Concurrent Misclassification and Out-of-Distribution Detection for Semantic Segmentation via Energy-Based Normalizing Flow (Supplementary material)

Denis Gudovskiy[1] **Tomoyuki Okuno**[2] **Yohei Nakata**[2]

[1]Panasonic AI Lab, Mountain View, CA, USA
[2]Panasonic Holdings Corporation, Osaka, Japan

1 IMPLEMENTATION DETAILS

Initialization. Convolutional parameters in the FED network $g(\boldsymbol{\theta})$ are initialized using the default scheme in PyTorch. ActNorm and iMap are reimplemented and initialized according to [Kingma and Dhariwal, 2018, Sukthanker et al., 2022] references. Distributional parameters in $g(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{U})$ are initialized with zero values. A subset of them ($\boldsymbol{\beta}$ and diag$(\boldsymbol{U})$) are passed through a SoftPlus activation, which results in a strictly non-negative values.

Training. FED training phase takes only few GPU-hours and has the following hyperparameters: AdamW optimizer with initial 1e-3 learning rate, which is reduced by a factor of 10 every 15,000 iterations. We use in total 50,000 iterations and a mini-batch size of 4. In addition, a warm-up phase with the learning rate gradually increasing from 1e-6 to 1e-3 is applied during first 4,000 iterations. We select the highest learning rate from the {1e-2, 1e-3, 1e-4} range using ablation study. Practically, the number of training iterations can be substantially decreased (e.g. to 20,000 iterations) without a significant drop in IDM/OOD metrics. We use the default image crop sizes during training: 512×1024 for DL-R101 and 1024×1024 for SF-B2 backbone.

Inference. Inference is done on full-size images without cropping for DL-R101 task backbone. We use the reference implementation for SF-B2 backbone, where 1024×1024 cropping with sliding is accomplished at test-time. Next, we discuss details about used test-time augmentation (TTA). TTA is a common technique to improve inference results for segmentation models and is available out-of-the-box in MMSegmentation library. In our case, we use TTA for input image resizing and averaging output scores without any other augmentations. We optionally apply TTA to FlowEneDet in order to increase IDM/OOD metrics at the expense of lower inference speed as reported in Section 5.4. During the training phase TTA doesn't require any modification: FED is trained by input/output tensors with $1/4\times$ spatial dimensions of image size. In other words, the $1/4\times$ rate is identical to the task's classifier resolution during training and inference without TTA. In case of the enabled TTA, inputs images are resized to have [$1/4\times$, $1/2\times$, $1\times$] resolution, while FED input/output tensors are internally upsampled by a factor of $4\times$ from the original $1/4\times$ resolution i.e. FED rates become [$1/4\times$, $1/2\times$, $1\times$] as well. Effectively, segmentation backbone processes images with the original or downsampled resolution, while FED operates at the original or upsampled resolution w.r.t. the training phase. This technique helps us to capture small- and large-scale OOD objects. A more compute-efficient approach is to train a set of multi-scale FED detectors with aggregation at the expense of marginally higher memory footprint.

2 EXTENDED ABLATION STUDY AND DISCUSSION ON LIMITATIONS

Table 8 presents an ablation study of various architectural tradeoffs for FED detector with SF-B2 backbone. We choose a more robust SF-B2 here instead of DL-R101 backbone because the latter shows similar trends on average, but has significantly higher metric's variances. Specifically, we evaluate: unconditional FED-U and conditional FED-C, full or diagonal covariance matrix $\boldsymbol{U}$, kernel size K (3×3, 7×7 or 11×11) for the flow's Conv2D layer that defines spatial receptive field, number of coupling blocks L (4 or 8), and the length P of condition vector $\boldsymbol{a}$ (32 or 128).

Note that the open-mIoU evaluation in Table 8 is different for the configuration with TTA and without TTA. The configurations without TTA are implemented exactly as described in Section 5.1 with the closed-set mIoU of 81.1%. However, IDM detection is not feasible for the multi-scale processing scheme described in Appendix 1, where the backbone and FED network are trained by inputs with a certain resolution scheme ($1\times$ and $1/4\times$, respectively), but tested with another resolution setup [$1/4\times$, $1/2\times$, $1\times$] both for backbone and FED network). Therefore, we derive a modified multi-scale scheme from the reference scheme for SegFormer TTA in MMSegmentation. During inference with the enabled TTA

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Table 8: Ablation study of architectural choices for FED SF-B2 variants when applied to OOD detection on FS L&F and Static **validation split** and IDM/OOD detection using CS **validation split**, %. The **best** and the second best results are highlighted. Design space is defined as follows: covariance matrix $\boldsymbol{U}$ is full or diagonal, kernel size K for the flow's Conv2D layer is 3×3, 7×7 or 11×11, number of coupling blocks L is 4 or 8, the size P of condition vector $\boldsymbol{a}$ is 32 or 128. Our default configuration: full-covariance $\boldsymbol{U}$, $K = 7 \times 7$, $L = 8$, and $P = 32$ for FED-C or $P = 0$ for FED-U.

Method	$\boldsymbol{U}$	K	L	P	FS L&F		FS Static		CS
					AP↑	FPR_{95} ↓	AP↑	FPR_{95} ↓	open-mIoU↑
FED-U	full	7×7	8	-	39.90	18.66	55.93	17.15	81.43
FED-C	full	7×7	8	32	41.15	11.10	47.56	37.53	77.61
FED-U (TTA)	full	7×7	8	-	41.75	10.05	66.60	8.94	81.77
FED-C (TTA)	full	7×7	8	32	56.11	**3.87**	52.61	14.91	79.40
FED-U (TTA)	full	3×3	8	-	42.28	9.94	65.98	9.09	81.13
FED-C (TTA)	full	3×3	8	32	51.98	6.88	53.98	13.69	79.14
FED-U (TTA)	full	11×11	8	-	40.36	9.98	**66.80**	**8.93**	82.66
FED-C (TTA)	full	11×11	8	32	**56.84**	4.19	51.47	16.93	76.34
FED-U (TTA)	diag	7×7	8	-	41.71	9.99	66.21	9.09	81.98
FED-C (TTA)	diag	7×7	8	32	51.62	4.04	55.66	13.15	81.13
FED-U (TTA)	full	7×7	4	-	41.57	9.92	66.21	9.15	82.00
FED-C (TTA)	full	7×7	4	32	49.54	4.63	50.65	15.89	71.86
FED-C (TTA)	full	7×7	8	128	26.00	17.22	32.57	22.24	**86.59**

for open-mIoU evaluation in Table 8, the backbone input rate ([$1/2\times$, $1\times$, $3/2\times$]) is consistent with the FED input rate [$1/8\times$, $1/4\times$, $3/8\times$]. Hence, we preserve the same $1/4\times$ rate for the FED network during train and inference phases to successfully detect misclassifications. This TTA scheme increases closed-set mIoU from 81.1% to 81.75%. For reference, we report modified OOD scores for this TTA scheme on FS validation dataset using [AuROC, AP, FPR_{95}] format:

- FED-U L&F: [97.83→98.51, 41.75→49.03, 10.05→7.66]
- FED-C L&F: [99.11→99.27, 56.11→52.92, 3.87→2.95]
- FED-U Stat: [98.30→97.80, 66.60→66.53, 8.94→10.31]
- FED-C Stat: [96.88→95.51, 52.61→52.78, 14.91→25.63]

In our ablation study in Table 8, we verify that the full covariance matrix $\boldsymbol{U} \in \mathbb{R}^{2\times2}$ outperforms the univariate $[\text{diag}(\boldsymbol{U})] \in \mathbb{R}^2$ approach in most cases. Similarly, the higher number of coupling blocks L results in better metrics. A 11×11 kernel size with larger receptive field is superior than our default 7×7 Conv2D layer in most cases. So, our default choice is suboptimal in the sense of performance metrics, but better in terms of inference speed and memory footprint. A transformer architecture with the global attention for the flow network can be an interesting future direction [Sukthanker et al., 2022] to resolve a problem with the limited receptive field in convolutional layers.

The length P of the condition vector $\boldsymbol{a}^P$ in the current FED-C plays an ambivalent role. The larger ($P = 128$) produces an excellent CS open-mIoU (86.59%) compared to the configuration with $P = 32$ (79.4%), but significantly underperforms in FS benchmark (17.22% FPR_{95} vs. 3.87% FPR_{95} for FS L&F). At the same time, the unconditional FED-U (i.e. $P = 0$) outperforms FED-C with $P = 32$ in FS Static and CS open-mIoU. Therefore, we observe that the most simplistic compute-free average pooling technique in FED-C model achieves state-of-the-art results in FS L&F and SMIYC, but underperforms in FS Static and CS's open-mIoU due to, possibly, two different reasons. We hypothesize that a larger P improves in-domain density estimation because latent-space embeddings contain more information about feature distribution, which is reflected in the excellent CS open-mIoU metric. At the same time, out-of-domain data can have a significant distributional shift. It seems to be the case in FS Static split, where FED-C underperforms compared to the embedding-unconditional FED-U model. Therefore, we conclude that FED-C approach is beneficial in general in comparison to FED-U. However, its current major limitation is in the feature pooling mechanism. We believe, FED-C results can be further improved and be more consistent across multiple datasets, if the pooled condition vector $\boldsymbol{a}$ satisfies the following: a) contains sufficient latent-space information for in-domain density estimation, and b) represents features that are robust to distributional shifts. We hope these observations will inspire follow-up research.

3 EXTRA QUALITATIVE RESULTS

Figure 4 shows additional qualitative results for our most low-complexity FED-U configuration with DL-R101 as well as MCD and SML. We plot confidence scores with a normalization to [0:1] range, where red (0) and blue (1) represent the most uncertain and certain areas, respectively. Normalization statistics are derived for each dataset before plotting detection predictions.

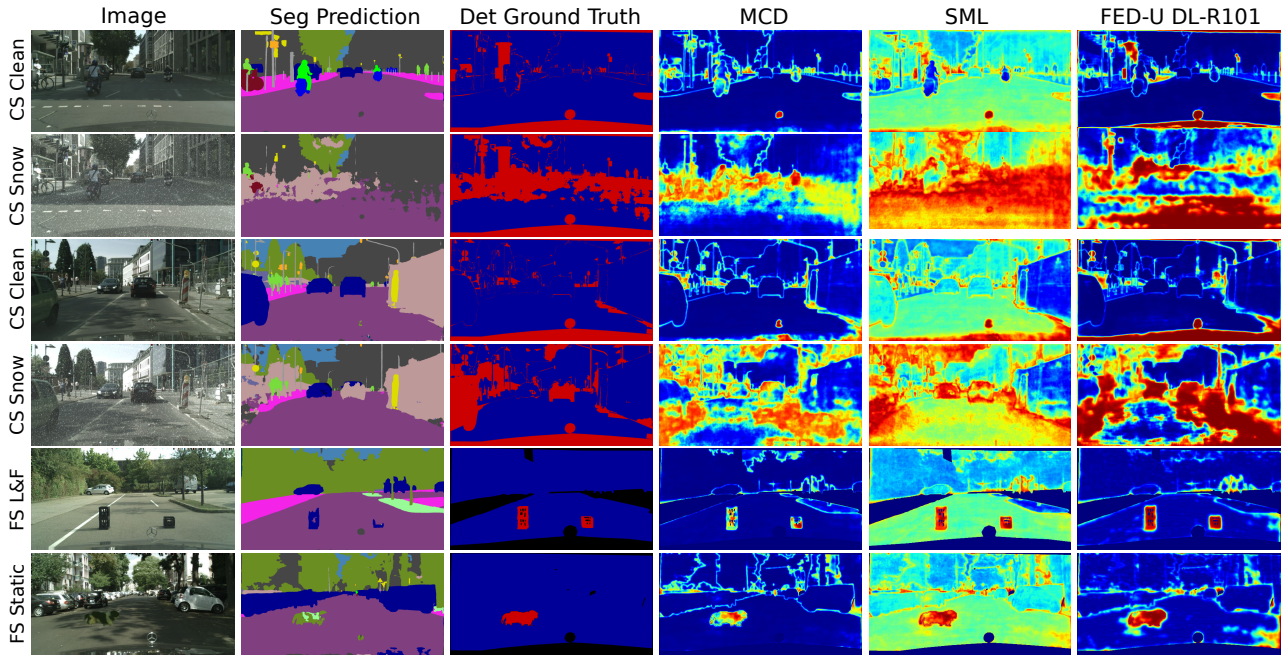

Figure 4: This figure shows from left to right: input image, DL-R101 segmentation prediction, IDM/OOD detection ground truth, and detection predictions for MCD [Mukhoti and Gal, 2018], SML [Jung et al., 2021] and our FED-U detector. Each input image example is from the corresponding validation dataset, specifically, from top to bottom: two Cityscapes (CS) images and the same images corrupted by the snow corruption from Cityscapes-C, an image from the Fishyscapes (FS) L&F and Static validation splits. Detector's task is to predict IDM/OOD pixels as red scores and correctly classified pixels as blue scores. Black area represents an ignored void class in FS datasets. Compared to other detectors, our FED-U separates IDM/OOD pixels more accurately. At the same time, IDM/OOD detection is quite challenging for heavily corrupted environment such as the snowy weather when the predicted segmentation becomes very imprecise.

We select two examples from the uncorrupted CS, and the corresponding CS-C validation dataset with the lowest severity snow corruption. The second column shows segmentation model predictions, and the third column highlights its correctly classified pixels (blue), the union of IDM and OOD pixel masks (red) i.e. the detection ground truth. Last two rows show images from FS L&F and Static validation datasets. Unlike CS, FS ground truth contains only OOD pixels (red), normal objects (blue), and the ignored during evaluation void class (black).

Our detector visually better matches detection ground truth masks. Notably, SML fails in assigning high confidence scores for in-domain positives (yellow and green instead of blue), and MCD is not consistent when assigning low confidence scores for OOD areas (green and blue instead of red). Finally, we emphasize that weather corruptions e.g. snow can pose a considerable difficulty for semantic segmentation performance as well as IDM/OOD detection. Certainly, decision-critical applications have to avoid operating in such extreme environment as soon as detector signals about broadly low-confident segmentation predictions.

References

Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized Max Logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *ICCV*, 2021.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

Jishnu Mukhoti and Yarin Gal. Evaluating Bayesian deep learning methods for semantic segmentation. *arXiv:1811.12709*, 2018.

Rhea Sanjay Sukthanker, Zhiwu Huang, Suryansh Kumar, Radu Timofte, and Luc Van Gool. Generative flows with invertible attentions. In *CVPR*, 2022.