Reliability in Semantic Segmentation: Can We Use Synthetic Data?

Thibaut Loiseau[2]†, Tuan-Hung Vu[1], Mickael Chen[1], Patrick Pérez[3], and Matthieu Cord[1,4]

[1] valeo.ai, Paris, France
[2] LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallee, France
[3] kyutai, Paris, France
[4] Sorbonne Université, Paris, France

Abstract. Assessing the robustness of perception models to covariate shifts and their ability to detect out-of-distribution (OOD) inputs is crucial for safety-critical applications such as autonomous vehicles. By nature of such applications, however, the relevant data is difficult to collect and annotate. In this paper, we show for the first time how synthetic data can be specifically generated to assess comprehensively the real-world reliability of semantic segmentation models. By fine-tuning Stable Diffusion [31] with only in-domain data, we perform *zero-shot generation* of visual scenes in OOD domains or inpainted with OOD objects. This synthetic data is employed to evaluate the robustness of pretrained segmenters, thereby offering insights into their performance when confronted with real edge cases. Through extensive experiments, we demonstrate a high correlation between the performance of models when evaluated on our synthetic OOD data and when evaluated on real OOD inputs, showing the relevance of such virtual testing. Furthermore, we demonstrate how our approach can be utilized to enhance the calibration and OOD detection capabilities of segmenters. Code and data are made public.

1 Introduction

Despite the rapid adoption of deep networks in safety-critical applications, reliability [26,36,38] has been an overlooked factor when designing and training these models. Recent efforts are geared toward enhancing model robustness under covariate shifts in data distributions [26, 30] and improving the model's ability to detect the *unknown* [10, 12, 41]. For open-world validation, in-domain data is no longer sufficient [16]; reliable and trustworthy systems demand more rigorous testing on diverse distributions, potentially exhibiting unknown objects. However, data collection campaigns can be quickly overwhelmed by the growing number of out-of-distribution (OOD) objects and conditions, sometimes extreme.

In this paper, we propose to leverage pretrained generative models, *e.g.* Stable Diffusion 1.5 (SD), to alleviate the need for real OOD data. We use SD as a

† Work done during an internship at valeo.ai

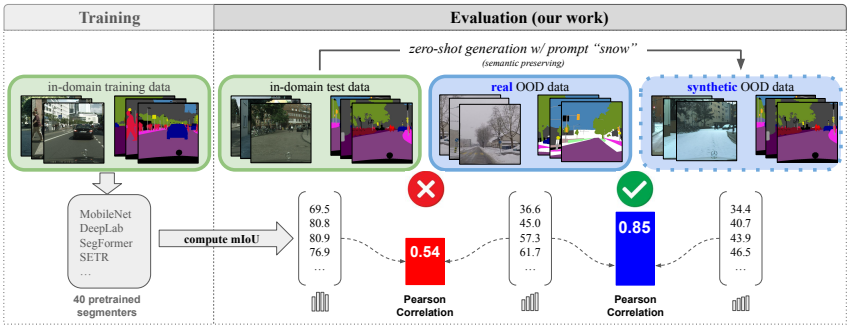

Fig. 1: Assessing 40 pretrained segmenters under covariate shifts. Segmentation models under scrutiny were trained on Cityscapes train set only (in-domain data). They are evaluated on (i) Cityscapes validation set, (ii) real OOD data, and (iii) proposed synthetic data. We observe a strong correlation between results on (ii) and (iii).

general validator, targeting multiple faces of reliability:
- First, we fine-tune SD on in-domain data to enable mask conditioning; in action, zero-shot prompting generates covariate shift images for testing. Fig. 1 illustrates the generation of synthetic data in OOD domains and its use for model evaluation. By assessing 40 pretrained segmenters under covariate shifts, our first contribution demonstrates how our generated synthetic OOD benchmark can act as a powerful proxy for the real OOD benchmark ACDC [33].
- Second, we explore a similar strategy for OOD object detection assessment: we inpaint objects of unknown classes into in-domain data by injecting appropriate prompts during the zero-shot inpainting process. Inpainted images and OOD masks are used to benchmark the 40 models to see how well they can recognize OOD objects. Our results are strongly correlated to the real OOD benchmark [4]. Our high-quality synthetic data is featured in the official BRAVO benchmark.
- Third, we demonstrate the usefulness of our two synthetic OOD benchmarks for hyperparameter tuning (here model calibration) and training. We calibrate the pretrained models to targeted OOD domains using our synthetic data and validate our strategy by comparing results to using real OOD data. We also train segmenters on inpainted data for OOD detection, obtaining competitive results.

2 Related Work

Covariate Shifts. Modern machine learning models, notably deep networks, fall short in preserving their robustness and estimating their prediction confidence in the presence of covariate shifts [26, 30]. Various benchmarks [5, 11, 16, 35, 36] have emerged to address the need for assessing models' reliability under different distributional shifts. Hendrycks et al. [11] propose a pioneering benchmark featuring data corrupted by various synthetic perturbations such as noise, blur, and brightness. Taori et al. [36] emphasize the importance of realistic shifts in

reliability assessment; they highlight the disparity between natural and synthetic shifts, asserting that there is minimal to no robustness transfer from synthetic to natural distribution shifts. Sign et al. [35] study low-shot robustness to natural distribution shifts, highlighting the robustness properties of advanced architecture and pretraining strategies. In [14], de Jorge et al. extend beyond the standard classification setup to address the reliability problem in semantic segmentation, with findings largely aligned with previous works on classification; interestingly, they point to the disconnection between robustness and confidence calibration, urging more attention to calibration during segmenter design and training. On a related line, prior works study the connections between in-domain and out-of-domain robustness [14, 23, 37]; they suggest either positive, none, or even negative correlations between ID and OOD robustness, which indeed largely depends on the type of shifts. In line with [14], we focus on segmentation as the visual perception task of interest. Different from previous works, we advocate for the use of advanced generative models to generate realistic synthetic data for testing segmenters under arbitrary covariate shifts. Taori et al. [36] criticize the synthetic robustness benchmarks and advocate for using real-shift data; while we agree on the importance of realism in testing, we demonstrate that the rapid advancement of generative models now permits very meaningful virtual assessment. Our goal is to study whether synthetic data can be a superior choice compared to ID data in correlation studies against OOD robustness – referred to as real-shift robustness here to avoid confusion with OOD detection. Inspired by [14], we also calibrate models but using synthetic OOD data instead.

OOD Object Detection. In addition to robustness and calibration, the ability to detect the "unknown" is equally important to assess for trustworthy systems [10,12,41]. In semantic segmentation, several datasets are available for evaluating OOD detection on the road, namely LostAndFound [27], StreetHazards [9] (synthetic), BDD-Anomaly [9], Fishyscapes [3] (synthetic), and SegmentMeIfYouCan [4]. Encountering and capturing images of OOD objects in real-world scenarios, without deliberately placing them on the road, is quite uncommon. Effectively, existing datasets are limited in scale, both in terms of the number of images and the variety of object classes. In this work, we propose leveraging advanced zero-shot inpainting techniques to augment an existing segmentation dataset with the insertion of OOD objects; this enables the generation of highly realistic synthetic data for testing and training OOD detection.

Synthetic Data for Testing. Generative models have been exploited to create training data for image classification [2, 8, 34], object detection [22], or semantic segmentation [7, 17, 18, 40, 43]. Only recently, a few works have delved into the topic of generative testing data. Li et al. [19] exploit diffusion models to realistically edit images, controlling over various object attributes. This approach enables stress-testing models and understanding their sensitivity to different attributes. Using SD, LANCE [29] generates counterfactual images capable of challenging any given perception model. To the extent of our knowledge, no existing works have proposed to generate testing data for segmentation reliability.

3 Reliability Under Covariate Shifts

In this section, we explore whether synthetic data can be used to assess the robustness of pretrained segmenters in the presence of covariate shifts to unseen OOD domains. We describe the data generation process and present the benchmarking results for a wide range of segmenters on our synthetic data. We demonstrate the validity of the approach using domains for which a real OOD dataset exists so that we can have access to a gold standard. We stress that we do not use the OOD data at any point in the method itself. This aspect is critical for the method to be applicable to benchmark robustness in the presence of extreme or hazardous conditions.

3.1 Generating images in arbitrary domains

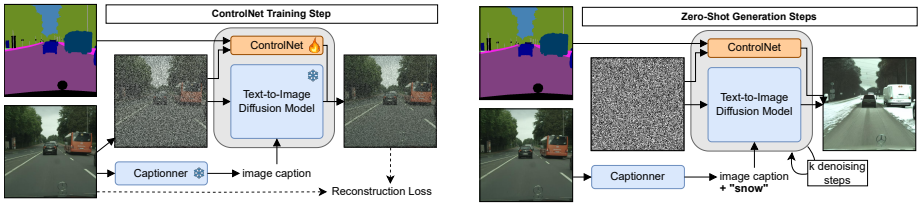

Fig. 2: Generating data with covariate shifts. Training (left) and Sampling (right) processes for producing the synthetic data with shifts. For training, only in-domain images and masks are used. For inference, we use the in-domain masks to generate OOD images. No real OOD data is required in the framework.

Our goal here is to obtain pairs of images and semantic masks, with the images belonging to visual domains for which we lack data. To this end, we leverage a pretrained text-to-image Stable Diffusion 1.5 (SD) model [31], repurposed as a semantic-conditioned model called ControlNet [42]. ControlNet is trained solely on images and segmentation ground truths from the Cityscapes dataset; at train time, the text prompts are captions automatically extracted using CLIP-interrogator [1]. As a result, the model is able to perform mask-to-image generation of driving scenes while retaining the ability to steer the generation through text prompting of Stable Diffusion.

To generate synthetic data, we prompt a trained ControlNet by the concatenation of OOD domain descriptions and CLIP-interrogator captions obtained from Cityscapes validation images. Also, segmentation masks in Cityscapes validation set are used to condition the generative process. Thanks to zero-shot prompting, the synthetic images are aligned with the semantic condition while displaying the visual properties of arbitrary OOD domains. Fig. 2 illustrates the training and generation steps. Detailed technical descriptions and more visualizations are in the Supplementary Material.

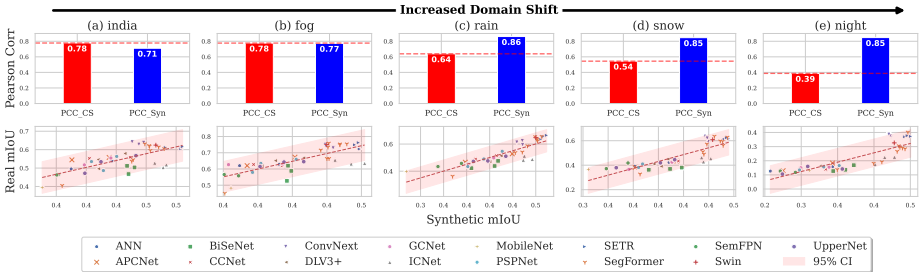

Fig. 3: Robustness correlation between real and synthetic covariate shifts across 40 **pretrained segmenters.** The tested models, see families in bottom legend, cover different architectures and sizes. (*top*) Pearson Correlation Coefficients of mIoUs between Cityscapes and real-shifts ('PCC_CS' ▮), and between synthetic shifts and real ones ('PCC_Syn' ▮). (*bottom*) Scatter plots of synthetic *vs*. real mIoUs along with the linear regression line accompanied by 95% confidence intervals ('CI'). (*a*-*e*) Five types of domain shifts from Cityscapes in-domain distribution, sorted by increasing gap as assessed by decreasing PCC_CS. The robustness results on synthetic data exhibit a strong correlation with those on real data, particularly in the case of the most distant shifts like 'snow' and 'night'. More details are provided as Supplementary Materials.

3.2 Robustness Assessment with Synthetic Data

With the pipeline outlined in Sec. 3.1, one can generate synthetic data to assess the robustness of pretrained segmenters in any unseen OOD domains through zero-shot prompting. To quantify robustness, we employ the traditional mean Intersection-over-Union (mIoU) score, measuring the correct overlap between semantic predictions and ground-truth masks. Given that our synthetic dataset comprises pairs of segmentation masks and synthetic images, one can straightforwardly derive synthetic scores for any pretrained segmenters. We here investigate whether synthetic performance can faithfully reflect the performance on real data in OOD domains under covariate shifts.

Experimental Setups. We address weather shifts and geographical shifts, which are often encountered in autonomous driving. These covariate shifts are exhibited in two existing real datasets: the Adverse Conditions Dataset (ACDC) [33] and the Indian Driving Dataset (IDD) [39]. We utilize those real data to quantify the quality of our synthetic data and to validate its usefulness.

Synthetic data are generated by conditioning on semantic masks from the Cityscapes validation set. To prompt ControlNet, we concatenate CLIP-interrogator's caption with a domain description following a simple template `[<caption>, in <domain>]` where `domain` is either 'india', 'fog', 'rain', 'snow' or 'night'.

For testing, we gather a collection of 40 publicly available segmenters *trained only on Cityscapes*, representative of different backbones, segmentation architectures, and sizes. The full list of models is in Supplementary Material.

Results. In Fig. 3, we present our main results. Our primary metric is the Pearson Correlation Coefficient (PCC) between the mIoUs on testing data and on

real-shift data from ACDC's splits or from IDD. The testing data can be either the Cityscapes validation set (CS) or our synthetic data (syn); the idea is to see which testing data –whether real CS or our synthetic one– correlates more with the real-shift data. Note that in the absence of OOD data, the Cityscapes validation set acts as the closest easily available proxy and is a reasonable predictor for OOD performance as pointed out by Jorge et al. [14].

We organize our results based on increasing domain gaps relative to the Cityscapes domain. The domain gaps are quantified by the Pearson correlation between Cityscapes mIoUs and real-shift mIoUs, annotated as PCC_CS and visualized as red bars in the subplots of Fig. 3. Moving from left to right, *i.e.* with growing domain gaps, we observe a widening discrepancy between PCC_CS and PCC_Syn. Here, PCC_Syn ( ) represents the Pearson correlation between synthetic mIoUs and real-shift mIoUs. In domains with small gaps, PCC_CS and PCC_Syn are relatively comparable. However, in domains with more adverse shifts, such as 'snow' and 'night', PCC_Syn outperforms PCC_CS significantly, exceeding PCC_CS in 'night' by more than double.

In Fig. 4, we analyze the results for the 'night' condition using the most robust models across different architectures, ranging from ConvNets to recent transformer networks. We use the Semantic-FPN score as the reference to normalize the scores of other models. This normalization aims to illustrate the relative improvement in robustness in terms of architecture. We rank the models from left to right based on their performance on real night data from the ACDC-night split. The consistently increasing trend of synthetic scores ( ) from left to right demonstrates a strong correlation with the ranking based on real scores. In contrast, Cityscapes scores ( ) are not indicative of night performance: a higher mIoU obtained on Cityscapes does not immediately translate into a higher mIoU at night.

Since synthetic data can be generated in any desired quantity, a natural question arises: how many images are sufficient? In addressing this question, we conducted experiments and presented the results in Fig. 5. Our empirical finding suggests that $\sim$ 500 synthetic images are adequate for a stable and reliable assessment of robustness.

Discussion. In their recent work on reliability in semantic segmentation, Jorge et al. [14] systematically quantified the robustness of segmenters on real-shift data; similarly to ours, they draw comparisons from ACDC and IDD datasets. One intriguing finding in this paper is that "[...] the larger the domain shift, the larger the improvement brought by more recent segmentation models", hinting at a correlation between model robustness on in-domain data and covariate-shift data; that corresponds to the CS baseline we consider here. In our study, we delve deeper into this correlation, choosing to separately address different weather types instead of grouping them all together as done in [14]. In domains exhibiting small gaps to Cityscapes, such as IDD or ACDC-Fog, our conclusion aligns with [14]. However, as domain gaps increase, the discrepancy between Cityscapes mIoUs and real-shift mIoUs becomes more pronounced, resulting in poor PCC_CS scores. On the contrary, synthetic mIoUs and real-shift mIoUs

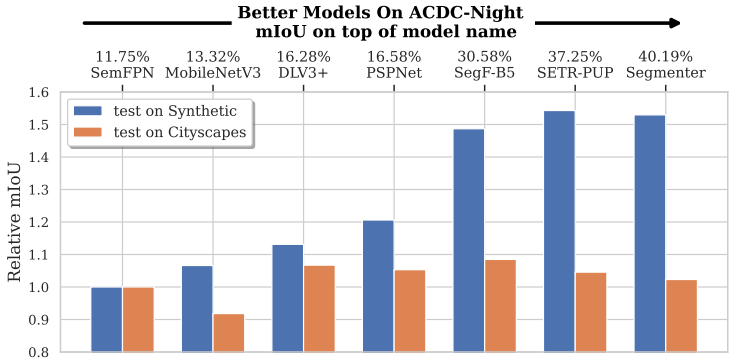

Fig. 4: Day-night shift. Models are ranked from left to right by their robustness on real night data – ACDC-Night mIoUs are shown on top of model names. For each presented architecture, the most robust model on Cityscapes is tested; the Semantic-FPN, DeeplabV3+, and PSPNet models have ResNet-101 as backbone. The Semantic-FPN model (lowest mIoU on ACDC-Night) serves as the reference for computing the relative mIoUs. Blue bars or orange bars show the relative mIoUs when testing on our synthetic data (▮) or testing on Cityscapes validation data (▮). Cityscapes scores are not reliable for ranking models in the night domain. Synthetic scores exhibit a stronger correlation with real night scores, as evidenced by the more consistently increasing trend in the blue bars from left to right.

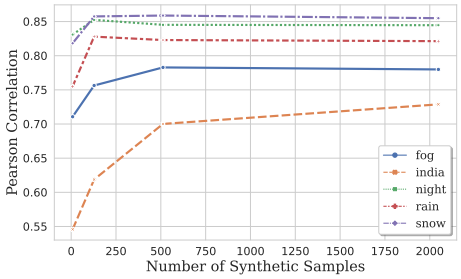

Fig. 5: Pearson Correlation *vs.* # Synthetic Samples. Using more synthetic samples contributes to increased stability in the results. Empirical plots demonstrate that approximately 500 samples are sufficient for a stable correlation assessment.

exhibit a strong correlation across shifts. *Our empirical study has validated that synthetic performance is a reliable indicator of model robustness in the presence of covariate shifts.*

With synthetic data, we are able to confirm observations from the literature on real data, such as (i) the robustness of transformer and ConvNext backbones and that (ii) within an architecture family, robustness correlates with the number of parameters and the robustness of the backbone.

Different generative models? The important ingredient in our pipeline is the generative model. In Tab. 1, we ablate by replacing the default SD1.5 model with other available ones: image-2-image GAN called TSIT [13], physics-based fog simulator [32], and a bigger SD variant called SDXL [28]. On 'night', 'rain', and 'snow', SD variants perform better than TSIT, while of note, TSIT was

	OOD expertise?	OOD data?	Night	Rain	Snow	Fog	India
GAN-based TSIT [13]	no	required	0.83	0.84	0.81	-	-
Physics-based Fog Sim. [32]	required	no	-	-	-	0.82	-
Ours w/ SD1.5 (default)	no	no	**0.85**	0.86	**0.85**	0.77	0.71
Ours w/ SDXL [28]	no	no	0.84	**0.90**	0.82	**0.89**	**0.93**

Table 1: Different generative models. Our pipeline achieves better performance (PCC with real-shifts) while does not require any OOD knowledge. Of note, it's very easy to apply to new unseen OOD domains using our pipeline, which remains challenging for GAN-based and physics-based models.

trained on real OOD data. The dedicated fog simulator performs very well, on par with our results using the larger model SDXL. Physics-based simulators are definitely valuable and this direction should be investigated further; however, such simulators require in-depth knowledge of OOD domains and hence are very difficult to design. Both GAN-based and physics-based approaches are limited in their scalability to many more OOD domains. Comparing SD1.5 *vs*. SDXL, we observe comparable results on challenging OOD domains while SDXL performs much better on 'fog' and 'india'; such results hint at the future potential of stronger and better generative models in further advancing virtual testing. Unfortunately, as SDXL is much more memory-demanding, we only limit our experiments with SDXL in this particular study.
In the Supplementary Material, we report the FID scores, which measure the direct distance between synthetic and real distributions; we also extend a discussion on some limitations of our framework.

3.3 Confidence Calibration with Synthetic Data

Confidence calibration is a crucial aspect of deep networks, particularly when employed in safety-critical applications such as autonomous driving. Jorge et al. [14] highlighted a disconnection between model robustness and calibration, asserting that "... despite the remarkable improvements in terms of robustness, recent models are not significantly better calibrated". Therefore, it is essential to devise techniques and protocols for recalibrating data, particularly in domains exhibiting covariate shifts. Drawing inspiration from this, we explore the feasibility of using synthetic data to recalibrate pretrained segmenters.

We perform temperature scaling using our synthetic data. Temperature scaling [6] is a well-established technique for calibrating pretrained models, typically conducted on a small validation set within the OOD domain. In our study, for each segmenter, we utilized the same sets of synthetic data generated in Sec. 3.2 to optimize temperature scaling factors, with one adjustment made for each covariate shift. For comparison, we replicate the process using real-shift data from ACDC and IDD.

Fig. 6 reports the calibration improvement for the 40 pretrained segmenters using either real-shift data (■) or synthetic data (■). The subplots are arranged

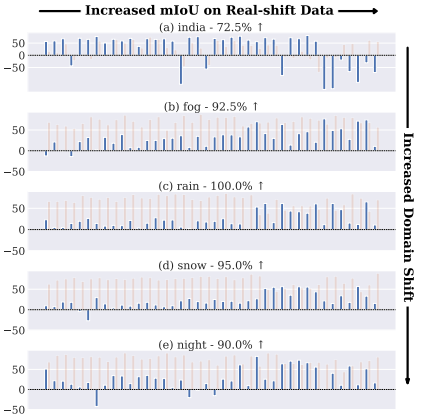

Fig. 6: ECE improvement using synthetic data. The 40 segmenters are calibrated using either real-shift data or synthetic data. For each model, the relative ECE improvement (%) over its non-calibrated version is computed, visualized by ■ (synthetic shift) and ■ (real shift). The subplots are ranked by model robustness on real-shift data (left-to-right) and by the increased domain shift (top-to-bottom). The titles of the subplots indicate the percentage of models that showed improvement with synthetic data.

in increasing domain gap order from top to bottom, with the segmenters ranked from left to right based on increasing robustness on real-shift data. The Expected Calibration Error (ECE) [24] quantifies the calibration results, with a lower ECE indicating a better-calibrated model. For better interpretation, we present the relative ECE improvement, computed as the percentage decrease in ECE after calibration compared to the original ECE without calibration. For example, a model with an ECE of 0.4 before calibration and an ECE of 0.2 after calibration will achieve a $(0.4 - 0.2)/0.4 = 50\%$ relative improvement.

We observe promising calibration results when employing our synthetic data. While not as good as real-shift data, synthetic data achieves a promising success rate of 72.5% on IDD and exceeds 90% on the four ACDC shifts. Interestingly, in weather shifts, we empirically observe that more robust models derive greater benefits when calibrated using our synthetic data; the reverse is observed for 'europe-india' geographical shift. While with real-shift data, robustness and calibration are not well correlated [14], our results suggest that a potential correlation might exist between the two factors when using synthetic data. We note that calibration with temperature scaling does not always guarantee ECE improvement. Such phenomenon may happen even using real data, especially under domain shifts as explored in prior work [26]. In the Supplementary Material, we provide more technical details and results.

3.4 On Practical Applicability

One significant advantage of our framework lies in its potential to address rare conditions simply through prompting. The practical applicability of generative testing is tremendous. Our results demonstrate promising signals; practitioners can begin assessing and ranking their pretrained models for a new, unseen OOD domain of interest *without the need for real data collection*. In practice, our proposed generative benchmarking can serve as the initial step in a full validation pipeline, helping filter out non-robust prototypes and thereby saving on total operational costs. Starting from complementing real-data validation, one

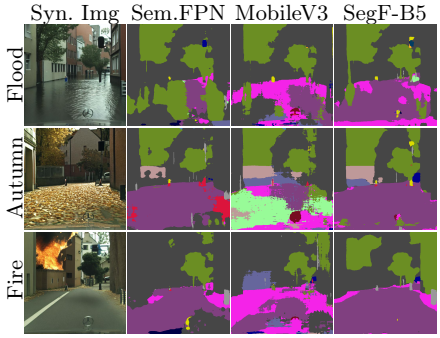

Fig. 7: Qualitative results. Examples of rare conditions generated for testing and predictions from different models. Results of the strong model like SegFormer-B5 are visibly better than the Semantic-FPN and MobileNetV3.

can envision a future where generative techniques mature to the point of fully replacing real-data validation. In Fig. 7, we visualize some synthetic images and model predictions under rare conditions, such as being flooded with water, having autumn leaves scattered across the road, or having a building on fire. We observe clear visual distinctions between weaker (Semantic-FPN and MobileNetV3) and stronger (SegF-B5) models, knowing that their Cityscapes scores do not differ significantly. More examples are provided in the Supplementary Material.

4 Reliability Against OOD Objects

We now address the reliability of segmentation models in the presence of Out-of-Distribution (OOD) objects. To begin, we explain our pipeline for inpainting random OOD objects into existing Cityscapes images. Following that, we demonstrate how one can utilize inpainted images for OOD detection assessment and for enhancing OOD detection.

4.1 Inpainting Anomaly Objects

We inpaint random objects into Cityscapes images. To this end, we initially sample a location — a square box to which we inpaint the new object. We crop the box, upsample its content to match the preferred output size of the generative model, and inpaint an object guided by a text prompt. In this step, we leverage Stable Diffusion inpainting capabilities, obtaining high-definition square images of the desired object. This image is then resized and pasted back into the original image, creating a final high-definition synthetic image. To ensure compositional consistency, we employ two techniques: Firstly, we divide the cropped box into two regions by center cropping it again. We inpaint only the inner region, leaving the outer region untouched, similar to the approach in RePaint [21]. Secondly, after composing the final image, we address any remaining inconsistencies by applying a light noise over the entire picture and performing reverse diffusion again. Details and visualizations are provided in the Supplementary Material.

After inpainting, it is necessary to extract the mask corresponding to the new object. To achieve this, we begin with a high-definition square image and

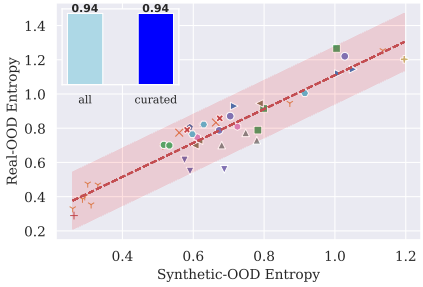

Fig. 8: Entropy Correlation. The top-left inset reports the Pearson correlations between real-OOD entropy *vs*. synthetic-OOD entropy, computed either on 'curated' ( ) or all synthetic inpainted images ( ). Evaluations are performed on the same model set used in Fig. 3, with similar markers.

apply the Grounded Segment Anything [15, 20], prompted with the name of the object. This process yields a mask within the square image, which can then be repositioned in the full image.

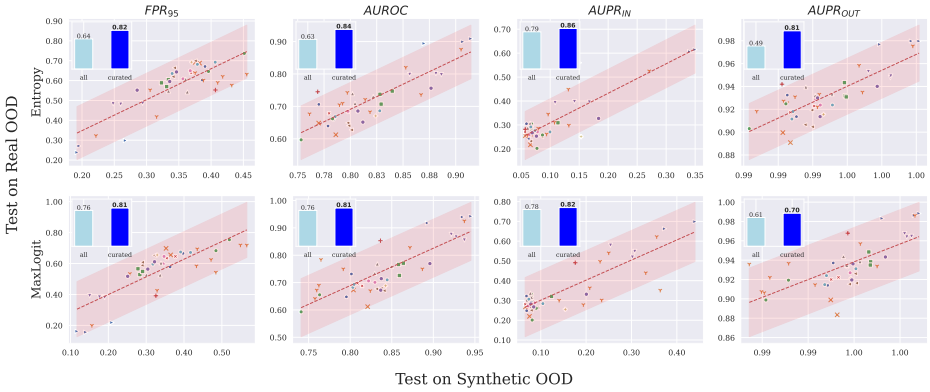

Fig. 9: Correlation in OOD Detection. Each subplot scatters computed anomaly scores of segmenters on real OODs (y-axis) and on synthetic OODs (x-axis). The top row shows the four anomaly metrics utilized: FPR_{95}, AUROC, AUPR_{IN}, and AUPR_{OUT}. The results are organized into two rows corresponding to two different confidence measures (i) Entropy and (ii) MaxLogit. In the top-left corner of each subplot, an inset plots Pearson correlations to real OOD for 'curated' ( ) and 'all' ( ) synthetic sets. Evaluations are performed on the same model set used in Fig. 3, with similar markers.

Our end-to-end generation pipeline is fully automatic. Through qualitative assessment, we achieve a satisfactory success rate in terms of generation realism; some inpainted images are illustrated in Fig. 10 and much more in the Supplementary Material. However, this still leaves a few generations with artifacts, characterized by either unusual compositions or unrealistic details. *We here question the criticality of realism in assessing OOD detection and, furthermore, in improving OOD detection.* To this end, we construct two different sets: (i) all 23,040 images generated automatically and (ii) 656 curated images where

we manually select the best images in terms of visual quality and realism; specifically, curators filter out strong color saturation differences or partial objects, *e.g.* animal heads. We note that the manual selection process for the curated set is not exhaustive and is constrained by our allocated resources; there are many more high-quality images in the 'all' set. In what follows, we present results using both curated and uncurated sets.

4.2 Assess OOD Detection

Experimental setup. To measure how the segmenters react to unseen OOD objects, we use standard anomaly detection metrics [41], which are False Positive Rate at 95% true positives (FPR_{95}), Area Under ROC curve (AUROC), and Area Under Precision-Recall curve (AUPR). AUPR are declined into $AUPR_{IN}$ and $AUPR_{OUT}$, which consider the in-distribution regions, respectively the out-of-distribution regions (the inpainted object), as positive regions to compute the Precision-Recall curves.

All segmenters in our study are not designed to produce confidence scores. We thus seek various techniques to derive confidence scores from pretrained models [9] and eventually narrow down the options to two measures: (i) Entropy of soft-probability predictions, and (ii) MaxLogit as the maximum logit value (before softmax) among the classes. While Entropy is the traditional measure of uncertainty, MaxLogit is a recent and surprising finding that has been proven to be more effective in estimating OOD confidence [9].

Results. For quantitative comparison, we leverage the SegmentMeIfYouCan (SMIYC) dataset [4], a recent dataset for OOD detection. We resort to the RoadAnomaly21 split in SMIYC, due to similarity in object scales to our synthetic data. We analyze the correlation between the OOD scores obtained on RoadAnomaly21 and one using our synthetic inpainted data. Fig. 8 reports our first analysis on the entropy in the OOD areas, either real or generated. For each model, we compute the Pearson Correlation (PCC) between real-OOD entropy and synthetic-OOD entropy; the computation is done on both 'curated' and 'all' sets. We observe a very high entropy correlation between real- and synthetic-OOD, reaching 0.94 PCC using both 'curated' and 'all' sets. In Fig. 11, we show a control experiment in which we inpaint in-domain object class 'car' into Cityscapes scenes, and we analyze models' responses to synthetic cars, real cars, and OOD objects.

We then analyze the correlation between real and synthetic anomaly metrics. Fig. 9 presents our primary findings and Fig. 10 illustrates some qualitative results. We observe a strong correlation with real scores when utilizing the 'curated' set for computing synthetic scores; the curated PCCs (▮) are consistently around 0.8 across multiple metrics, irrespective of the two confidence measures. Although the correlations are somewhat weaker when using all uncurated synthetic data (▮), such results remain acceptable, particularly when no effort is dedicated to curation.

Our results validate the potential of utilizing realistic synthetic data, inpainted with anomaly objects, for assessing OOD detection. In OOD testing, it is quite

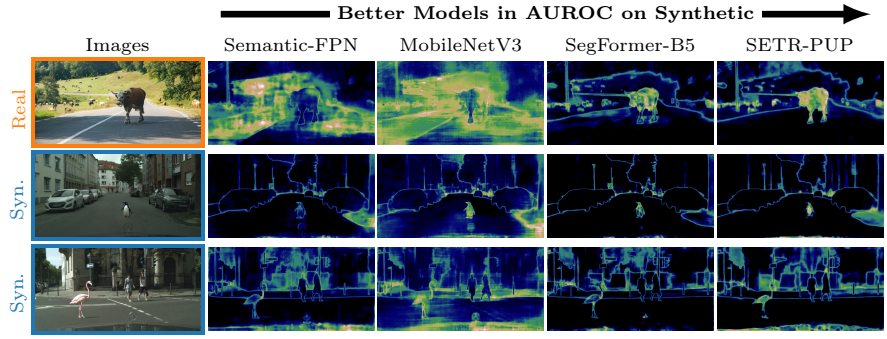

Fig. 10: Qualitative results. Confidence maps are visualized for the four exemplified models on real data (first row) and synthetic inpainted data (second and third rows). Hotter colors correspond to higher OOD likelihood. Ideally, results should exhibit hot colors in OOD areas and cold colors everywhere else. We observe a strong correlation in model reactions to real and synthetic OOD regions, particularly for more recent and robust models. We note that the real data in SMIYC also exhibit distributional shifts against Cityscapes; which already causes confusion to weak models in the background.

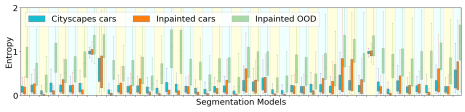

Fig. 11: In-domain vs OOD inpainted objects. Models' responses to synthetic cars are close to real cars, and far from synthetic OOD objects

important to use high-quality synthetic inpainted data. Nonetheless, even non-curated synthetic data can offer an acceptable estimation of real performance with minimal curation costs.

Method	AUROC (↑)	AUPR (↑)	FPR95 (↓)
RbA [25] Swin-B	95.6	78.4	11.8
+ COCO [25]	**97.8**	**85.3**	8.5
+ Ours (curated)	97.2	84.9	**8.1**
+ Ours (all)	97.3	84.8	8.2
RbA [25] Swin-L	96.4	79.6	15.0
+ COCO [25]	**98.2**	**88.7**	8.2
+ Ours (curated)	97.2	88.0	**7.9**
+ Ours (all)	98.1	88.6	8.3

Table 2: Improving OOD detection on real SMIYC benchmarking using our synthetic data. All results are obtained using the published code and default parameters.

4.3 Improve OOD Detection

In this experiment, we investigate if synthetic inpainted data can be used to enhance a deep network's ability to detect OOD objects. To this end, we adopt

the state-of-the-art RbA [25] approach for OOD detection and train RbA models on our synthetic data.

OOD metrics are computed on RoadAnomaly21 and reported in Tab. 2. The RbA models trained on our data significantly outperform the vanilla RbA model. We reach comparable performance to the RbA variants that leverage the external COCO dataset for augmentation. Notably, there are no clear differences between using 'curated' or 'all' sets. *We conjecture that, unlike benchmarking, training for OOD detection does not demand a high degree of realism from synthetic data.* This explains why the simple strategy of copy-pasting COCO objects [25] already proves effective. All results are consistent across the two addressed backbones. Fig. 10 illustrates a few qualitative results.

Discussion. In our work, we take all available levers and study the extent to which synthetic data can be used for evaluation and specific training purposes. We acknowledge the fact that the models used in this work were trained on substantial amounts of data. Of note, we do not claim the efficacy of synthetic data in all aspects, particularly regarding the total amount of training data required. Our findings are limited to highlighting the significant potential of published generative models in the task of reliability assessment. Advances in efficient training of generative models may address concerns regarding data quantity but are beyond the scope of this work.

5 Takeaways

In this work, we explore the potential of synthetic data in reliability assessment for semantic segmentation networks. We introduce two automatic zero-shot pipelines to generate data in OOD domains and to inpaint OOD objects for virtual reliability assessment. Our promising results encourage further collective investigations into this research problem, paving the way for synthetic system validation, especially in safety-critical applications. We summarize here our findings:

- ▷ *Reliability Under Covariate Shifts:* synthetic data can help assess the relative robustness of models in real-life covariate shifts, especially when shifts to the training condition are significant. Synthetic data can well complement real data in system validation, helping reduce the total operational cost. Pretrained models can be calibrated using synthetic data to better estimate prediction confidence in any shifted domains.
- ▷ *Reliability Against OOD Objects:* synthetic data is useful in both OOD testing and OOD training; however, the demands on synthetic data quality differ in these two cases. In OOD testing, the best result estimations are obtained with the most realistically inpainted data, which may require a certain amount of curation time for qualitative assessment. The curation task is not time-demanding and can be done quickly with a reasonable budget. On the other hand, for OOD training, no curation is actually needed to achieve improvements.

Acknowledgements

This work is supported by ELSA - European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. We thank the authors of RELIS [14] for providing us their pretrained checkpoints.

References

1. clip-interrogator. `https://github.com/pharmapsychotic/clip-interrogator` (2023)
2. Besnier, V., Jain, H., Bursuc, A., Cord, M., Pérez, P.: This dataset does not exist: training models from generated images. In: ICASSP (2020)
3. Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: The fishyscapes benchmark: Measuring blind spots in semantic segmentation (2021)
4. Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R., Fua, P., Salzmann, M., Rottmann, M.: Segmentmeifyoucan: A benchmark for anomaly segmentation. In: NeurIPS (2021)
5. Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F.A., Brendel, W.: Partial success in closing the gap between human and machine vision. NeurIPS (2021)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML (2017)
7. Hariat, M., Laurent, O., Kazmierczak, R., Zhang, S., Bursuc, A., Yao, A., Franchi, G.: Learning to generate training datasets for robust semantic segmentation. In: WACV (2024)
8. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? (2023)
9. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: ICML (2022)
10. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: ICCV (2021)
11. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
12. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
13. Jiang, L., Zhang, C., Huang, M., Liu, C., Shi, J., Loy, C.C.: Tsit: A simple and versatile framework for image-to-image translation. In: ECCV (2020)
14. de Jorge, P., Volpi, R., Torr, P.H., Rogez, G.: Reliability in semantic segmentation: Are we on the right track? In: CVPR (2023)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
16. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: ICLR (2021)
17. Le Moing, G., Vu, T.H., Jain, H., Pérez, P., Cord, M.: Semantic palette: Guiding scene generation with class proportions. In: CVPR (2021)

18. Li, D., Ling, H., Kim, S.W., Kreis, K., Fidler, S., Torralba, A.: Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In: CVPR (2022)
19. Li, X., Chen, Y., Zhu, Y., Wang, S., Zhang, R., Xue, H.: Imagenet-e: Benchmarking neural network robustness via attribute editing. In: CVPR (2023)
20. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
21. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: Repaint: Inpainting using denoising diffusion probabilistic models. In: CVPR (2022)
22. Marathe, A., Ramanan, D., Walambe, R., Kotecha, K.: Wedge: A multi-weather autonomous driving dataset built from generative vision-language models. In: CVPRW (2023)
23. Miller, J.P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P.W., Shankar, V., Liang, P., Carmon, Y., Schmidt, L.: Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In: ICLR (2021)
24. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: AAAI (2015)
25. Nayal, N., Yavuz, M., Henriques, J.F., Güney, F.: Rba: Segmenting unknown regions rejected by all. In: ICCV (2023)
26. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. NeurIPS (2019)
27. Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., Mester, R.: Lost and found: detecting small road hazards for self-driving vehicles. In: IROS (2016)
28. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
29. Prabhu, V., Yenamandra, S., Chattopadhyay, P., Hoffman, J.: Lance: Stress-testing visual models by generating language-guided counterfactual images. In: NeurIPS (2023)
30. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: ICLR (2019)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
32. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. IJCV (2018)
33. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV (2021)
34. Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: CVPR (2023)
35. Singh, A., Sarangmath, K., Chattopadhyay, P., Hoffman, J.: Benchmarking low-shot robustness to natural distribution shifts. In: ICCV (2023)
36. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. NeurIPS (2020)
37. Teney, D., Lin, Y., Oh, S.J., Abbasnejad, E.: Id and ood performance are sometimes inversely correlated on real-world datasets. arXiv (2022)
38. Tran, D., Liu, J., Dusenberry, M.W., Phan, D., Collier, M., Ren, J., Han, K., Wang, Z., Mariet, Z., Hu, H., et al.: Plex: Towards reliability using pretrained large model extensions. arXiv (2022)

39. Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C.: Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: WACV (2019)
40. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: ICCV (2023)
41. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al.: Openood: Benchmarking generalized out-of-distribution detection. NeurIPS (2022)
42. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
43. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. In: CVPR (2021)