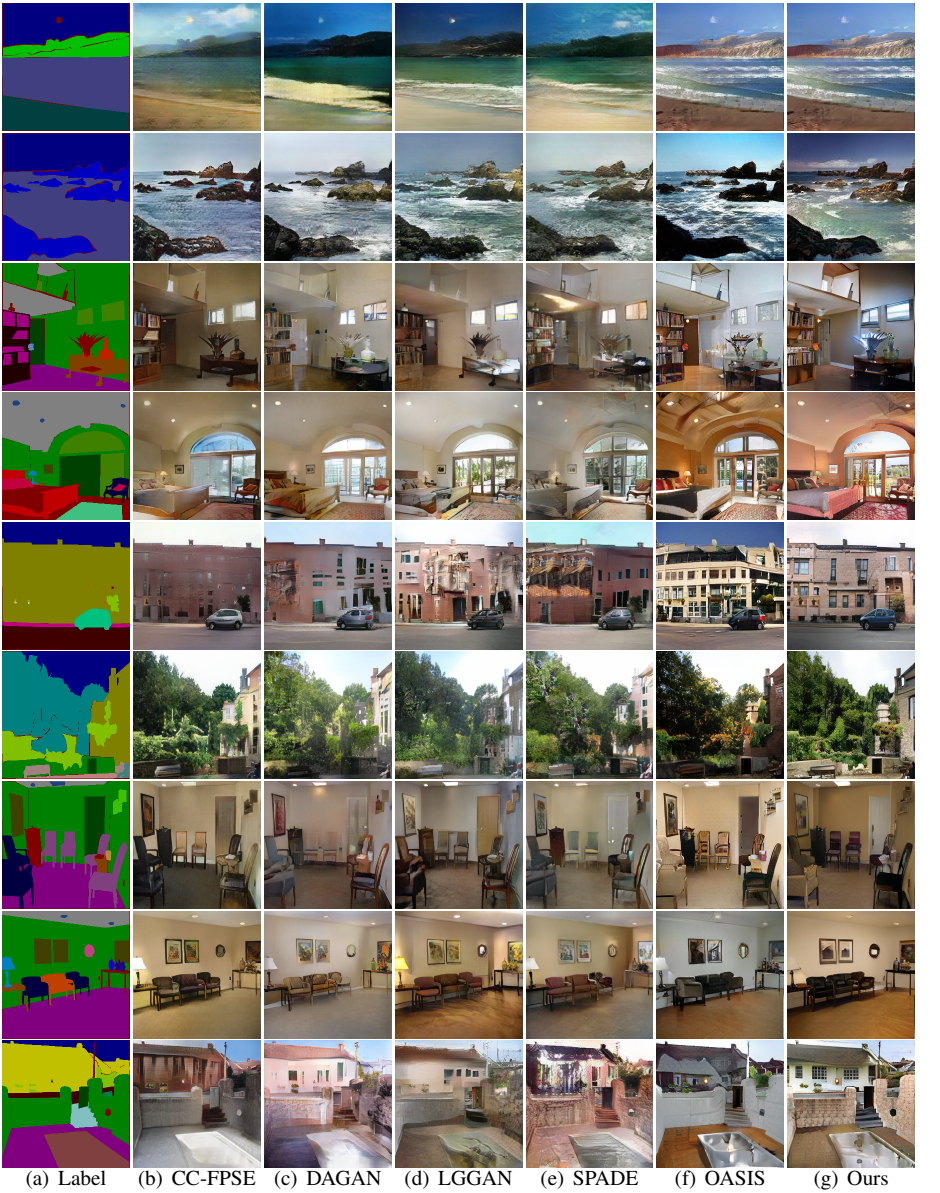
Figure 1: Generated objects of different sizes. The objects on the left hand side are generated by OASIS [3] whereas the objects on the right hand side are generated by our method. Our method generates more realistic objects for all sizes.

(a) Label (b) GT (c) 1 (d) 2 (e) 3 (f) 4

Figure 2: Images generated with four different random noise tensors z.

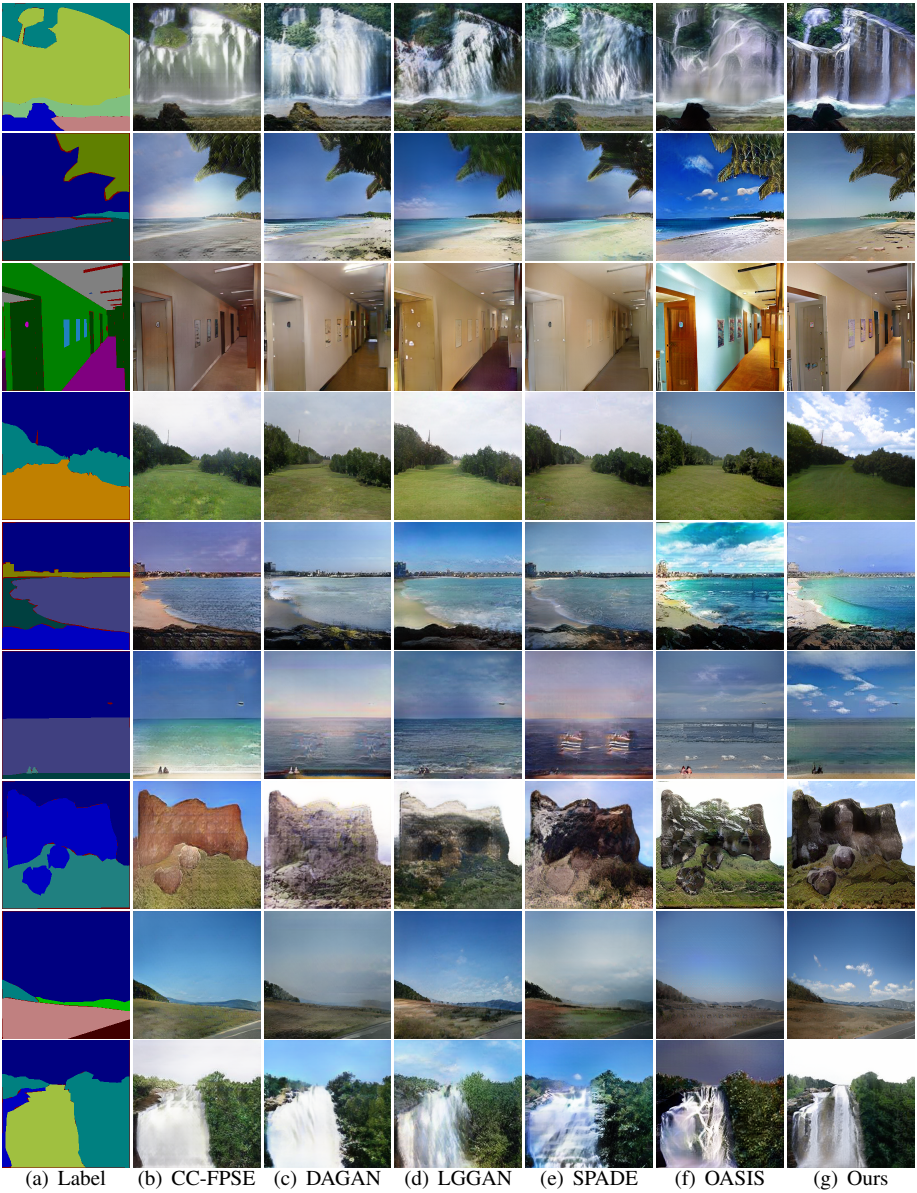
Operation	Input	Size	Output	Size
Concatenate	z y	(64,256,256) (N,256,256)	z_y	(64+N,256,256)
Conv2D	$interp(z\_y)$	(64+N,8,8)	up_0	(1024,8,8)
SPADE-ResBlock	up_0 s_0	(1024,8,8) (64,8,8)	up_1	(1024,16,16)
SPADE-ResBlock	up_1 s_1	(1024,16,16) (64,16,16)	up_2	(512,32,32)
SPADE-ResBlock	up_2 s_2	(512,32,32) (64,32,32)	up_3	(256,64,64)
SPADE-ResBlock	up_3 s_3	(256,64,64) (64,64,64)	up_4	(128,128,128)
SPADE-ResBlock	up_4 s_4	(128,128,128) (64,128,128)	up_5	(64,256,256)
Conv2D, LeakyRelu, TanH	up_5	(64,256,256)	x	(3,256,256)

Table 8: Image synthesis pyramid of the generator. N refers to the number of semantic classes, z is noise sampled from a unit Gaussian, y is the label map, and $interp$ interpolates a given input to the appropriate spatial dimensions of the current layer. s_0-s_5 are from the spatial adaptation learning pyramid shown in Table 6.



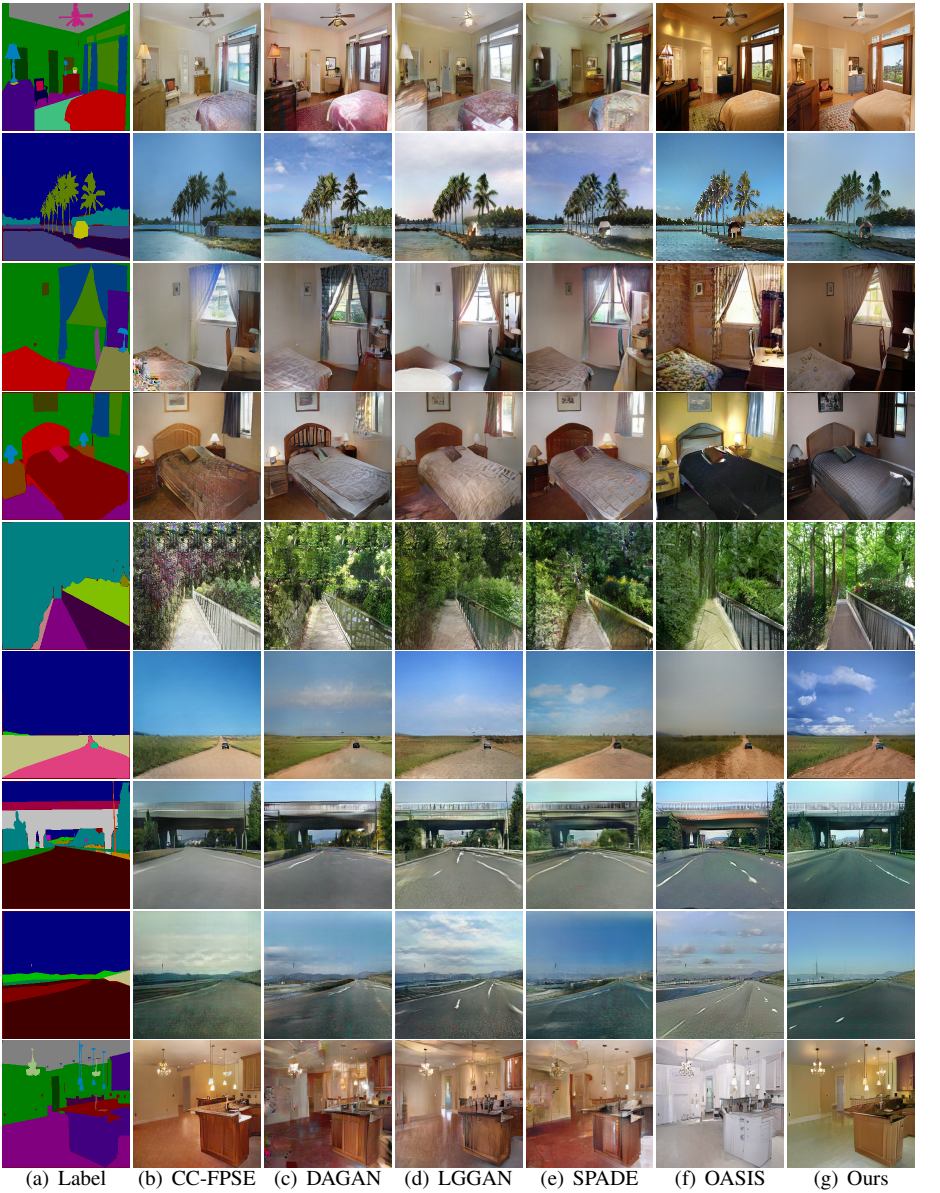
(a) Label (b) CC-FPSE (c) DAGAN (d) LGGAN (e) SPADE (f) OASIS (g) Ours

Figure 3: Qualitative results for ADE20K.



(a) Label (b) CC-FPSE (c) DAGAN (d) LGGAN (e) SPADE (f) OASIS (g) Ours

Figure 4: Qualitative results for ADE20K.



(a) Label (b) CC-FPSE (c) DAGAN (d) LGGAN (e) SPADE (f) OASIS (g) Ours

Figure 5: Qualitative results for ADE20K.

(a) Label (b) CC-FPSE (c) DAGAN (d) LGGAN (e) SPADE (f) OASIS (g) Ours

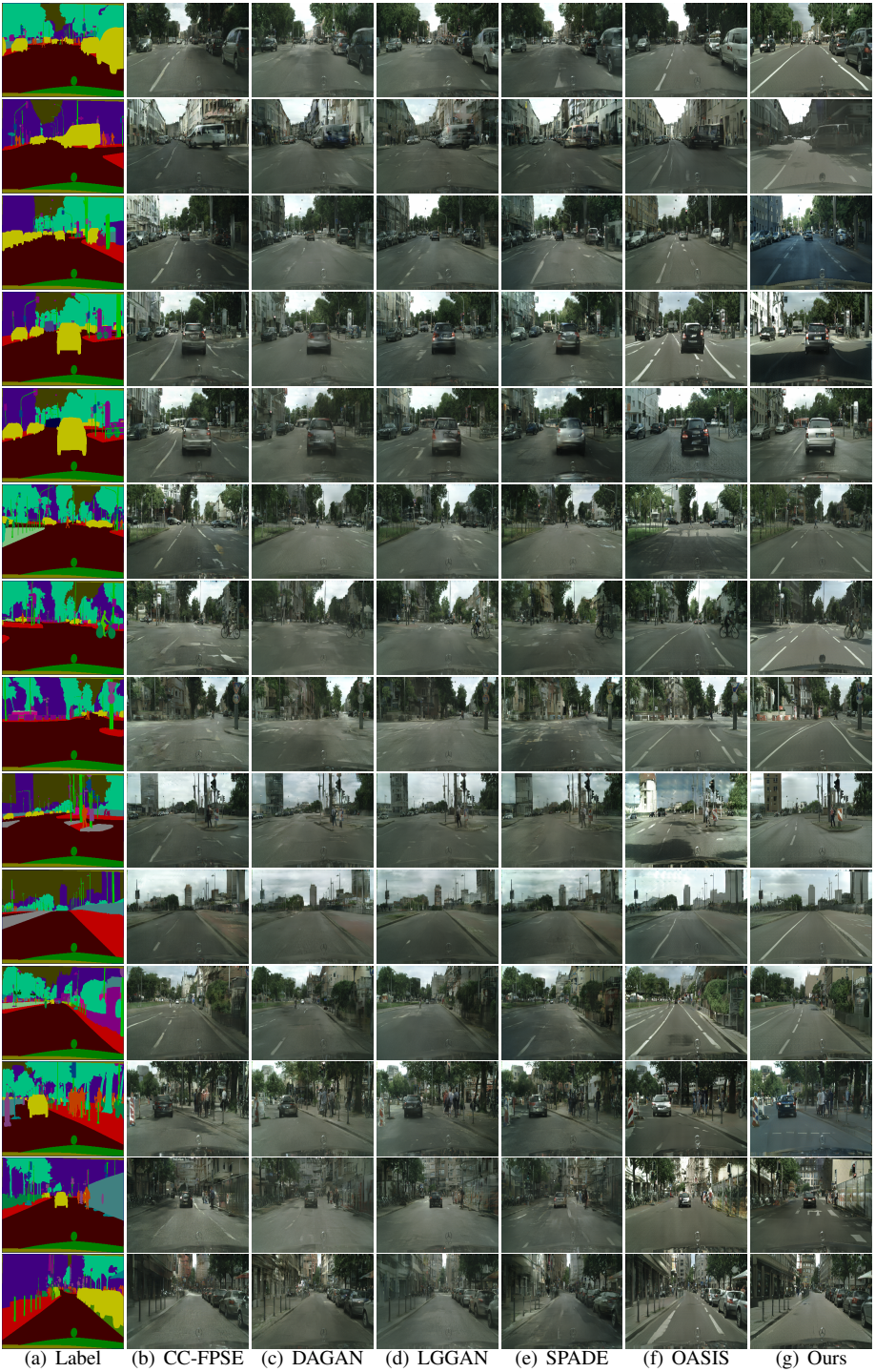
Figure 6: Qualitative results for ADE20K.

(a) Label (b) CC-FPSE (c) DAGAN (d) LGGAN (e) SPADE (f) OASIS (g) Ours

Figure 7: Qualitative results for ADE20K.

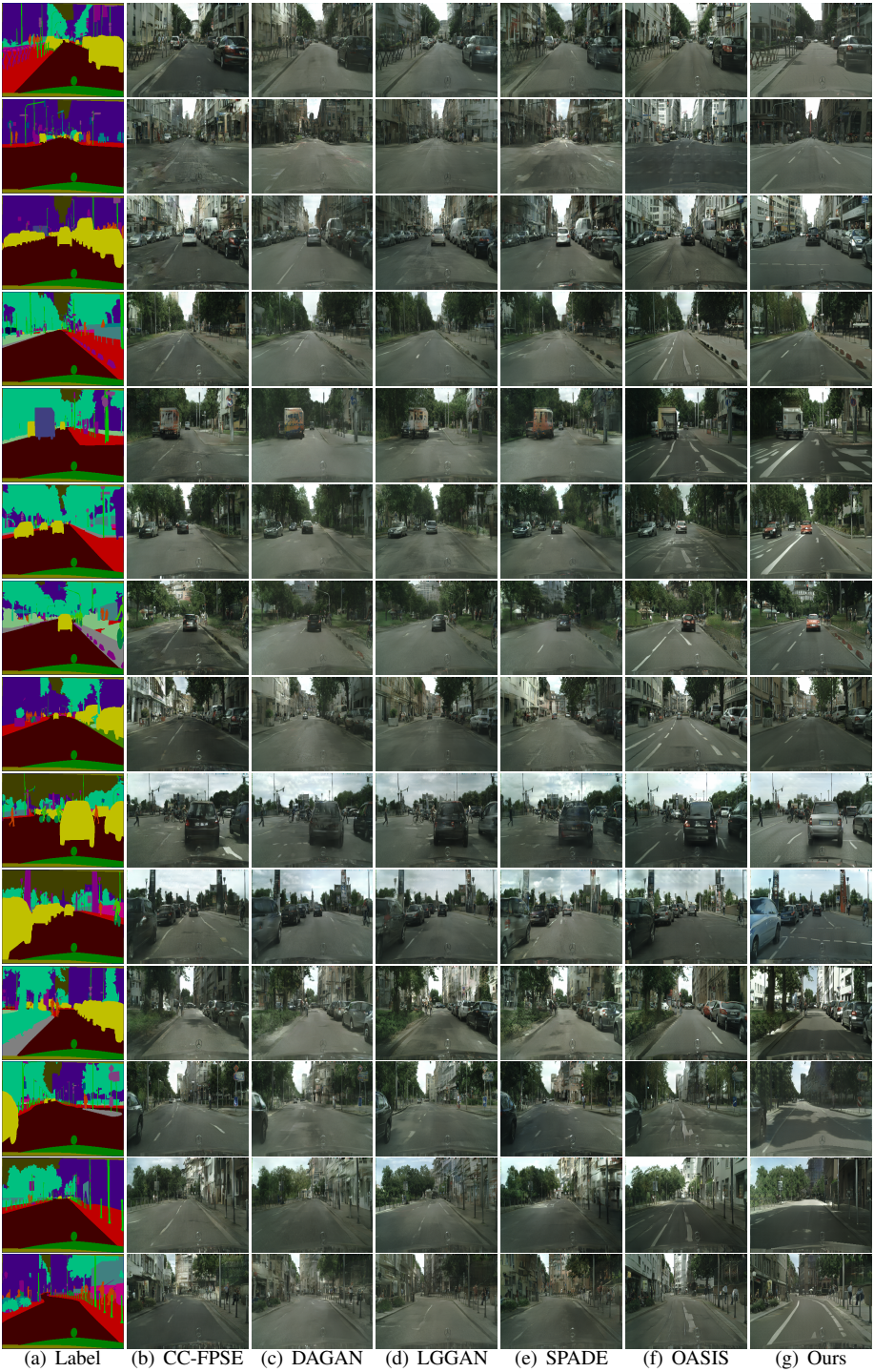
(a) Label (b) CC-FPSE (c) DAGAN (d) LGGAN (e) SPADE (f) OASIS (g) Ours

Figure 8: Qualitative results for ADE20K.



(a) Label (b) CC-FPSE (c) DAGAN (d) LGGAN (e) SPADE (f) OASIS (g) Ours

Figure 9: Qualitative results for Cityscapes.



(a) Label (b) CC-FPSE (c) DAGAN (d) LGGAN (e) SPADE (f) OASIS (g) Ours

Figure 10: Qualitative results for Cityscapes.