Real-Time Anomaly Segmentation for Road Scenes

Federico Bussolino
Politecnico di Torino
s317641@studenti.polito.it

Muhammad Sarib Khan
Politecnico di Torino
s298885@studenti.polito.it

Francine Tatiana Olanga Ombala
Politecnico di Torino
s319110@studenti.polito.it

Abstract

Semantic segmentation has become one of the most interesting task in computer vision. Its robustness becomes even more critical in computer vision challenges that fall under the category of open-world scenario like autonomous driving. Currently existing deep neural networks have a suboptimal performance on out-of-distribution objects which are not observed in the training process. In this paper we compare different ood-detection methods applied to semantic segmentation task. As baseline we will use classical scoring system such as MSP, MaxLogit and Entropy and temperature scaling. We then test ENet, BiSeNet and ERFNet using their background class as an anomaly score. Experimental results proved the important correlation between mIoU of the network and its capability to correctly predict an anomaly. Architecture like BiSeNet, that combine in output layers at multiple scale showed also better generalization performance on object of different size. Finally usage of on purpose loss functions for anomaly detection like logit normalization loss and enhanced isotropic maximization loss allow a great boost in performances.

1. Introduction

In recent years the interest towards the application of computer vision and machine learning to autonomous driving has grown significantly in scientific community. One crucial task of computer vision, particularly when applied to autonomous driving, is to determine which object is present in every part of the image. This can be achieved in 3 ways:

- **Object detection**: Regress bounding box for singular objects in the image and assign them a class prediction;
- **Semantic segmentation**: Make prediction at a pixel-level for each object class in the image;
- **Instance segmentation**: Similar to semantic segmentation but different object of the same class must be classified as separate entities.

First semantic segmentation architectures were based on an encoder-decoder structure. First architectures of this type have been developed since 2016 like ENet [13] and UNet [16], more recent works like ERFNet [15] improved performances thanks to various architectural ideas such as factorized residual blocks and dilated convolution, that lead to computationally lighter blocks preserving accuracy and this allowed to build deeper network. Later on, in 2018 BiseNet [18] achieved optimal results both in term of accuracy and speed, due to the fact that the encoder alone is able to produce high-quality segmentation map at 1/8 of the original image resolution and decoder is replaced by a simple upscale operation. In 2020, with the advent of transformer in the field of computer vision [4], semantic segmentation also benefited from this technology. Despite their accuracy transformer architectures tend to be slower and computationally expensive, that's why we focused on lighter methods in our analysis.

In this paper we will address **anomaly detection**: a complementary task to semantic segmentation that aims to solve the reliability problem when we encounter out of distribution data. Particularly, we are interested in detecting if pixels contain objects that are present in training dataset or not, starting from the output of a semantic segmentation network.

We will evaluate various anomaly detection methods by training deep neural network on cityscapes dataset [3] and measure anomaly detection performances on same dataset as in [14] that are: RoadAnomaly21, RoadObstacle21 [2], fishyscape static, fishyscape Lost and Found [1] and RoadAnomaly [9].

This test datasets differ by:

- Image size and proportion;

- Synthetic image patch vs. real object;
- Position of obstacles (only on road vs. anywhere);
- Size of anomalous object (rarity of anomaly).

Thanks to this differences this is a comprehensive benchmark for anomaly segmentation on the road scenario. The code implementation is available at https://github.com/Busso00/AML_ood_sem_segm

2. Related works

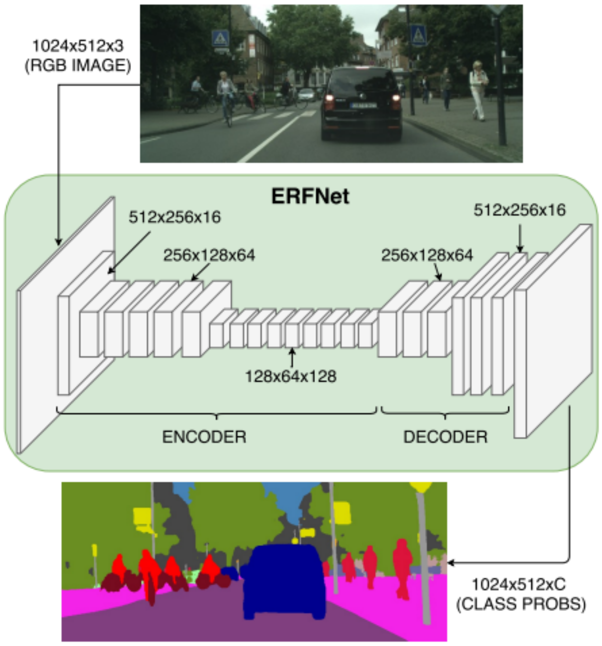


Figure 1. Erfnet structure: a classical encoder-decoder architecture.

First network based on encoder-decoder structure have proved their effectiveness in semantic segmentation since 2016 with ENet [13]. One important turning point for the accuracy of the segmentation networks was the introduction of more robust backbone, which include residual connection and batch normalization, in this sense Resnet-18 [6] is present both in ERFNet [15] and in BiSeNet [18]. To address the task of anomaly segmentation we first encompass baseline method like MSP [8], maxLogit [7], maxEntropy [10], we treat then post-processing score calibration methods like MSP with temperature scaling [5] as well as void classifiers trained on the previous cited segmentation architectures [13] [15] [18] and finally on-purpose loss function for anomaly detection like logit normalization loss [17] and enhanced isotropy maximization loss [12].

3. Scoring Methodology

3.1. Maximum Softmax Probability (MSP)

MSP consist of taking maximum softmax probability as anomaly metric: if low the pixel will be labeled as anomalous. The downside of this method are multiple:

- Overconfidence of softmax when classifying anomalous data can lead to classify data that deviate slightly from the normal data as important anomaly;
- Maximum softmax score is prone to degradating when we have lot of classes;
- The score depend solely on one class probability.

3.2. MaxLogit

MaxLogit solve the first two problems of MSP. It is also more computationally efficient and in our test obtained best result among baseline method.

3.3. Entropy

Using entropy as anomaly score solve in particular last two problems of MSP. It is probably the best choice when network output doesn't tend to be very confident, or when lot of classes are present.

$$E(f(x)) = -\sum_{j\in C} f_j(x)\log(f_j(x)). \tag{1}$$

3.4. MSP with temperature scaling

Address first problem of MSP, but more straightforward method like MaxLogit proved to work better in our benchmark. Temperature scaling consist of dividing the logit by $T > 0$ constant before applying MSP. However this showed not to be the best score calculation method since it address only overconfidence problem of MSP.

3.5. Void classifiers

Void classifiers represent an alternative approach to anomaly segmentation, relying heavily on the diversity of the background class in the training data. In these models, the anomaly score is determined by the softmax output corresponding to the background class. This method has shown to be particularly effective, especially when the output is derived from feature aggregation at multiple scales, as we can see in architectures like BiSeNet.

3.6. Enhanced isotropy maximization loss

Isotropy maximization loss [11] present several features that improve ood-detection performance:

- Its objective is to obtain output logits that have equal impact on class prediction in order to mitigate overconfidence of softmax scores. To obtain this it substitute the original logits with their L2 norm distance learnable set of parameters like follows:

$$\mathbf{f}_\theta(\mathbf{x}) - - > \|\mathbf{f}_\theta(\mathbf{x}) - \mathbf{p}_\phi\| \tag{2}$$

This allow also more flexible decision boundaries;

- To increase entropy of logits the entropic scale is applied only at train time so we can regularize magnitude of outputs at test time;
- In enhanced version of isomax loss [12] distances are normalized (so we calculate the cosine similarity): in this way we further regularize the newly obtained scores from 2. Moreover a learnable distance scale d_s is added for the same purpose;

The loss expression obtained is then:

$$\mathcal{L}_{\text{IsoMax+}} = -\log\left(\frac{\exp\left(-E_s|d_s|\left\|\widehat{\mathbf{f}_\theta}(\mathbf{x}) - \widehat{\mathbf{p}_\phi}^{\mathbf{k}}\right\|\right)}{\sum_j \exp\left(-E_s|d_s|\left\|\widehat{\mathbf{f}_\theta}(\mathbf{x}) - \widehat{\mathbf{p}_\phi}^{\mathbf{j}}\right\|\right)}\right) \tag{3}$$

Then output of the network at test time will be:

$$-|d_s|\left\|\tilde{\mathbf{f}}_\theta(\mathbf{x}) - \tilde{\mathbf{p}}_{\mathbf{s}}^{\mathbf{k}}\right\| \tag{4}$$

A downside of this method is the important impact on latency of the network since it requires to calculate distance from a prototype of the same size of output logits.

3.7. Logit normalization

This training methodology consist in applying an on purpose loss for anomaly detection [17]:

$$\mathcal{L}_{\text{logit\_norm}}(f(\mathbf{x};\theta), y) = -\log\left(\frac{e^{f_y/(\tau\|f\|)}}{\sum_{i=1}^{k} e^{f_i/(\tau\|f\|)}}\right) \tag{5}$$

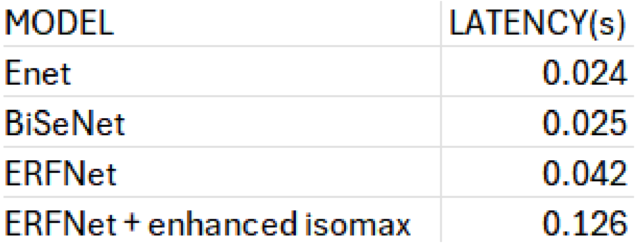
where y is ground truth class, k are the number of classes, so with $\tau = 1$ we normalize logit before applying a cross-entropy loss. This allow us to optimize only the direction of logits avoiding the overgrowth (or excessive reduction) of their magnitude, addressing the first two downside of MSP. When applying this technique 2 considerations turn out to be important:

- May be beneficial to reduce the impact of other regularization technique like weight decay, since logit normalization already reduce model expressiveness;
- This loss alone is not very suitable to imbalanced tasks, so we added weighted the cross-entropy loss applied after logit normalization;
- As pointed out in [17] chosing a proper scale factor for logit normalization is also important.

4. Training and evaluation methodology

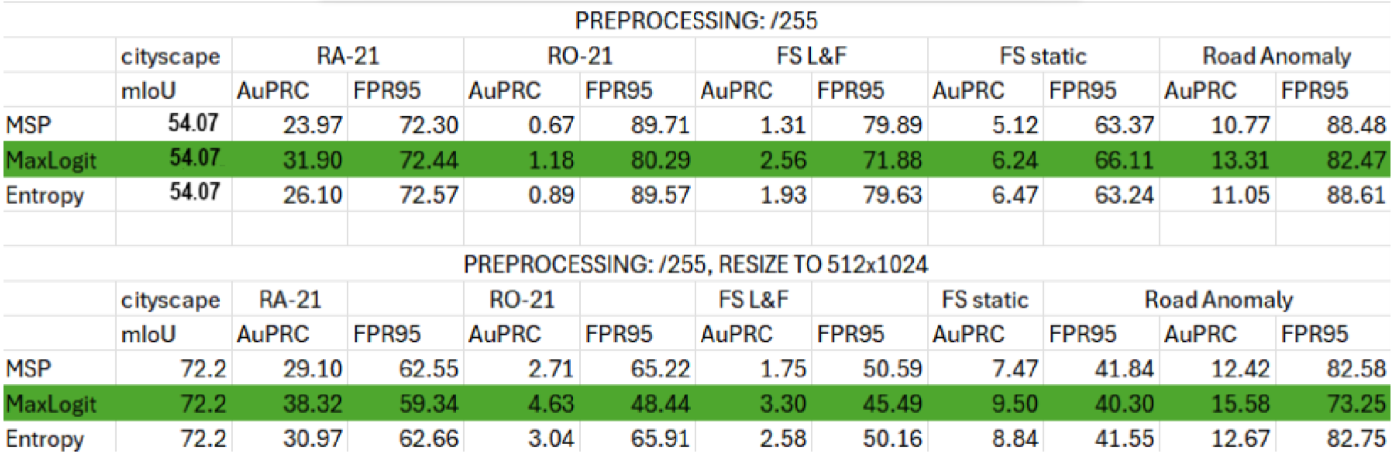
All networks are trained on cityscapes train split, mIoU is measured on cityscapes validation split taking into account 19 classes and excluding background.
Since we noticed that the network proposed obtained better mIoU values at the scale of 512x1024, probably due to their receptive field size, we decided to compare also result obtained by reshaping the image to this size, moreover when testing BiSeNet some reshape were necessary to match a proportion that the network can accept.
Training of BiSeNet deserves particular attention: in our case we trained with a multi-scale augmentation which allow to achieve higher mIoU on cityscapes even without resize.
Training with logit normalization loss alone required some adjustments: we added a weight to this loss and reduced weight decay to achieve a higher mIoU. However, we noticed that this was not sufficient for the training to converge to the original mIoU. In fact, only by incorporating CrossEntropy and Focal loss were we able to achieve a good mIoU.
ERFNet with enhanced isotropy maximization loss has been trained for only 40 epochs with batch size of 4. It was not trained with same procedure as other methods (that were trained with larger batch size and for more epochs) due to limited computational resources.
The joint training of ERFNet with enhanced isomax loss, Cross Entropy and Focal loss didn't converged as expected, this could be due to absence of weights in Focal loss, another reason behind that behaviour is that Cross Entropy and Focal loss "pull" the model to another direction that not necessarily maximize isotropy.

5. Experimental results



MODEL	LATENCY(s)
Enet	0.024
BiSeNet	0.025
ERFNet	0.042
ERFNet + enhanced isomax	0.126

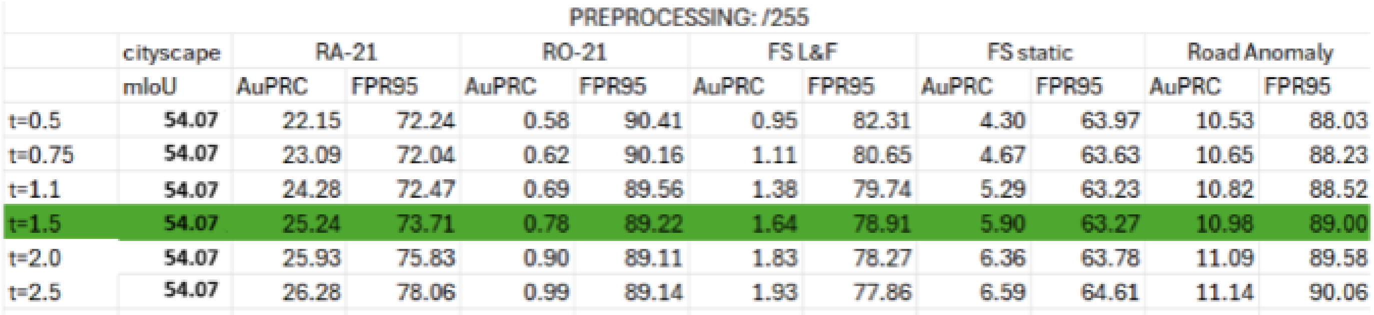
Figure 2. Since algorithms for autonomous driving are required to run in real-time we report latency of the various methods. The latency is measured on T4 GPU in Colab environment with Pytorch 2.4.0 and CUDA library 12.1. We advocate that latency differencies between MSP, maxLogit, entropy and temperature scaling can be irrelevant.



PREPROCESSING: /255										
	cityscape	RA-21		RO-21		FS L&F		FS static		Road Anomaly
	mIoU	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
MSP	54.07	23.97	72.30	0.67	89.71	1.31	79.89	5.12	63.37	10.77	88.48
MaxLogit	54.07	31.90	72.44	1.18	80.29	2.56	71.88	6.24	66.11	13.31	82.47
Entropy	54.07	26.10	72.57	0.89	89.57	1.93	79.63	6.47	63.24	11.05	88.61

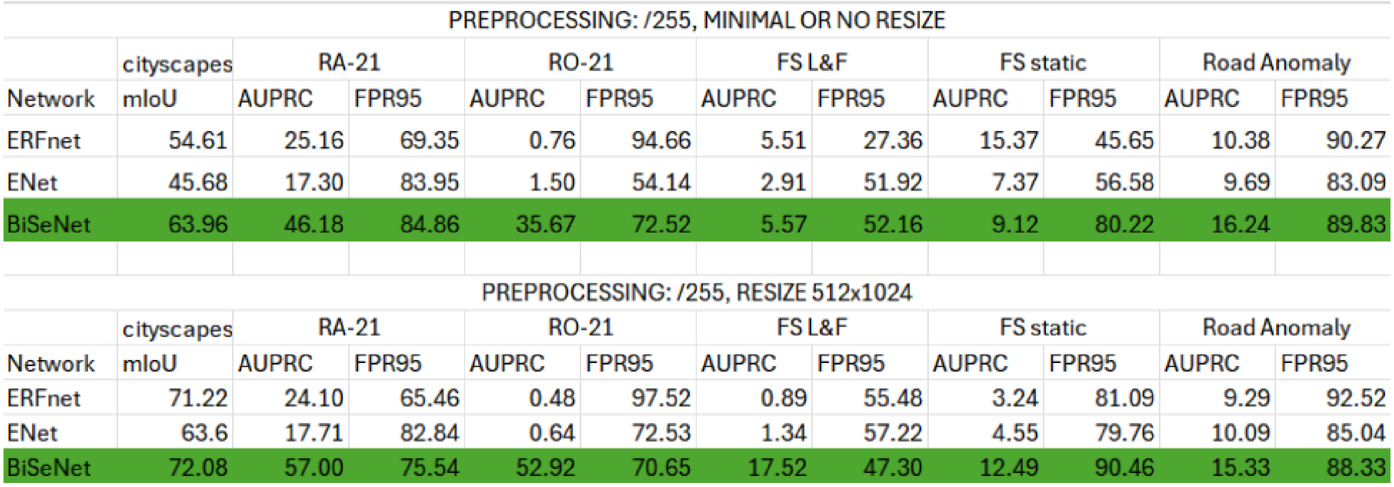
PREPROCESSING: /255, RESIZE TO 512x1024											
	cityscape	RA-21		RO-21		FS L&F		FS static		Road Anomaly	
	mIoU	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
MSP	72.2	29.10	62.55	2.71	65.22	1.75	50.59	7.47	41.84	12.42	82.58
MaxLogit	72.2	38.32	59.34	4.63	48.44	3.30	45.49	9.50	40.30	15.58	73.25
Entropy	72.2	30.97	62.66	3.04	65.91	2.58	50.16	8.84	41.55	12.67	82.75

Figure 3. Baseline experiments: show the effect of maximum softmax probability (MSP), max logit and entropy as anomaly score with different preprocessing steps. Results shows that mIoU can be largely impacted by the size of the image and this has also consequence on anomaly detection metrics. Between all methodologies applied to calculate the score we identify maxLogit as the best one overall.



PREPROCESSING: /255											
	cityscape	RA-21		RO-21		FS L&F		FS static		Road Anomaly	
	mIoU	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95	AuPRC	FPR95
t=0.5	54.07	22.15	72.24	0.58	90.41	0.95	82.31	4.30	63.97	10.53	88.03
t=0.75	54.07	23.09	72.04	0.62	90.16	1.11	80.65	4.67	63.63	10.65	88.23
t=1.1	54.07	24.28	72.47	0.69	89.56	1.38	79.74	5.29	63.23	10.82	88.52
t=1.5	54.07	25.24	73.71	0.78	89.22	1.64	78.91	5.90	63.27	10.98	89.00
t=2.0	54.07	25.93	75.83	0.90	89.11	1.83	78.27	6.36	63.78	11.09	89.58
t=2.5	54.07	26.28	78.06	0.99	89.14	1.93	77.86	6.59	64.61	11.14	90.06

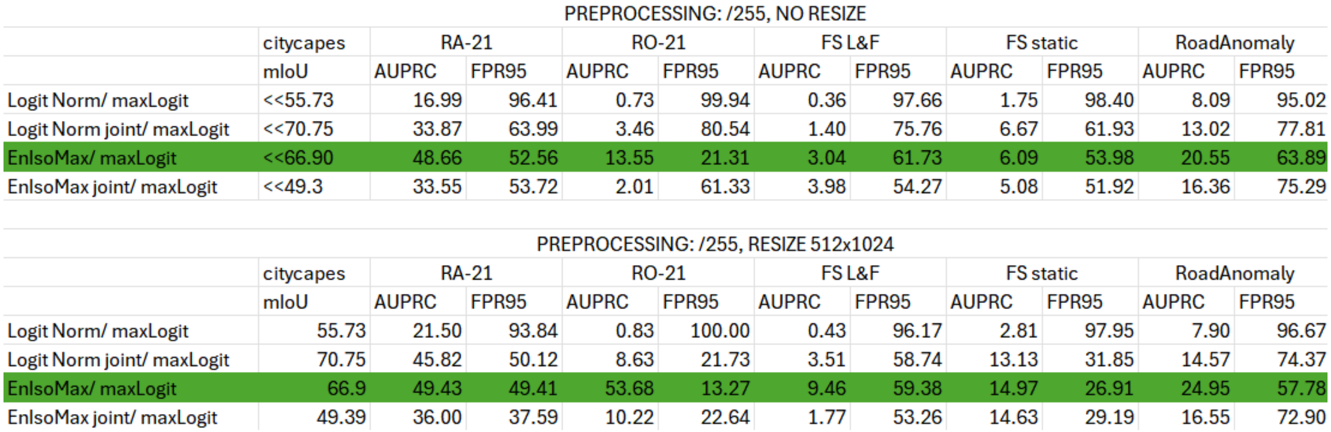
Figure 4. MSP with temperature scaling experiments with different preprocessing steps. We observe that temperature scaling with a value slightly different from 1 does not have a significant impact on anomaly detection performance.



PREPROCESSING: /255, MINIMAL OR NO RESIZE											
	cityscapes	RA-21		RO-21		FS L&F		FS static		Road Anomaly	
Network	mIoU	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95
ERFnet	54.61	25.16	69.35	0.76	94.66	5.51	27.36	15.37	45.65	10.38	90.27
ENet	45.68	17.30	83.95	1.50	54.14	2.91	51.92	7.37	56.58	9.69	83.09
BiSeNet	63.96	46.18	84.86	35.67	72.52	5.57	52.16	9.12	80.22	16.24	89.83

PREPROCESSING: /255, RESIZE 512x1024											
	cityscapes	RA-21		RO-21		FS L&F		FS static		Road Anomaly	
Network	mIoU	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95
ERFnet	71.22	24.10	65.46	0.48	97.52	0.89	55.48	3.24	81.09	9.29	92.52
ENet	63.6	17.71	82.84	0.64	72.53	1.34	57.22	4.55	79.76	10.09	85.04
BiSeNet	72.08	57.00	75.54	52.92	70.65	17.52	47.30	12.49	90.46	15.33	88.33

Figure 5. Void classifier experiments with different architecture and preprocessing steps shows that BiSeNet consistently outperformed the others in anomaly detection, we also notice it achieves an higher mIoU at original cityscapes size of 1024x2048.



PREPROCESSING: /255, NO RESIZE

	citycapes	RA-21		RO-21		FS L&F		FS static		RoadAnomaly	
	mIoU	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95
Logit Norm/ maxLogit	<<55.73	16.99	96.41	0.73	99.94	0.36	97.66	1.75	98.40	8.09	95.02
Logit Norm joint/ maxLogit	<<70.75	33.87	63.99	3.46	80.54	1.40	75.76	6.67	61.93	13.02	77.81
EnIsoMax/ maxLogit	<<66.90	48.66	52.56	13.55	21.31	3.04	61.73	6.09	53.98	20.55	63.89
EnIsoMax joint/ maxLogit	<<49.3	33.55	53.72	2.01	61.33	3.98	54.27	5.08	51.92	16.36	75.29

PREPROCESSING: /255, RESIZE 512x1024

	citycapes	RA-21		RO-21		FS L&F		FS static		RoadAnomaly	
	mIoU	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95	AUPRC	FPR95
Logit Norm/ maxLogit	55.73	21.50	93.84	0.83	100.00	0.43	96.17	2.81	97.95	7.90	96.67
Logit Norm joint/ maxLogit	70.75	45.82	50.12	8.63	21.73	3.51	58.74	13.13	31.85	14.57	74.37
EnIsoMax/ maxLogit	66.9	49.43	49.41	53.68	13.27	9.46	59.38	14.97	26.91	24.95	57.78
EnIsoMax joint/ maxLogit	49.39	36.00	37.59	10.22	22.64	1.77	53.26	14.63	29.19	16.55	72.90

Figure 6. Anomaly segmentation results of ERFNet trained with Logit Normalization loss, Logit normalization + Cross Entropy and Focal loss, Enhanced Isotropy Maximization loss and Enhanced Isotropy Maximization loss + Cross Entropy and Focal loss. Anomaly score used is maximum logit.

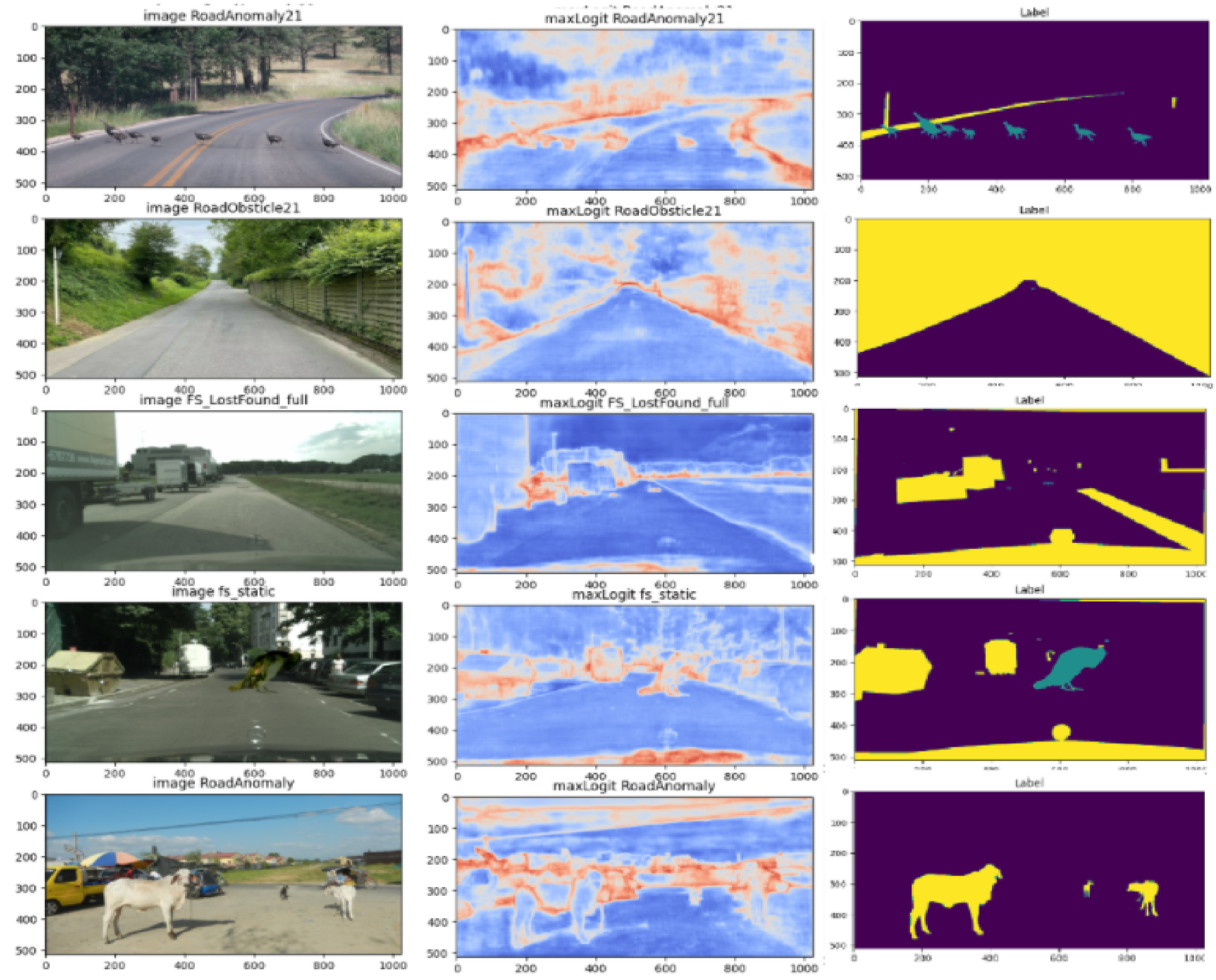


Figure 7. Qualitative results of maxLogit baseline on a scale of 512x1024
on first image of each test dataset.

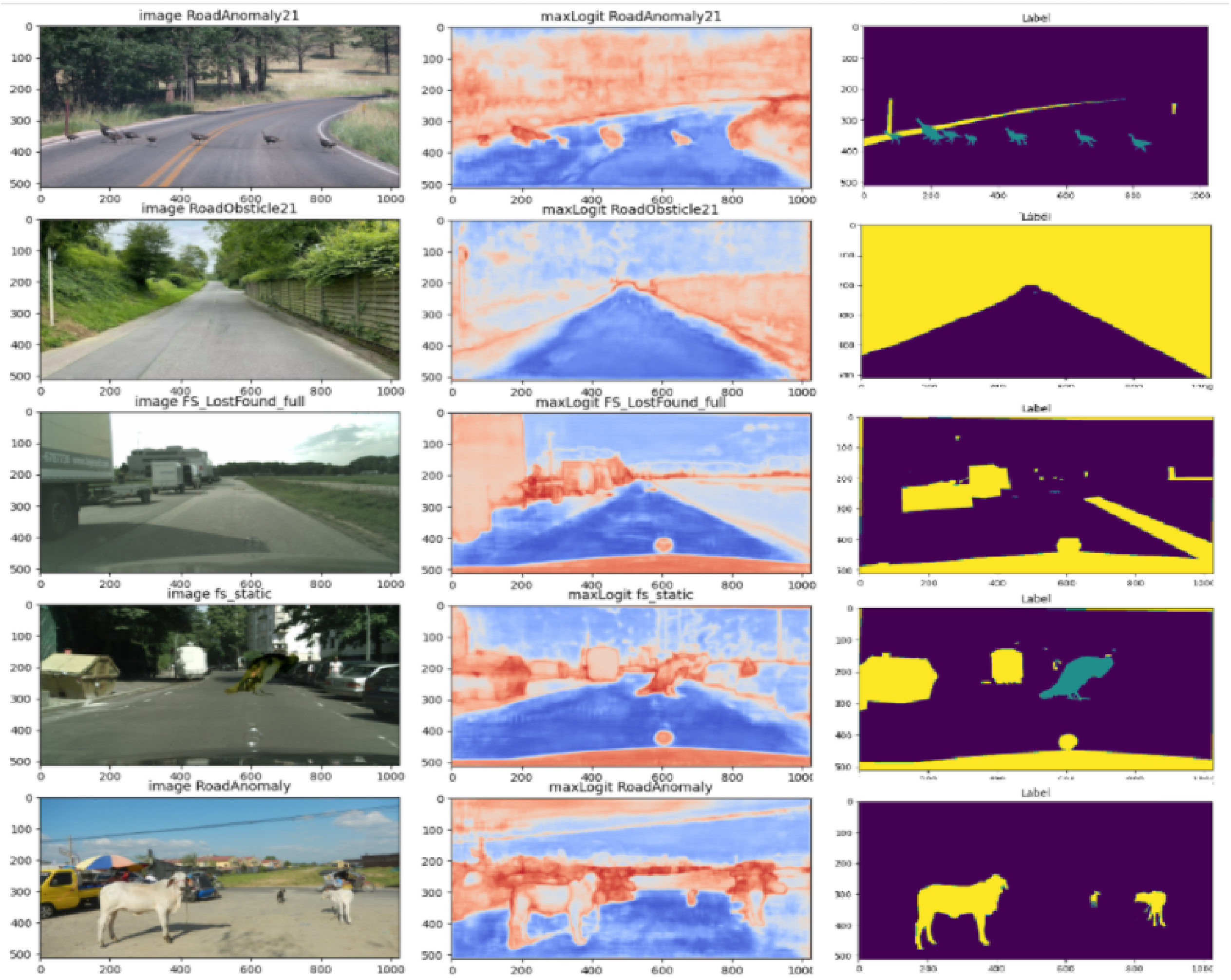


Figure 8. Qualitative results of best performing model: **ErfNet with enthanced IsoMax loss rescaled to 512x1024** on first image of each test dataset.

6. Results discussion and conclusion

Anomaly detection often relies on deep aspects of the model's feature representations and its ability to identify deviations from learned patterns. Temperature scaling mainly helps with posterior calibration (how confident the model is about its predictions), which doesn't directly improve the model's capacity to recognize anomalous patterns. This is likely why MSP, even with temperature scaling, was not the most effective method for calculating the score. Instead, maxLogit and maxEntropy, for the reason explained in the scoring methodology, addressed the problem more effectively.

BiSeNet architecture showed to be the best one both in term of accuracy and speed, we think that this network's anomaly detection preformance also were better due to multi-scale aggregation that allow good performance at different size of anomalous objects. Probably optimizing this architecture with on purpose losses will bring more benefit to anomaly detection capability, although, since optimizing BiSeNet is not as simple as for encoder-decoder architectures it would become even harder with different loss function.

Logit normalization loss, despite limiting the expression capability of the network, can be used jointly with weighted cross-entropy loss and focal loss to help convergence.

Enhanced isotropy maximization loss proved to be the most effective one, this is probably thanks to the fact that it doesn't apply constraint on last convolutional layer's output direcly. As a downside, since it requires to add a weight layer to the model, it lead to relevant latency increase and memory occupation that can be a problem particularly in case of deployment in real world application.

We found out that except some methods AUPRC was difficult to optimize for RoadObstacle21 and Fishyscapes Lost and Found, this is due to the strongly imbalanced distribution of anomaly class in this 2 dataset, however when applying enhanced isotropy maximization loss we were able to obtain appreciable results.

We also noticed that better performing methodologies tend to optimize FPR95 metric more easily rather than AUPRC and this can be noticed also from differencies between Figure 5 and Figure 8 The need for more robust anomaly detection and the ability to handle dynamic environments remains a challenge. Future research could focus on incorporating more advanced attention mechanisms or learning strategies to handle rare or unseen objects better.

A natural extension of this work can be to try BiSeNet with logit normalization or enhanced isotropy maximization loss depending wether we want to achieve lower latency or better outlier detection capabilities. A way to consistently improve overall segmentation performance can be the usage of transformer architectures, but due to their computational load a better approach for a real-time application can be knowledge-distillation on lighter network with the addition of loss functions designed for anomaly detection.

References

[1] Hermann Blum, Paul-Edouard Sarlin, Juan I. Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *CoRR*, abs/1904.03215, 2019. 1

[2] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation, 2021. 1

[3] 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. *CoRR*, abs/1604.01685, 2016. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[7] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings, 2022. 2

[8] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018. 2

[9] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis, 2019. 1

[10] David Macedo, Tsang Ing Ren, Cleber Zanchettin, Adriano L. I. Oliveira, and Teresa Ludermir. Entropic out-of-distribution detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2021. 2

[11] David Macedo, Tsang Ing Ren, Cleber Zanchettin, Adriano L. I. Oliveira, and Teresa Ludermir. Entropic out-of-distribution detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2021. 2

[12] David Macêdo and Teresa Ludermir. Enhanced isotropy maximization loss: Seamless and high-performance out-of-distribution detection simply replacing the softmax loss, 2022. 2, 3

[13] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *CoRR*, abs/1606.02147, 2016. 1, 2

[14] Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, and Barbara Caputo. Unmasking anomalies in road-scene segmentation, 2023. 1

[15] Eduardo Romera, José M. Álvarez, Luis M. Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2018. 1, 2

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 1

[17] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization, 2022. 2, 3

[18] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *CoRR*, abs/1808.00897, 2018. 1, 2