




Pixel-Wise Detection of Road Obstacles on Drivable Areas by Introducing Maximized Entropy to Synboost Framework

Carlos David Ardon Munoz^(✉), Masashi Nishiyama, and Yoshio Iwai

Tottori University, 101 Minami 4-chome, Koyama-cho, Tottori 680-8550, Japan
m22j4003x@edu.tottori-u.ac.jp, {nishiyama,iwai}@tottori-u.ac.jp
<https://www.tottori-u.ac.jp/>

Abstract. Semantic segmentation plays a crucial role in understanding the surroundings of a vehicle in the context of autonomous driving. Nevertheless, segmentation networks are typically trained on a closed-set of inliers, leading to misclassification of anomalies as in-distribution objects. This is especially dangerous for obstacles on roads, such as stones, that usually are small and blend well with the background. Numerous frameworks have been proposed to detect out-of-distribution objects in driving scenes. Some of these frameworks use softmax cross-entropy measurements as an attention mechanism for a dissimilarity network to find anomalies. However, a significant limitation arises from the segmentation network's tendency toward overconfidence in its predictions, resulting in low cross-entropy in regions where anomalies are present. This suggests that normal cross-entropy is a low-quality prior for anomaly detection. Therefore, for the task of detecting stones on roads, we propose utilizing a fined-tuned segmentation network with a changed target, from semantic segmentation to maximize the cross-entropy in anomalous areas. With this, we feed the dissimilarity network with a better prior image. Furthermore, due to the lack of datasets with enough samples of stones for pixel-wise detection, we synthetically added stones on images of driving scenes to create a dataset for fine-tuning and training. The results of our comparative experiments showed that our model attains the highest average precision while having the lowest false positive rate at 95% true positive rate when evaluating on a real-stone image dataset.

Keywords: Anomaly classification · Semantic segmentation · Computer vision

1 Introduction

Understanding surroundings through images is an important task in applications such as autonomous vehicles and robots, where it is vital to identify abnormal objects that may be a danger to vehicles and must be avoided. For this study,

we focus on a subclass of anomalies, stones. Because those are a type of obstacle commonly found on roads and dangerous to vehicles. Additionally, keeping roads clear of obstacles, such as stones, is a continuous task done by local governments and municipalities. Who regularly conduct car patrols with experts to manually identify obstacles so they can take appropriate corrective action. However, by requiring the intervention of experts, the efficiency of detection is limited since the frequency of patrols is normally low. This increases the time that obstacles are dangerous to vehicles, raising the probability of accidents and damage. Thereby, developing a support system for detecting this type of obstacle can help to overcome these problems by allowing automatic detection during patrolling.

Given the critical need of efficient anomaly detection systems, many techniques have been applied to perform anomaly detection from a single monocular RGB image. For example, uncertainty measurements in semantic segmentation, such as softmax entropy and the difference between the two largest softmax values (softmax distance), have been used to statistically detect areas of low segmentation reliability and classify them as anomalies [1]. Additionally, generating a photo-realistic image (image re-synthesis) from a semantic segmentation map and comparing it with the original image can be used to detect anomalies. As the segmentation network will not be able to understand anomalies, they will not appear in the reconstruction [2]. Thus, the areas with significant differences between the two images are classified as anomalies. Finally, the framework Synboost [3] complements the results of both techniques by using a dissimilarity network to find differences between the input and synthesized images with softmax entropy and softmax distance as attention mechanisms. However, when anomalies blend well with its surroundings or only span a few pixels in the image, which is the case for most stones on the road, the softmax cross-entropy is low and does not contribute much with the framework’s prediction ability.

Therefore, in Sect. 2, we propose a method to overcome the limitations of Synboost [3] and improve detection accuracy of anomalies in a constrained setting, namely, small anomalies laying on the drivable area (Fig. 1). We call these kinds of anomalies, road obstacles. In Sect. 3, we describe the process of training the dissimilarity network of Synboost using a dataset with synthetically added obstacles to compensate for the lack of anomalies examples in usual datasets. Furthermore, we introduce to our framework a neural network that strives to maximize cross-entropy in regions where there are anomalies [4]. With this, we increase the cross-entropy even for small anomalies that previously were unnoticeable on softmax entropy and softmax distance images, in consequence, they produce a better contribution to the final prediction, acting as attention mechanisms. Finally, in Sect. 4, we present our conclusions, highlighting how our framework enhances the pixel-wise detection accuracy for stones and significantly reduces false positives compared to previous methods.



Fig. 1. Drivable area. The region-of-interest are roads and sidewalks, which are highlighted in white in this image. We employed semantic segmentation inference to identify the contours of this area. Any anomaly completely enclosure within the drivable area is a road obstacle for this research.

2 Proposed Framework

2.1 Related Methods

We built our framework on top of Synboost [3]. This framework combines two techniques for anomaly detection. The first one is using uncertainty measurements, softmax entropy and softmax distance. The second one is image re-synthesis, that compares the original input with a generated image to find differences between both. We expect that anomalous regions in the original image look different in the generated one. The outputs of uncertainty measurements and image re-synthesis are used as inputs for a dissimilarity network that carries out a binary classification for each pixel with classes for anomaly and non-anomaly. This framework achieves state-of-the-art performance for anomaly detection with minimal computational cost for training since it is designed to use pre-trained models and only requires training for the dissimilarity network.

We chose this model to build upon our framework because it addresses different scenarios when dealing with semantic segmentation anomalies [3]. The first scenario occurs when an anomaly is misclassified as any of the inlier classes, resulting in low entropy due to overconfidence but a significantly different re-synthesized image if the object is large. The second scenario involves over-segmentation of the anomaly, with different sections assigned to different classes, leading usually to higher entropy. The final scenario occurs when the anomaly goes undetected and is classified as part of its surroundings. In this case, the anomalous area appears different in the re-synthesized image, particularly if it is sufficiently large. However, Synboost still has a low accuracy for detecting objects that blend with the background (low entropy) and only encompass a few

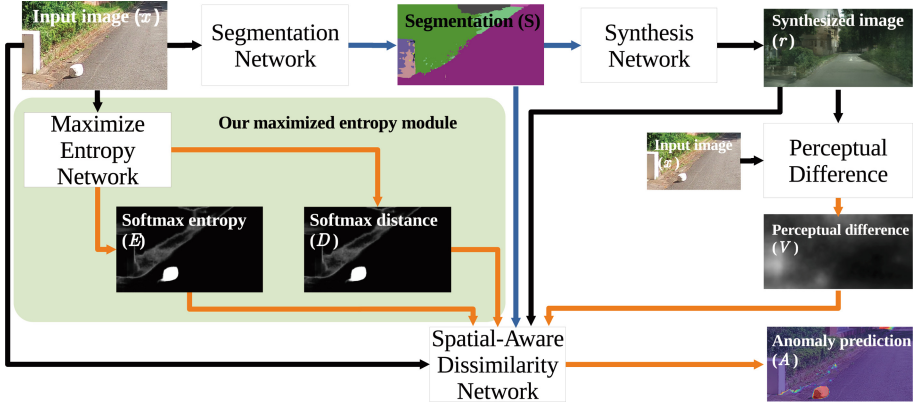


Fig. 2. Outline of our framework architecture. In Synboost [3], softmax entropy and softmax difference images are taken from the semantic segmentation network. In our framework, we introduced to the Synboost framework a Maximized entropy network [4] fine-tuned using the loss function showed in Eq. (2). Thus, this module specializes in creating prior images to feed into the Spatial-Aware Dissimilarity Network that produces the anomaly prediction (highlighted with white contours for better visualization).

pixels in the images (similar synthesized image), which is the case for most road obstacles, including stones, as both approaches fail to detect the anomaly.

2.2 Our Framework Architecture

Figure 2 shows an outline of the proposed framework. We use a pre-trained segmentation network with a WideResNet38 backbone, which is trained on Cityscapes dataset [5] according to [6]. For the synthesis network, we take a pre-trained model from the CC-FPSE framework [7], which is a conditional generative adversarial network. In the case of the perceptual difference V , we simply take a pre-trained VGG19 on ImageNet dataset as a feature extractor to find differences between the original and generated images, following the same procedure as Synboost [3]:

$$V(x, r) = \sum_{i=1}^N \frac{1}{M_i} \left\| F^{(i)}(x) - F^{(i)}(r) \right\|_1, \quad (1)$$

where $F^{(i)}$ is the i -th layer with M_i elements in the VGG network with N layers. This equation computes the perceptual difference as the sum of the absolute differences between the feature representations of the original image x and the generated image r across all layers of the VGG network. Therefore, it captures the perceptual variations and discrepancies between the two images in feature space.

Furthermore, because neural networks tend to be overconfident in their predictions, the cross-entropy can be low even in regions where an anomaly exists. In consequence, the uncertainty measures may not contribute much to the final prediction. To deal with this problem, we take a segmentation network pre-trained on the Cityscapes dataset and fine-tune it following the procedure of [4]. With this, we aim to create a better-quality prior image as input for the dissimilarity network that maximizes the cross-entropy on regions where there are anomalies, while minimizing it on regions where there are no anomalies. To accomplish this, during fine-tuning, we modify the loss function to minimize the target:

$$L = (1 - \lambda)\mathbb{E}_{(x,y)\sim\mathcal{D}_{in}}[\ell_{in}(f(x), y(x))] + \lambda\mathbb{E}_{(x')\sim\mathcal{D}_{out}}[\ell_{out}(f(x'))], \quad (2)$$

where

$$\ell_{in}(f(x), y(x)) = - \sum_{j\in C} \mathbb{1}_{j=y(x)} \log(f_j(x)), \quad (3)$$

$$\ell_{out}(f(x')) = - \sum_{j\in C} \frac{1}{q} \log(f_j(x')), \quad (4)$$

where, $(x, y) \sim \mathcal{D}_{in}$ stands for an in-distribution example, while $(x') \sim \mathcal{D}_{out}$ is an out-of-distribution example. $f(x)$ indicates the softmax probabilities for a predicted class and $y(x)$ the corresponding ground truth. λ is a value in the range of $[0, 1]$ that controls the weight between the two single objectives. $\mathbb{1}_{j=y(x)}$ is an indicator function that yields 1 when $j = y(x)$ and zero otherwise. C is the set of q classes. The minimization of the single objective for the out-of-distribution part is equivalent to maximizing entropy in anomalous regions as stated in [4].

We can interpret this new objective as follows: for in-distribution pixels, minimizing Eq. (3), which represents the standard cross-entropy loss function when using one-hot encoding, is equivalent to maximizing the softmax value for the true class. On the other hand, for out-of-distribution pixels, Eq. (4) yields the lowest loss value when the predictions in the softmax layer are uniformly distributed across all classes. Consequently, for outlier pixels, we encourage the one-hot encoding vector to have the same value in every class. This is the reason why the logarithm in Eq. (4) is divided by the number of classes q . Such a vector will yield a high value when evaluated with Shannon entropy, hence the term ‘‘Maximized entropy’’. Shannon entropy, also known as relative entropy, is defined as follows:

$$E(f(x)) = - \sum_{j\in C} f_j(x) \log(f_j(x)), \quad (5)$$

where $f(x)$ indicates the softmax probabilities predicted by the segmentation model. We use this equation to obtain the entropy image E in Fig. 2.

In Table 1, we present an overview of the models used in our framework. The segmentation, synthesis, and perceptual difference networks are pre-trained models that can be seamlessly integrated into the framework. Regarding the Maximized entropy network, it starts as a segmentation network pre-trained on Cityscapes that we subsequently fine-tune on our composite training set using

Table 1. Model and datasets overview.

Model	Training set	Training details
Segmentation	Cityscapes	Pre-trained
Synthesis	Cityscapes	Pre-trained
Perceptual difference	ImageNet	Pre-trained
Maximized entropy	Our composite dataset	Fine-tuned
Dissimilarity	Our composite dataset	Trained

Eq. (2) as loss function. Lastly, the Dissimilarity network is trained from random initialization, employing our composite training set.

2.3 Ensemble

Because the prior images used as inputs to the dissimilarity network are also anomaly predictions, it is feasible to combine them with the dissimilarity network output generated in the last step of Fig. 2. To accomplish this, we use a weighted sum to generate a more robust prediction (A_e). We refer to this procedure as ensemble, and it can be applied to both Synboost and our model using the following equation in a pixel-wise manner:

$$A_e = w_1A + w_2E + w_3D + w_4V, \quad (6)$$

where A , E , D , and V represent the output from the dissimilarity network, softmax entropy, softmax distance, and perceptual difference images, respectively. To find the values for the weights w_1 , w_2 , w_3 , and w_4 to combine these images for ensemble, we applied a grid search restricted to values that satisfy $w_1 + w_2 + w_3 + w_4 = 1$.

This means that prior images can be used in two ways: first, as inputs for the dissimilarity network where they serve as attention mechanisms, and second, as part of an ensemble together with the dissimilarity network output.

3 Experiments

3.1 Dataset

For the training set, we started from the same methodology used in Synboost training. Consequently, we took Cityscapes training set (2975 images) using the objects labeled as void class as samples for anomalies and inferred semantic segmentation for image S in Fig. 2. To train the dissimilarity network to find differences between the original and synthesized images, we added a copy of each image and randomly swapped labels for known objects in the ground truth semantic segmentation map before image synthesis. These objects look quite different in the synthesized image and are also used as anomalies samples. For



Fig. 3. Stones included as road obstacles.

this part of the dataset, we used the modified ground truth segmentation for image S in Fig. 2.

On top of that, we synthetically added one stone on a random position of the drivable area per image. The stone dataset (Fig. 3) consists of images of four stones (83, 87, 120 and 70 images per stone, respectively) taken from different orientations and illumination conditions (left only, right only, and both). We used stones 1 and 2 for training, while stones 3 and 4 were used only for evaluation. In addition, we used six images from different spots in our University’s campus as backgrounds to synthetically add stones. Applying image augmentation techniques (horizontal flip, cropping, and adjusting brightness) we added 192 images to the training set. Finally, to conduct experiments with real-stone images, we placed stones, and we took 153 images from seven spots in our University’s campus. In sum, the training set contains 6142 images when only using composite images and 6295 images when including real-stone images, as shown in Table 2. Lastly, in Fig. 4 we show some examples of composite images.

For evaluation, we built a 147 real-stone images dataset. We placed stones 3 and 4 and took photos in six spots in the University’s campus that were not used for training. With this, all evaluations for experiments are run using places and stones unseen during training. Additionally, using a real-obstacle dataset for evaluation prevents our experiments from yielding unrealistic results caused by the use of synthetic data during training. For instance, a model might unrealistically excel at identifying merely pasted objects rather than genuine obstacles [8].

Table 2. Composition of training set.

Type	Source	Number of images
Composite	Cityscapes	5950
Composite	Our background set	192
Total composite images		6142
Real	Our real-stone set	153
Total composite + real-stone images		6295

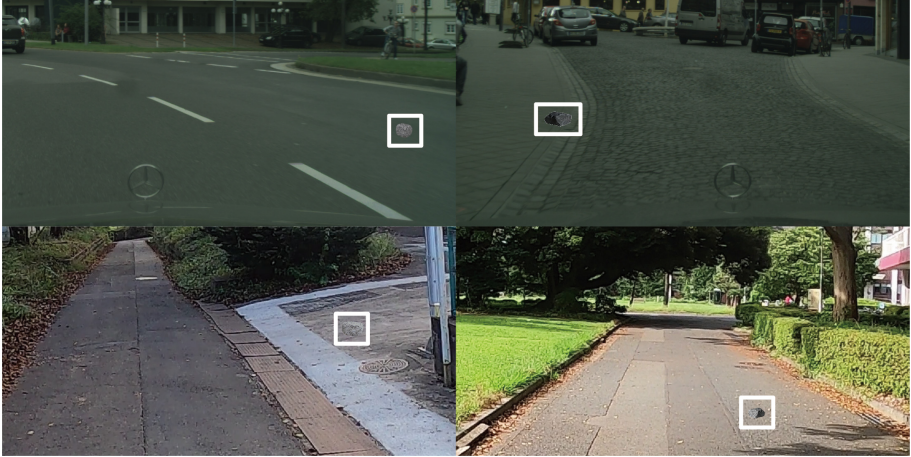


Fig. 4. Composite images. Examples of composite image used for training. In the top row there are examples of backgrounds from Cityscapes. While in the bottom row, there are examples of our backgrounds. The stones synthetically added are highlighted by a solid line rectangle.

3.2 Experimental Conditions

The region-of-interest (ROI) for evaluation of anomaly detection in our experiments comprises all pixels that belong to road or sidewalks. As well as regions assigned to other classes enclosed within the road or sidewalk. While the road obstacles that we want to detect are objects on the drivable area that do not belong to any known classes from the Cityscapes dataset [5]. We used average precision (AP) as the primary metric, as well as false positive rate at 95% true positive rate (FPR95), that are fitted for highly imbalance classification problems, such as anomaly detection.

For the first experiment, we used our dataset with only composites images to train Synboost’s and our model’s dissimilarity network, as well as fine-tuning a segmentation network to maximize entropy on anomalous areas. For the second experiment, we added 192 images with real-stone images into the training set, which improved performance for all models despite the small sample size.

Table 3. Hyperparameters used for training and fine-tuning.

	Dissimilarity network training	Maximized entropy network fine-tuning
Hyperparameter	Value	Value
Learning rate	0.0001	0.00001
Betas (Adam)	0.5, 0.999	0.9, 0.999
Epochs	30	20
Batch size	8	8

Table 4. Best weights for Eq. (6) for our model according to a grid search.

Model	Training set	w_1	w_2	w_3	w_4
Synboost	Composite images only	1	0	0	0
	Composite + real-stone images	1	0	0	0
Ours	Composite images only	0.85	0.15	0	0
	Composite + real-stone images	0.55	0.45	0	0

We trained the dissimilarity networks for 30 epochs five times per model from random initializations. In the case of entropy maximization, we fine-tuned the segmentation network for 20 epochs with $\lambda = 0.9$ in Eq. (2). In both cases, we selected the epoch with the lowest lost value on the evaluation set as the best epoch. In Table 3, we show the hyperparameters used for training and fine-tuning.

The experiments were conducted on a computer running Ubuntu 20.04.4 LTS, with Python 3.6.9, PyTorch 1.10.1, CUDA 11.4, and a NVIDIA GeForce RTX 3090 GPU with 24 GB of RAM.

3.3 Results

The outcome of a grid search for the best ensemble weights for Synboost and our model when using Eq. (6) are shown in Table 4. In the case of Synboost, we obtained that $w_1 = 1$ for both training sets, which means that it only uses the dissimilarity network output. In other words, Synboost does not benefit from ensemble. On the other hand, for our model we obtained that the dissimilarity output (w_1) is combined with softmax entropy (w_2) to create the final prediction. However, w_3 and w_4 are zero for both training sets which means that softmax distance and perceptual difference are not used for ensemble with our model.

Consequently, using the ensemble weights during testing, the results of the best networks when evaluating on real-stone images are shown in Table 5, where Synboost achieved its best performance without ensemble while our model did it using ensemble. All models improve when a small set of real-stone images are added to the training set.

Table 5. Performance comparison between best networks.

Method	Composite images only		Composite + real-stone images	
	AP \uparrow	FPR95 \downarrow	AP \uparrow	FPR95 \downarrow
Synboost [3]	64.2	7.07	67.0	4.92
Maximized entropy [4]	76.8	2.02	85.5	0.30
Ours	86.6	0.29	91.0	0.13

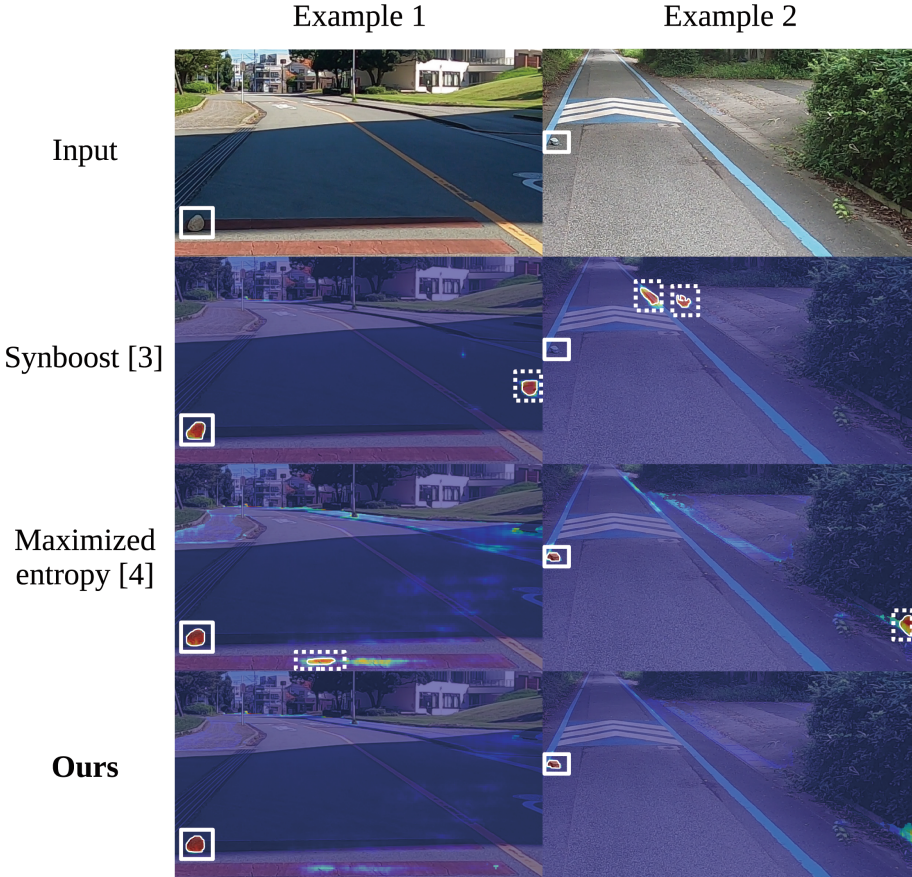


Fig. 5. Examples of anomaly predictions. The first row depicts the input image, the ground-truth anomaly is highlighted by a solid line rectangle for visualization purposes only. For the inferences, false positives are highlighted by a dotted line rectangle. The model’s anomaly prediction is indicated by a contour line. In the case of Synboost (second row), it detected the obstacle for the first example but failed to detect it for the second. In both cases, it produced false positives. For Maximized entropy (third row), it spotted correctly the obstacles in both cases, however, it also yielded false positives. In contrast, our model (fourth row) successfully recognized the obstacles in both cases without false positives.

Our experiments demonstrated that our model outperforms previous models in terms of achieving the highest average precision (AP) while simultaneously having the lowest false positive rate at 95% true positive rate (FPR95) when evaluated on a real-stone dataset. In Fig. 5, we present examples of pixel-wise anomaly prediction, where each image contains a single road obstacle. The contour lines indicate the pixels classified as anomalies by each model. Ground-truth anomalies are highlighted with solid line rectangles, while false positives are marked with dotted line rectangles.

4 Conclusions

In this paper, we presented a framework to detect small anomalies laying on roads, such as stones. We demonstrated that using maximized entropy, as an attention mechanism for a dissimilarity network, improves the average precision and false positive rate at 95% true positive rate in pixel-wise anomaly detection. Furthermore, the difference between the results from the two experiments, due to the inclusion of the small sample of real-stone images in the training set, indicates that there is still room for improvement in the composite dataset creation.

Regarding the use of ensemble, the values found from grid-search confirms that standard softmax entropy do not contribute much to the final anomaly prediction of road obstacles, as we found that the highest AP values for Synboost were achieved without taken into consideration any of the prior images for ensemble. While our model with maximized entropy uses softmax entropy images along with the dissimilarity network output in ensemble to create the final anomaly prediction. In other words, while normal softmax entropy serves as an attention mechanism for the dissimilarity network, it doesn't improve the final result when using ensemble. On the other hand, maximized entropy is beneficial in both scenarios.

In future work, we intend to conduct experiments using a composite dataset with additional types of road obstacles aside from stones to create a more general anomaly detection system. For this purpose, we plan to create a composite dataset using objects not present in the in-distribution label set of Cityscapes, treating them as obstacles. We also aim to develop a framework with local image synthesis, constraint to the drivable area to avoid synthesizing whole images and thus, reducing the computational of inferences.

References

1. Rottmann, M., et al.: Prediction error meta classification in semantic segmentation: detection via aggregated dispersion measures of Softmax probabilities. In: 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, pp. 1–9 (2020)
2. Lis, K., Nakka, K.K., Fua, P., Salzmann, M.: Detecting the unexpected via image resynthesis. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, pp. 2152–2161 (2019)

3. Di Biase, G., Blum, H., Siegart, R., Cadena, C.: Pixel-wise anomaly detection in complex driving scenes. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 16913–16922 (2021)
4. Chan, R., Rottmann, M., Gottschalk, H.: Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 5108–5117 (2021)
5. Cordts, M. et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 3213–3223 (2016)
6. Zhu, Y. et al.: Improving semantic segmentation via video propagation and label relaxation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 8848–8857 (2019)
7. Liu, X., Yin, G., Shao, J., Wang, X.: Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: Advances in Neural Information Processing Systems, Red Hook, NY, USA (2019)
8. Bogdoll, D., Uhlemeyer, S., Kowol, K., Zollner, M.: Perception datasets for anomaly detection in autonomous driving: a survey. In: 2023 IEEE Intelligent Vehicles Symposium (IV), Anchorage, AK, USA, pp. 1–8 (2023)