Segmentations-Leak: Membership Inference Attacks and Defenses in Semantic Image Segmentation Supplementary Materials

Yang He[1,2], Shadi Rahimian[1], Bernt Schiele[2], and Mario Fritz[1]

[1] CISPA Helmholtz Center for Information Security
[2] Max Planck Institute for Informaticss
Saarland Informatics Campus, Germany
{yang.he, shadi.rahimian, fritz}@cispa.saarland, schiele@mpi-int.mpg.de

1 Details of using *Mapillary Vistas*

In our main paper, we show attack results on *Cityscapes* dataset [2], where we train an attacker with *Mapillary Vistas* dataset [3] in the independent attack setting. *Mapillary Vistas* has 65 defined categories in its label space, while *Cityscapes* has only 19 categories. In the supplementary materials, we provide the details for reproducibility how we utilize *Mapillary Vistas* in our experiments.

To perform a successful attack, we merge the label space in *Mapillary Vistas* compatible to *Cityscapes*. *Mapillary Vistas* provides a fine-grained label space, which covers the concepts and classes in *Cityscapes*. For example, *Mapillary Vistas* recognizes a rider as bicyclist, motorcyclist or other rider. In our experiments, we re-label the categories of bicyclist, motorcyclist and other rider of *Mapillary Vistas* into the label of rider in *Cityscapes*. Overall, the complete label transformations from *Mapillary Vistas* to *Cityscapes* used in our experiments are reported in Table. 1.

Table 1: Label transformations from *Mapillary Vistas* to *Cityscapes*.

ID_{MV}	Class	ID_{City}	Class
13, 24, 41	Road, Lane Marking - General, Manhole	0	Road
2, 15	Curb, Sidewalk	1	Sidewalk
17	Building	2	Building
6	Wall	3	Wall
3	Fence	4	Fence
45, 47	Pole, Utility Pole	5	Pole
48	Traffic Light	6	Traffic Light
50	Traffic Sign (Front)	7	Traffic Sign
30	Vegetation	8	Vegetation
29	Terrain	9	Terrain
27	Sky	10	Sky
19	Person	11	Person
20, 21, 22	Bicyclist, Motorcyclist, Other Rider	12	Rider
55	Car	13	Car
61	Truck	14	Truck
54	Bus	15	Bus
58	On Rails	16	Train
57	Motorcycle	17	Motorcycle
52	Bicycle	18	Bicycle
Others	Others	255	Ignored Label

Besides, Fig. 1 draws some examples of the ground truth and the outputs of Deeplab-v3+ [1] trained on merged labels, which is employed as one of the shadow models in our independent attacks.

Fig. 1: Visualization of *Mapillary Vistas* of raw images, original label images, transformed label images with Table 1 and predicted label images from the shadow model of Deeplab-v3+.

2 Visualization of segmentation

To compare the models with different defenses qualitatively, we show some segmentation examples of PSPNet [5] and UperNet [4] in Fig. 2 and Fig. 3 respectively. We highlight the details for some patches with rich boundaries. Comparing the models with Gaussian noises to the original model, we observe the boundaries become increasingly noisy with stronger noises. However, adding Gaussian noise will not change major parts, and it only decreases mIoU a little bit, as shown in Fig. 4 of our main paper. Therefore, adding Gaussian noises on posteriors is visually justified, in particular when we apply noises with a small variance, which still prevents information leakage.

Further, comparing the model with dropout during test to the original model, we can see that there are no corrupted boundaries with noisy-points, but very strong distortions and decreases performance significantly.

In the end, when we compare the model trained with DPSGD, we cannot observe any clear artifacts and we also maintain the segmentation performance.

References

1. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018) 1
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 1
3. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017) 1
4. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV (2018) 3
5. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017) 3

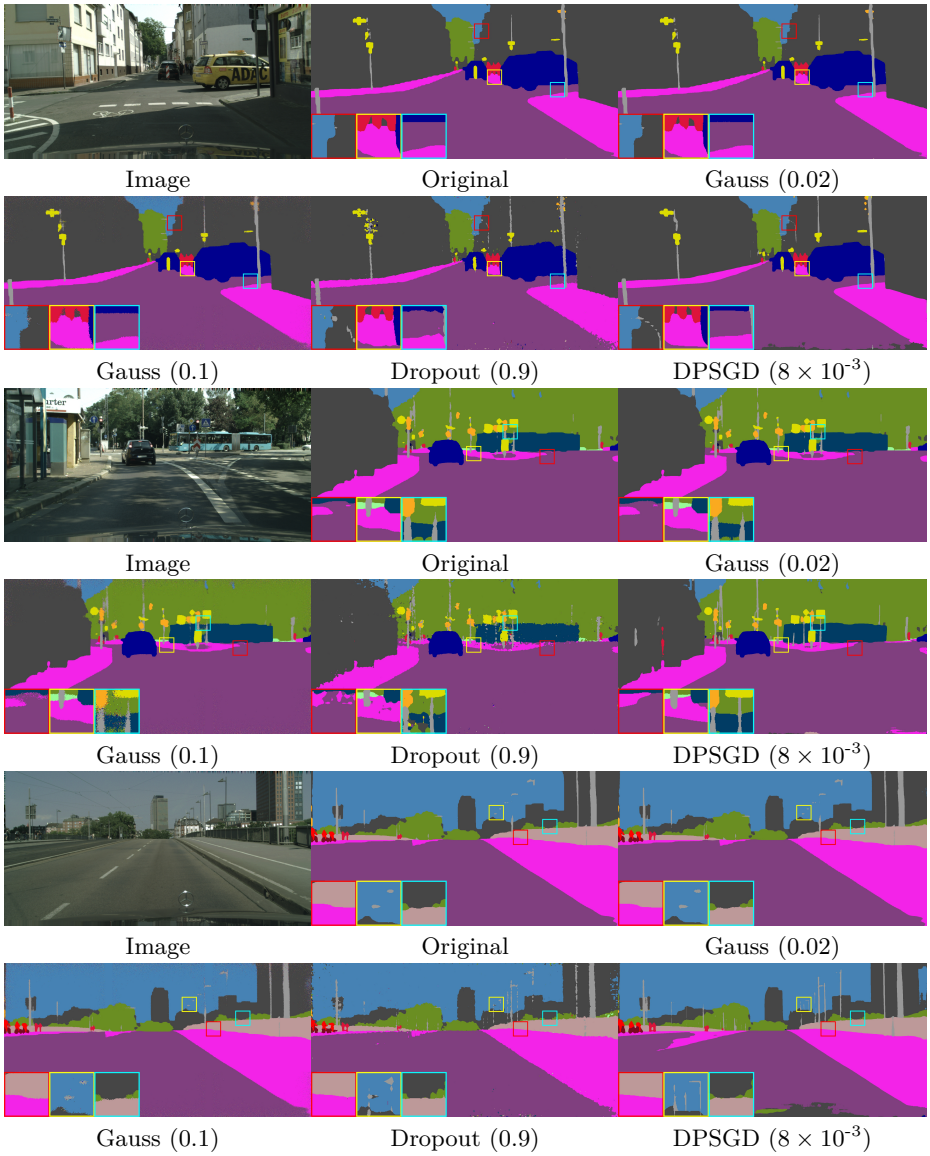

Fig. 2: Visualization results of PSPNet. We show the segmentation results with different defenses.

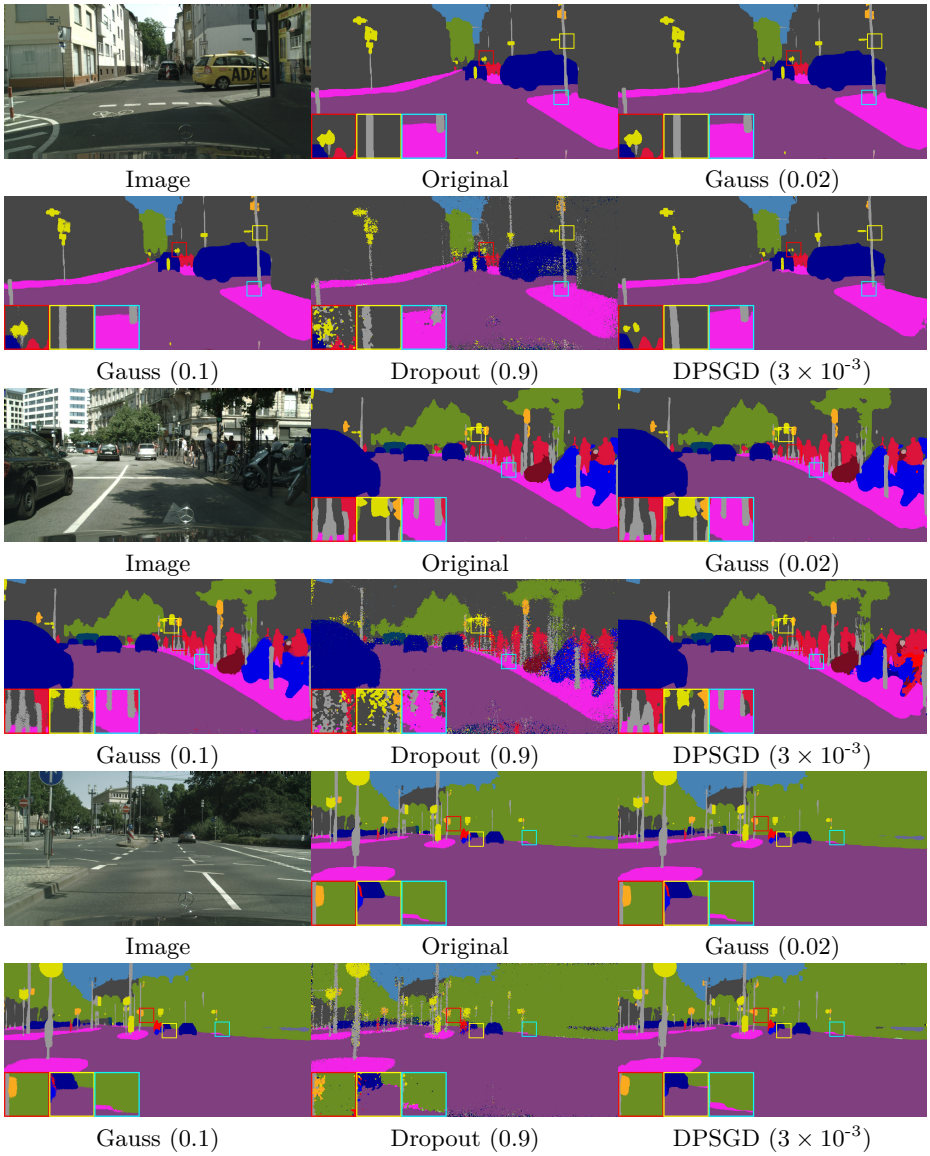

Fig. 3: Visualization results of UperNet. We show the segmentation results with different defenses.