GMMSeg: Gaussian Mixture Models for Deep Generative Semantic Segmentation
Supplemental Material

Chen Liang[1,3*†], **Wenguan Wang**[2*], **Jiaxu Miao**[1], **Yi Yang**[1]

[1]CCAI, Zhejiang University [2]ReLER, AAII, University of Technology Sydney [3]Baidu Research

https://github.com/leonnnop/GMMSeg

In this document, we provide the following items that shed deeper insight on our contributions:

- §S1: Detailed training parameters.
- §S2: More experimental results.
- §S3: More qualitative visualization.
- §S4: Discussion of legal/ethical considerations and limitations.

S1 Detailed Training Parameters

We evaluate our GMMSeg on six base segmentation architectures. Four of them, *i.e.*, DeepLab$_{\text{v3+}}$ [1], OCRNet [2], Swin-UperNet [3], SegFormer [4], are presented in our main paper. And the two additional base architectures, *i.e.*, FCN [5] and Mask2Former [6], are provided in this supplemental material (*cf.* §S2). We follow the default training settings in the official Mask2Former codebase and MMSegmentation for Mask2Former and other base architectures respectively. In particular, we train FCN, DeepLab$_{\text{v3+}}$ and OCRNet using SGD optimizer with initial learning rate 0.1, weight decay 4e-4 with polynomial learning rate annealing; we train Swin-UperNet and SegFormer using AdamW optimizer with initial learning rate 6e-5, weight decay 1e-2 with polynomial learning rate annealing; we train Mask2Former using AdamW optimizer with initial learning rate 1e-4, weight decay 5e-2 and the learning rate is decayed by a factor of 10 at 0.9 and 0.95 fractions of the total training steps.

S2 More Experimental Results

More Base Segmentation Architectures. We first demonstrate the efficacy of our GMMSeg on two additional base segmentation architectures, *i.e.*, FCN [5] and Mask2Former [6], with quantitative results summarized in Table S1. We train FCN based models with the according training hyperparameter settings mentioned in §S1 and strictly follow the same training and inference setups in our main manuscript (*cf.* §4.1). Furthermore, for a fair comparison with Mask2Former, the backbone, *i.e.*, Swin$_{\text{Large}}$ [3], is pretrained with ImageNet$_{\text{22K}}$ [10]. For ADE$_{\text{20K}}$/COCO-Stuff/Cityscapes, we train Mask2Former based models using images cropped to 640 × 640/640 × 640/1024 × 1024, for 160K/80K/90K iterations with 16/16/16 batch size. We adopt sliding window inference on Cityscapes with a window size of 1024 × 1024 and we keep the aspect ratio of test images and rescale the short side to 640 on ADE$_{\text{20K}}$ and COCO-Stuff.

Table S1: Additional quantitative results (§S2) on ADE$_{\text{20K}}$ [7] `val`, Cityscapes [8] `val`, and COCO-Stuff [9] `test` in mean IoU.

Method	Backbone	**ADE$_{\text{20K}}$**	**Citys.**	**COCO.**
FCN [CVPR15] [5]	ResNet$_{101}$	39.9	75.5	32.6
`GMMSeg`		**41.8** ↑1.9	**76.7** ↑1.2	**34.1** ↑1.5
Mask2Former [CVPR22] [6]	Swin$_{\text{Large}}$	56.1	83.3	51.0
`GMMSeg`		**56.7** ↑0.6	**83.8** ↑0.5	**52.0** ↑1.0

[*] Equal contributions.

[†] Work partly done during an internship at Baidu Research.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table S2: Quantitative results (§S2) on Fishyscapes (FS) Lost&Found `test` and Static `test`.

Method	Re-training	Extra Network	OoD Data	**FS Lost&Found**		**FS Static**	
				AP ↑	FPR_{95} ↓	AP ↑	FPR_{95} ↓
Density - Single-layer NLL [12]	✗	✓	✗	3.01	32.9	40.86	21.29
Density - Minimum NLL [12]	✗	✓	✗	4.25	47.15	62.14	17.43
Density - Logistic Regression [12]	✗	✓	✓	4.65	24.36	57.16	13.39
Image Resynthesis [15]	✗	✓	✗	5.70	48.05	29.6	27.13
Bayesian Deeplab [16]	✓	✗	✗	9.81	38.46	48.70	15.05
OoD Training - Void Class [17]	✓	✗	✓	10.29	22.11	45.00	19.40
Discriminative Outlier Detection Head [18]	✓	✓	✓	31.31	19.02	96.76	0.29
Dirichlet Deeplab [19]	✓	✗	✓	34.28	47.43	31.30	84.60
SynBoost [20]	✗	✓	✓	43.22	15.79	72.59	18.75
MSP [21]	✗	✗	✗	1.77	44.85	12.88	39.83
Entropy [22]	✗	✗	✗	2.93	44.83	15.41	39.75
kNN Embedding - density [12]	✗	✗	✗	3.55	30.02	44.03	20.25
SML [14]	✗	✗	✗	31.05	21.52	53.11	19.64
GMMSeg-$\text{DeepLab}_{\text{V3+}}$	✗	✗	✗	**55.63**	**6.61**	**76.02**	**15.96**

Here FCN is a famous fully convolutional model that is in line with the per-pixel dense classification models we discussed in the main paper (*cf*.§3.1). Besides, of particular interest is the Mask2Former, which is an attentive model proposed very recently that formulates the task as a mask classification problem, where a mask-level representation is learned instead of pixel-level. However, it still relies on a discriminative softmax based classifier for mask classification. We equip Mask2Former by replacing the softmax classification module with our generative GMM classifier.

As seen, GMMSeg consistently boosts the model performance despite different segmentation formulations, *i.e*., pixel classification or mask classification, verifying the superiority of our GMMSeg that brings a paradigm shift from a discriminative softmax to a generative GMM. Notably, with Mask2Former-$\text{Swin}_{\text{Large}}$ as base segmentation architecture, our GMMSeg earns mIoU scores of **56.7%/83.8%/52.0%**, establishing new state-of-the-arts among $\text{ADE}_{20\text{K}}$/Cityscapes/COCO-Stuff.

Anomaly Segmentation Result on Fishyscapes Lost&Found `test` and Static `test`. We additionally report the anomaly segmentation performance of our Cityscapes [8] trained GMMSeg built upon $\text{DeepLab}_{\text{V3+}}$ [1]-ResNet_{101} [11] on Fishyscapes [12] Lost&Found `test` and Static `test`. Fishyscapes Static is a blending-based dataset built upon backgrounds from Cityscapes and anomalous objects from Pascal VOC [13], that contains 30/1,000 images in `val`/`test` set. The `test` splits of Fishyscapes Lost&Found and Static are privately held by the Fishyscapes organization that contain entirely unknown anomalies to the methods. The results are summarized in Table S2, and are also publicly available in anonymous on the official leaderboard[3]. We categorize the methods by checking whether they require retraining, extra segmentation networks or utilize OoD data, following [12, 14].

As seen, without any add-on post-calibration technique, GMMSeg significantly surpasses the state-of-the-art methods by even larger margins on the challenging `test` set compared to results on `val` set, *i.e*., **+24.58%/+14.91%** in AP and **+22.91%/+3.68%** in FPR_{95} on Fishyscapes Lost&Found/Static `test`. Notably, GMMSeg even outperforms all other benchmark methods that employ additional training networks/data on Fishyscapes Lost&Found `test`, verifying the strong robustness to unexpected anomalies on-road due to the accurate data density modeling of GMMSeg.

Impact of Memory Capacity. In Table S3, we further explore the influence of the memory capacity, *i.e*., the amount of pixel representations stored for class-wise EM estimation, with $\text{DeepLab}_{\text{V3+}}$-$\text{ResNet}_{101}$ on $\text{ADE}_{20\text{K}}$ `val` trained for 80K iterations. For the first row, where the memory size is set to 0, the EM is only performed within mini-batches. Not surprisingly, data distribution estimated at such a local scale is far from accurate, leading to inferior results. With enlarged memory capacity, the performance is increased. When the performance reaches saturation, the stored pixel samples are sufficient enough to represent the true data distribution of the whole training set.

Table S3: Impact of memory size, evaluated on $\text{ADE}_{20\text{K}}$ [7] `val`.

# Sample	mIoU (%)
0	40.3
8K	45.1
16K	45.4
32K	46.0
48K	46.0

[3] https://fishyscapes.com/results

S3 More Qualitative Visualization

Semantic Segmentation. We illustrate the qualitative comparisons of GMMSeg equipped SegFormer [4]-MiT$_{\text{B5}}$ against the original model on ADE$_{\text{20K}}$ [7] (Fig. S1), Cityscapes [8] (Fig. S2) and COCO-Stuff [9] (Fig. S3). It is evident that, benefiting from the accurate data characterization modeling, GMMSeg is less confused by object categories and gives preciser predictions than SegFormer.

Anomaly Segmentation. We then show more qualitative results of MSP [22]-DeepLab$_{\text{V3+}}$ [1] and GMMSeg-DeepLab$_{\text{V3+}}$ on Fishyscapes Lost&Found `val`. As observed, different from MSP, GMMSeg gets rid of being overwhelmed by overconfident predictions and successfully identifies the anomalies.

S4 Discussion

Asset License and Consent. We use three semantic segmentation datasets, *i.e.*, ADE$_{\text{20K}}$ [7], Cityscapes [8], COCO-Stuff [9], and two anomaly segmentation datasets, *i.e.*, Fishyscapes [12], Road Anomaly [15], that are all publicly and freely available for academic purposes. We implement all models with MMSegmentation [23] and official Mask2Former [6] codebases. ADE$_{\text{20K}}$ (`https://groups.csail.mit.edu/vision/datasets/ADE20K/`) is released under a CC BSD-3; Cityscapes (`https://www.cityscapes-dataset.com/`) is released under this License; COCO-Stuff v1.1 (`https://github.com/nightrome/cocostuff`) is released under Flickr Terms of use for images and CC BY 4.0 for annotations; Road Anomaly (`https://www.epfl.ch/labs/cvlab/data/road-anomaly/`) is released under CC BY 4.0; All assets mentioned above release annotations obtained from human experts with agreements. Fishyscapes (`https://fishyscapes.com/`) is released under CC BY 4.0. This dataset is synthesized and re-organized from existing datasets that we are not capable to trace every detail; MMSegmentation codebase (`https://github.com/open-mmlab/mmsegmentation`) is released under Apache-2.0 license. Mask2Former codebase (`https://github.com/facebookresearch/Mask2Former`) is released under MIT license.

Limitation Analysis. One limitation of our approach is that the EM based generative parameter estimation needs extra optimization loops in each training iteration which would reduce the training efficiency in terms of time complexity. However, in practice, we find one EM loop per training iteration is good enough for global model convergence, which only brings a minor computational overhead, *i.e.*, $\sim$5% training speed delay. We will dedicate to designing more powerful algorithms with further improvements in both efficiency and efficacy.

Broader Impact. This work introduces the first generative semantic segmentation framework that shows promising results in both closed-set and open-world scenarios. On positive side, the approach advances model accuracy of semantic segmentation and can certainly have a wide range of real-world applications, *e.g.*, precision agriculture, robot navigation, *etc*. The strong robustness to anomalous objects further warrants its potential for usage in safety-critical applications, *i.e.*, autonomous driving. On negative side, the generated results can be fed into other algorithms for malicious purposes, *e.g.*, identifying the minority groups. Though beyond the scope of this paper, we will organize a gated release of our models to make sure that they are not being used beyond academic research purposes.

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) S1, S2, S3, S8

[2] Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: ECCV. (2020) S1

[3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. (2021) S1

[4] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS. (2021) S1, S3, S5, S6, S7

[5] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) S1

[6] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. CVPR (2022) S1, S3

[7] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. (2017) S1, S2, S3, S5

[8] 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) S1, S2, S3, S6

[9] Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR. (2018) S1, S3, S7

[10] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. IJCV **115**(3) (2015) 211–252 S1

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) S2

[12] Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: The fishyscapes benchmark: Measuring blind spots in semantic segmentation. IJCV (2021) S2, S3, S8

[13] Everingham, M., Eslami, S., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV (2015) S2

[14] Jung, S., Lee, J., Gwak, D., Choi, S., Choo, J.: Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In: ICCV. (2021) S2

[15] Lis, K., Nakka, K., Fua, P., Salzmann, M.: Detecting the unexpected via image resynthesis. In: ICCV. (2019) S2, S3

[16] Mukhoti, J., Gal, Y.: Evaluating bayesian deep learning methods for semantic segmentation. arXiv preprint arXiv:1811.12709 (2018) S2

[17] DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865 (2018) S2

[18] Bevandić, P., Krešo, I., Oršić, M., Šegvić, S.: Simultaneous semantic segmentation and outlier detection in presence of domain shift. In: GCPR. (2019) S2

[19] Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. NeurIPS (2018) S2

[20] Di Biase, G., Blum, H., Siegwart, R., Cadena, C.: Pixel-wise anomaly detection in complex driving scenes. In: CVPR. (2021) S2

[21] Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. arXiv preprint arXiv:1911.11132 (2019) S2

[22] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR. (2017) S2, S3, S8

[23] Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020) S3

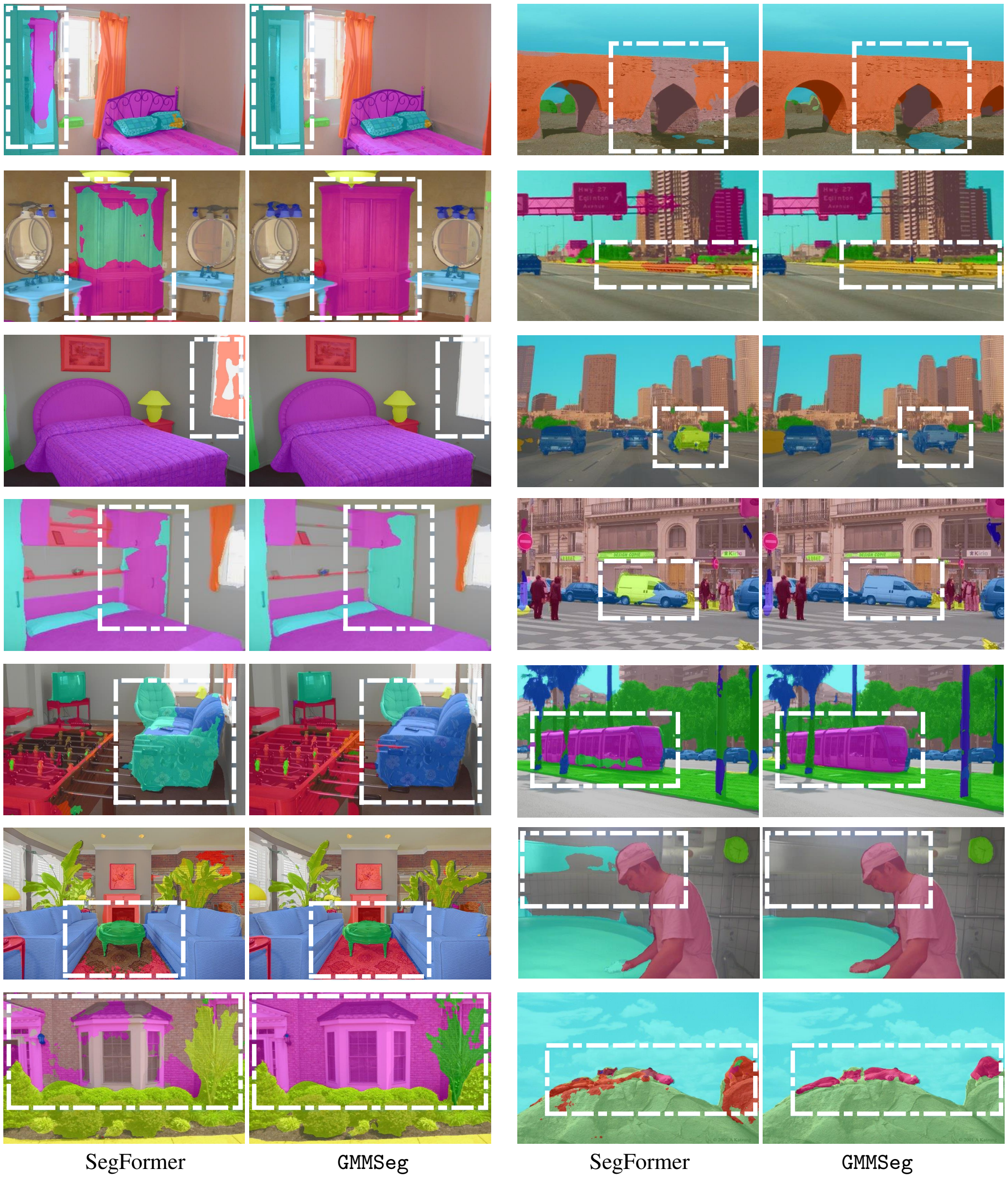


Figure S1: Qualitative results (§S3) of SegFormer [4] and our GMMSeg on ADE$_{20K}$ [7].

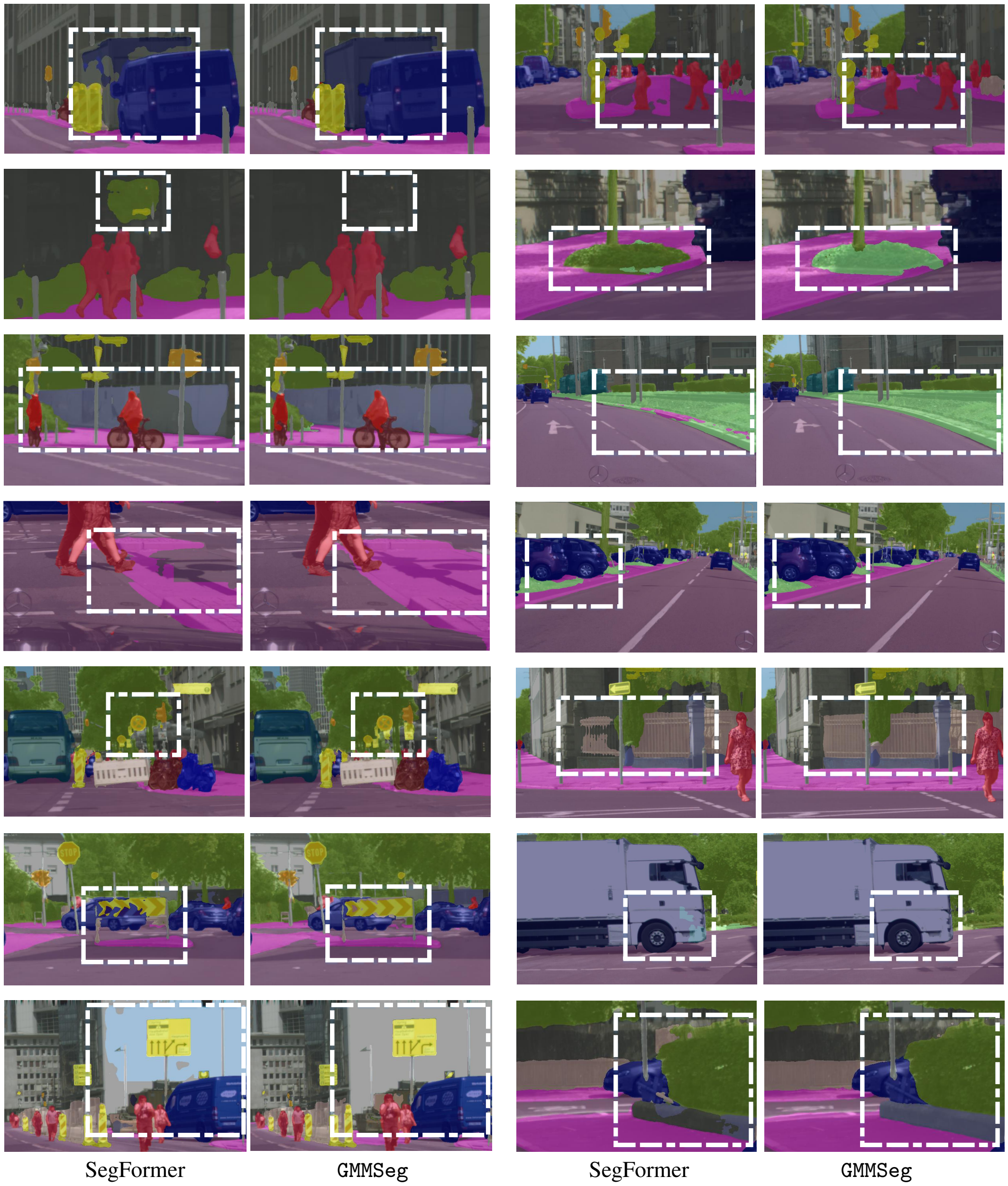
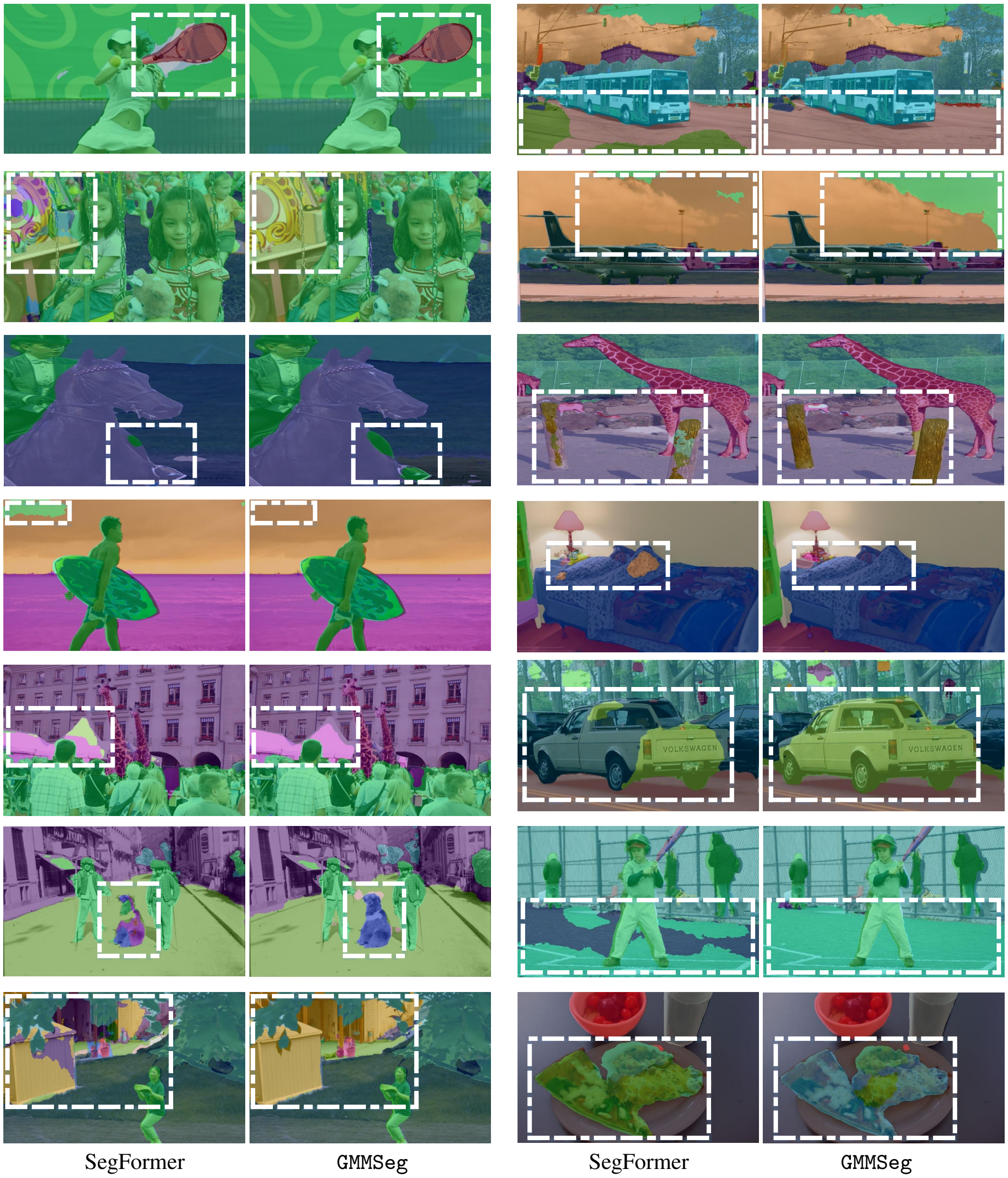


Figure S2: Qualitative results (§S3) of SegFormer [4] and our GMMSeg on Cityscapes [8].



SegFormer GMMSeg SegFormer GMMSeg

Figure S3: Qualitative results (§S3) of SegFormer [4] and our GMMSeg on COCO-Stuff [9].

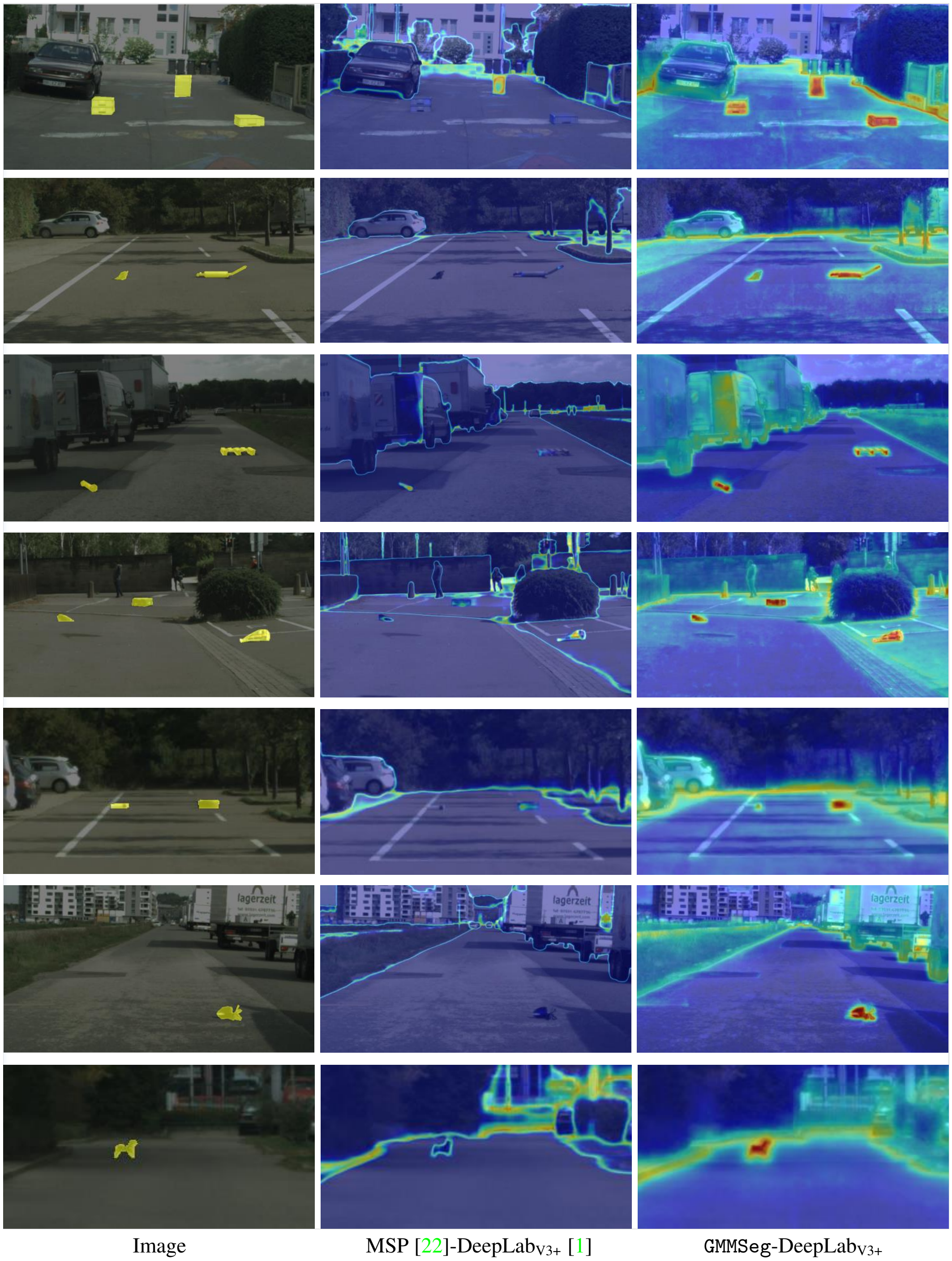


Image MSP [22]-DeepLab$_{V3+}$ [1] GMMSeg-DeepLab$_{V3+}$

Figure S4: Qualitative results (§S3) of anomaly heatmaps on Fishyscapes Lost&Found [12] `val`.