Learning to generate training datasets for robust semantic segmentation

Marwane Hariat,[1,*] **Olivier Laurent**,[1,2,*] **Rémi Kazmierczak**,[1] **Shihao Zhang**,[3]
Andrei Bursuc,[4] **Angela Yao**[3] & **Gianni Franchi**[1,†]
U2IS, ENSTA Paris, Institut Polytechnique de Paris,[1] SATIE, Université Paris-Saclay,[2]
National University of Singapore,[3] valeo.ai[4]

Figure 1. **Illustration of image synthesis models under different perturbations.** Compared to previous work [101], our results express more natural textures and details, even under anomalies or shifts in the input distribution.

Abstract

Semantic segmentation methods have advanced significantly. Still, their robustness to real-world perturbations and object types not seen during training remains a challenge, particularly in safety-critical applications. We propose a novel approach to improve the robustness of semantic segmentation techniques by leveraging the synergy between label-to-image generators and image-to-label segmentation models. Specifically, we design Robusta*, a novel robust conditional generative adversarial network to generate realistic and plausible perturbed images that can be used to train reliable segmentation models. We conduct in-depth studies of the proposed generative model, assess the performance and robustness of the downstream segmentation network, and demonstrate that our approach can significantly enhance the robustness in the face of real-world perturbations, distribution shifts, and out-of-distribution samples. Our results suggest that this approach could be valuable in safety-critical applications, where the reliability of perception modules such as semantic segmentation is of utmost importance and comes with a limited computational budget in inference. We release our code at github.com/ENSTA-U2IS/robusta.*

[*]equal contribution, [†] gianni.franchi@ensta-paris.fr

1. Introduction

Semantic segmentation is an essential perception task that is commonly used in safety-critical applications such as autonomous driving [68, 74] or medical imaging [73, 78]. To fulfill safety requirements [62], it is crucial to not only produce accurate segments but even more so make reliable predictions in the face of perturbations [46], distribution shifts [52, 96, 105], uncommon situations [115] and out-of-distribution (OOD) objects [11, 44]. Modern Deep Neural Networks (DNNs) achieve impressive performance in segmentation tasks [2, 22, 40, 79, 91]; however, they struggle to generalize to data samples not seen during training, e.g., image corruptions [46, 85], adversarial attacks [21, 102], or change of style [31]. In the face of such events, they often produce overconfident probability estimates [38, 43, 77] even when they are wrong, which can impede the detection of failure modes and thus their adoption in the industry.

Technically, modern high-capacity DNNs trained with uncertainty and calibration awareness could effectively learn to deal with such situations if they were seen during training. However, such samples from the long tail are extremely numerous, difficult, and expensive to acquire in the real world. In the absence of such data, new data augmentation techniques have been devised towards mitigating this problem: generating perturbations on the input data [114, 116], the

latent space [20], or the output space [67]. They may slightly relieve this problem by increasing the size of the training set with richer samples. Still, they may not always yield realistic images and are often architecture-dependent or unsuitable for semantic segmentation. Meanwhile, generative models are making steady progress towards the ambitious goal of generating abundant and high-quality data for training data-hungry DNNs [9, 59, 88, 93]. Despite the realism reflected by excellent FID scores [50], training on such synthetic images is non-trivial and less effective than with real images [6, 89, 99]. Very recent methods based on the latest generative models [42, 97] are starting to show encouraging results for image classification. However, such approaches have not shown yet that they can improve the performance and robustness of semantic segmentation models dealing with high-resolution images and complex scenes.

In this paper, we pave the way for improving the robustness of semantic segmentation models against input perturbations as well as their ability to detect outliers, i.e., objects with no associated labels in the train set. To this end, we propose leveraging the symmetry between label-to-image conditional generative adversarial networks (cGANs) [34, 55] and image-to-label models to train robust segmentation networks [17, 70]. We argue that it is possible to use label-to-image cGANs at training time to enrich datasets with perturbations and outliers and improve segmentation models. The generated data must be of sufficient quality and diversity to enable the classifier - trained on it - to generalize to new and unseen real-world situations and to prevent overfitting to the specific anomalies of the training set.

A significant difficulty we face is that cGANs are not robust to variations in the input labels. They may generate images with artifacts when given label maps with anomalies, hindering their utility for semantic segmentation training. To address this challenge, we advance *Robusta*, a novel cGAN architecture that leverages attention layers [106] and subnetworks [109]. Robusta can produce high-quality images even from label maps with corruptions and anomalies. We exploit Robusta synthetic images to construct a dataset for training an observer network [5] to detect anomalies and failures of the segmentation model.

Contributions. (1) We propose a new strategy to improve the robustness and OOD detection performance of semantic segmentation models by leveraging the symmetry of label-to-image cGANs and image-to-label segmentation networks: we use cGAN images with outliers to train a robust segmentation model; **(2)** We design the first framework to evaluate the robustness of label-to-image cGANs against perturbations and uncommon inputs, and we use it to investigate the robustness of multiple cGAN methods against three new dataset generation techniques; **(3)** We propose *Robusta*, a new cascaded cGAN with improved robustness compared to state-of-the-art cGANs on the proposed framework. Our approach is expected to enhance the reliability and safety of autonomous driving systems by enabling more accurate OOD and failure detection in real-world situations.

2. Related Work

Conditional GANs for label-to-image translation: Various label-to-image translation techniques have been proposed [81] with two main categories: paired strategies [55, 81, 84, 108, 118], where images and labels are aligned, and unpaired strategies [58, 119]. In this work, we focus on paired label-to-image translation models that synthesize RGB images from semantic inputs. Most approaches in this area are based on cGANs [55, 84].

Pix2Pix [55] is a strong baseline and was one of the first cGAN-based techniques applied to label-to-image translation, but it often fails to generate high-quality images using a single GAN. Pix2PixHD [109] upgrades the quality of the generated images by incorporating enhanced multi-scale generators and discriminators.

SPADE (SPatially ADaptive DEnormalization) [84] advances spatially-adaptive normalization layers that leverage semantic label maps to modulate the activations in the normalization layers, resulting in significant improvements for datasets with strong global redundancy. Rather, DP-GAN [98] shares the spatially adaptive normalization parameters, learned at each scale of the down-sampling part, with the up-sampling layers. More recently, OASIS [101] incorporates an adversarial component and pixel-level discriminators, and PITI [108] pre-trains a diffusion model on a large dataset of various images while leveraging the latent space for the downstream tasks.

Anomaly detection for semantic segmentation: Anomaly segmentation is challenging as it requires precise anomaly localization and can be tackled by cGAN-based methods. Various cGAN methods [17, 70, 107, 111] have been proposed to detect anomalies. These methods utilize cGANs at test time, leading to significant runtime costs [104]. In contrast, our approach only uses cGANs in the offline stage to generate synthetic images for training an efficient anomaly-aware segmentation network, avoiding the runtime costs associated with cGAN-based approaches.

Other techniques employ uncertainty quantification methods such as ensembles [60], their efficient derivatives [24, 30, 39, 53, 61, 110], and Monte-Carlo Dropout [27, 28, 76]. Some approaches, such as DNP [29], compute distances between feature representations produced by the segmentation encoder network [90] using nearest-neighbors. Others, like PEBAL [104], utilize abstention learning [72] at the pixel level through energy-based models [65].

Some methods [7, 12, 37] involve using auxiliary data in the wake of Outlier Exposure [49] for classification. Going further, Obsnet [5] generates failures with local adversarial

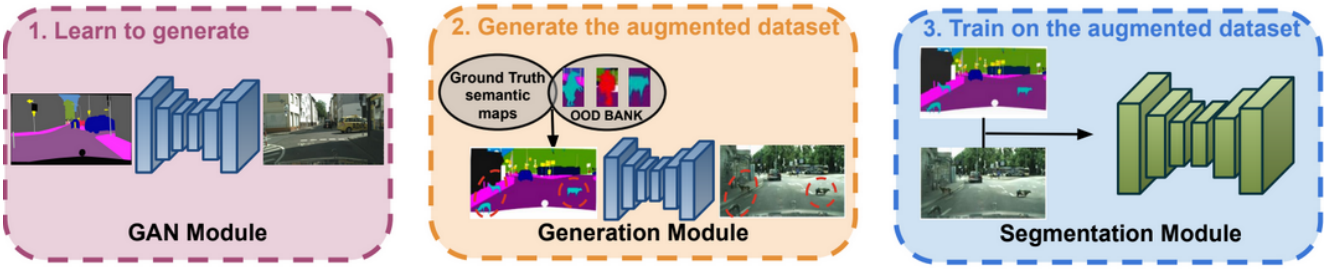

Figure 2. **Illustration of the pipeline.** The pipeline diagram depicts three steps. Firstly, we train the new Robusta model. Secondly, we utilize Robusta to create a diverse and high-quality dataset that includes various objects of interest. Finally, we train a segmentation model on this augmented dataset.

attacks and NflowJS [35, 36], generate a negative dataset using normalizing flows [18, 82, 92]. In contrast, our method employs cGANs to generate additional data with non-typical content and scene layouts, and generates outliers as a proxy for out-of-distribution. This generated data is used downstream to train robust segmentation models.

Dataset synthesis for semantic segmentation: The lack or limited amount of anomalous data in real-world datasets motivates the synthesis of datasets for semantic segmentation. StreetHazards [45] and MUAD [26] are recent works that propose solving this problem with fully synthetic datasets generated with computer graphics simulators.

To increase the volume of data of a given dataset and test the robustness of segmentation models, some researchers propose using GANs to modify images with challenging conditions of interest. For instance, Rainy [105] and Foggy Cityscapes [96] respectively add rain and fog to Cityscapes' urban scenes, and CoMoGAN [86] allows changing the picture's time of the day with continuous transformations.

An alternative to these methods consists in directly generating complete scene layouts. SB-GAN [1] and PGAN-CGAN [51] are recent methods proposed for this task, sequentially generating semantic label maps and converting them into images using classical conditional GANs. SemanticPalette [64] improves upon these methods with a pipeline that controls the relative importance of the different classes in the images.

3. Problem setup and method overview

Our objective is to improve the ability of semantic segmentation models to handle unexpected and uncommon situations and objects that arise due to aleatoric and epistemic uncertainty sources. For the scope of this paper, we define a robust semantic segmentation model as one that can produce accurate predictions even when confronted with uncommon samples coming from a distribution with strong uncertainties, shifted from the original training distribution of the model. To consolidate the robustness of this segmentation model, we introduce a three-step pipeline, which we outline below as well as in Figure 2.

Step 1: We train a new cGAN called Robusta to generate additional data from the original segmentation training set (see section 4.1). **Step 2:** We leverage Robusta to create a new high-quality dataset by conditioning it on label maps with different objects, such as road signs placed in unconventional locations, aiming to expand the diversity of the training set (see section 4.2). **Step 3:** We boost the generalization ability of the segmentation model by training it on this new augmented dataset (see section 5).

To the best of our knowledge, we are the first to explore the robustness of generative models in the context of semantic segmentation for autonomous driving scenes. To this extent, we introduce a novel architecture that enhances the reliability of generated images. Our primary goal is, indeed, to develop a robust label-to-image generative model capable of producing realistic corner-case images to improve the robustness of segmentation networks. To assess the effectiveness of these models and our pipeline (see Figure 2), we employ three benchmarks: first, we quantify the quality of the generated images under non-perturbed conditions (purple block) in Appendix **C**; then we measure the robustness of the generative model with perturbed label-maps (yellow block) in sections 4.2 and 6.2; finally, we evaluate the efficiency of synthetic images in enhancing the robustness of semantic segmentation (blue block) in section 6.3.

4. Enhancing data generation to robustify segmentation

The efficient training and robustification of the semantic segmentation network rely on generating high-quality data (see Appendix **F.1** for more details). This requires generating both in-distribution data with perturbations and different textures, as well as high-quality out-of-distribution data. As outlined in section 3, it is essential to enhance the robustness of the generating models. Specifically, the quality of the output must remain satisfactory even when given inputs outside

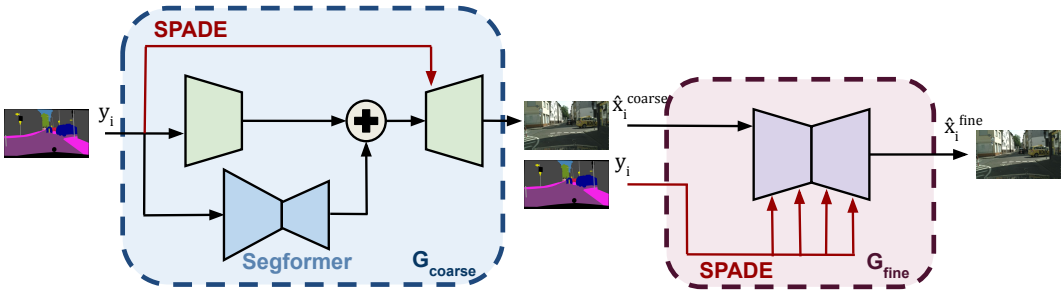

Figure 3. **Illustration of Robusta's generation process.** First, we train the networks G_{coarse}, producing low-resolution images. We add another generator, G_{fine}, to improve the image quality from the output of G_{coarse}. The **+** operation corresponds to a concatenation.

of its training distribution (corrupted label maps). In section 4.1, we propose a novel cGAN-cascade called *Robusta* to improve the robustness of label-to-image translation. We test the usefulness of our method thanks to the first **protocol to evaluate the robustness of generative processes** that we detail in section 4.2. To enhance the robustness of the cGAN, we have integrated attention layers [106] and decomposed the GAN into two sub-GANs.

4.1. Robusta: a new cGAN cascade for improved robustness

4.1.1 Architecture

We propose *Robusta* a novel architecture based on cGANs designed to address the artifacts presented in Figure 1 that arise when generating outlying label maps in the context of label-to-image translation. To overcome these artifacts, we decompose our translation cascade into two generators G_{coarse} and G_{fine}, as illustrated in Figure 3, that are trained sequentially. We design G_{coarse} for handling artifacts related to label-to-image translation and G_{fine} artifacts related to output image quality.

The architecture of G_{coarse} is based on Pix2PixHD [109] and includes two encoder-decoders (in blue and green in Figure 3). These encoder-decoders have the same input dimensions, but the output of the blue module has the size of the latent space of the green module. For better robustness and performance, we choose Segformer [112] for the blue module as encoder-decoder, which has proven effective in semantic-segmentation tasks. Segformers are based on ViTs [19] and encode inputs in feature maps, which promotes contextual information learning [32, 69]. Moreover, ViTs have been shown to be more robust learners [3, 4, 8]. We think that the favorable Transformer's robustness also extends to image generation, and support this hypothesis with the ablation studies conducted in Appendix **E.1**. We argue that this approach is particularly relevant for tackling translation artifacts. Additionally, we incorporate a SPADE [84] layer after the concatenation of the output of the Segformer (in blue) and the encoding of the main generator (in green) to re-inject lost spatial semantic information before decoding.

To further improve the quality of the final image, we add a generator, G_{fine}, which takes as input the images provided by G_{coarse} and generates high-quality outputs. In G_{fine}, we substitute all batch normalizations [54] with spatially-adaptive normalization (SPADE) [84] layers to accurately incorporate the information present in semantic label maps.

To obtain a better image quality with higher-frequency textures [83] (see Appendix **E.1.3** for more details), we use U-Net [94, 95], a residual convolutional neural network [41]. G_{fine} is trained separately from G_{coarse} using the loss (3), G_{coarse}'s weights being frozen at this time.

Appendix **A** details the links between Pix2Pix, Pix2PixHD, SPADE, and Robusta, and Appendix **E.1** provides ablation studies on the architecture.

4.1.2 Training

Standard procedure. Let $F_{\boldsymbol{\theta}}(\mathbf{x})$ denote the application of a DNN F of weights $\boldsymbol{\theta}$ to an input image $\mathbf{x}$, $D_{\boldsymbol{\theta}}(\cdot)$ a discriminator and $G_{\boldsymbol{\theta}}(\cdot)$ a generator. The weights of the discriminator and the generator are denoted by $\boldsymbol{\theta}_D$ and $\boldsymbol{\theta}_G$ respectively and are omitted when obvious. Label-to-image models consist of cGANs trained with the following losses:
(a) A GAN loss:

$$\mathcal{L}^{\text{BCE}}_{\text{cGAN}}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_D) = \mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x} \mid \mathbf{y})] + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z} \mid \mathbf{y})))], \quad (1)$$

where $\mathbf{y}$ is the conditional information. In label-to-image translation, $\mathbf{y}$ is the input label, $\mathbf{x}$ is the target RGB image, and $\mathbf{z}$ is the sampled latent variable.
(b) An L1 loss between the output of the generator and the target images, denoted $\mathcal{L}_{L1}(\boldsymbol{\theta}_G)$.
(c) A feature-matching loss $\mathcal{L}_{\text{FM}}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_D)$ measuring the L1 norm between the feature maps of the real and generated images extracted from the layers of the discriminator D.
(d) A perceptual loss $\mathcal{L}_{\text{VGG}}$ measuring the L1 norm between

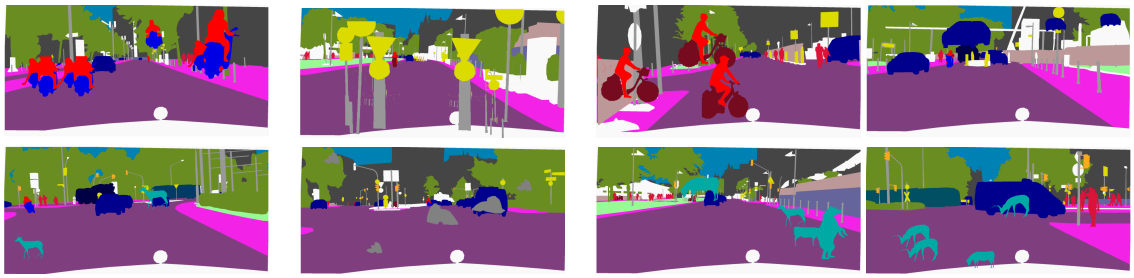

Figure 4. **Perturbed label maps.** Different label maps from Corrupted-Cityscapes (top) and Outlier-Cityscapes (bottom).

the feature maps of the real and generated images extracted from the layers of a VGG network [100] pre-trained on ImageNet [16].

Training Robusta. We train Robusta in two steps beginning with the generator G_{coarse}. To do so, we optimize the loss $\mathcal{L}_{\text{coarse}}$ on a dataset made of the original label maps as inputs and original images as targets. We minimize the linear combination of losses (2) of Wang et al. [109].

$$\mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{cGAN}}^{\text{BCE}} + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{FM}\mathcal{L}_{\text{FM}} + \lambda_{\text{VGG}}\mathcal{L}_{\text{VGG}}. \quad (2)$$

When the training of G_{coarse} is complete, the synthesized images are used to feed and train the second generator G_{fine}, which does not use $\mathcal{L}_{L1}$ and is based on $\mathcal{L}_{\text{cGAN}}^{\text{MSE}}$ instead of $\mathcal{L}_{\text{cGAN}}^{\text{BCE}}$ as proposed for SRGAN [66]. This corresponds to:

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{cGAN}}^{\text{MSE}} + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}} + \lambda_{\text{VGG}}\mathcal{L}_{\text{VGG}}. \quad (3)$$

Please refer to the Appendix **A** for details on the losses and the definition $\mathcal{L}_{\text{cGAN}}^{\text{MSE}}$, and to Appendix **B** for training details. Furthermore, Appendix **E.2** includes ablation studies on Robusta's training, and Appendix **F.3** details the training time and the computational cost of Robusta.

4.2. Robustness assessment protocol for label-to-image translation

To the best of our knowledge, we present the first protocol to evaluate the robustness of label-to-image generators quantitatively. This type of network is susceptible to perturbations in input data, prompting us to design three novel methodologies to gauge the efficacy of label-to-image generators derived from semantic segmentation datasets. These methodologies involve modifying the original dataset to quantify the robustness of generators against diverse sources of corruption.

When working on Cityscapes [14] (CS), we can process the images of the validation set and produce three versions: *Corrupted-CS*, *Outlier-CS*, and *Morphological-CS*. Note that the following processes naturally extend any other semantic segmentation dataset.

Corrupted-CS randomly adds Cityscapes objects, such as road signs and cyclists, into the label maps. A robust label-to-image network should be able to properly render the additional objects that are included in its training set.

Outlier-CS adds objects that do not belong to the distribution of the original dataset. As Cityscapes is composed of urban scenes, we add shapes of bears, deer, cows, and rocks. With this dataset, we evaluate the capacity of the network to generate renderings from unknown shapes.

Last, *Morphological-CS* corrupts the segmentation maps by applying morphological operators on the labels. We detail the morphological operations in Appendix **D**. As for CIFAR-C [46] and Cityscapes-C [75], we study the robustness of the network to varying intensities of perturbations of the input images. In our case, this corresponds to increasing the size of the structuring element of the morphological dilation.

In these three variations, we expect a robust label-to-image translator to produce reliable and high-quality outputs even when perturbed by modified label maps. Hence, we measure the quality of the generated images using the Fréchet Inception Distance (FID). We also assess the semantic segmentation performance of a pre-trained model on the original dataset but inferring on the modified version, with the mean Intersection over Union (mIoU). This value should remain high if the rendering conveys the semantic meaning correctly. Corrupted label maps are available in Figure 4.

Our method generates artifacts and out-of-domain object maps, but we claim that they are relevant to enrich training data and robustify training. We view these perturbations of the true label maps distributions as targeting the critical long-tail anomalies. For instance, flying cars could be related to car crashes in real-world images. In our training pipeline (on Figure 2), achieving good quality on these datasets is crucial for the subsequent training of the segmentation methods. Therefore, we use these datasets as a benchmark for evaluating the robustness of our generator models.

The generating process of the conditional GAN is thoroughly explained in Appendix **A**. Additionally, section 5 provides a detailed account of the techniques used for generating images of the *Outlier-CS* and *Corrupted-CS* datasets.

5. New training sets for robust segmentation

Thanks to Robusta, it is now possible to generate high-quality data with outliers and perturbations to extend existing segmentation datasets while limiting artifacts. The primary goal is to harness this generated data and refine the training process to obtain more resilient segmentation models.

To train segmentation networks, we begin with a dataset of training examples $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{|\mathcal{D}|}$ of $|\mathcal{D}|$ pairs of images $\mathbf{x}_i \in \mathbb{R}^{C\times H\times W}$ and label maps $\mathbf{y}_i \in [\![0, N_C[\![^{H\times W}$ modeled as a realization of a joint distribution $\mathcal{P}(X, Y)$. We denote C, H, and W as the number of channels, the height, and the width of the image, respectively, and N_C is the number of classes in the dataset.

As described in 3, the generation pipeline G detailed in Figure 3 is designed to generate realistic images that correspond to the input label maps $\mathbf{y}_i$, producing a $\hat{\mathbf{x}}_i$ image. In this section, we aim to use G to generate a synthetic dataset as depicted in Figure 2. We use Cityscapes as our original training set for illustration. In section 4.2, we previously introduced *Corrupted-CS* and *Outlier-CS*. Here, we provide more details about their use and definition. In the following subsections, we clarify how G generates these two sets.

Improving generalization with *Corrupted-CS*: To enhance robustness, we generate *Corrupted-CS* using Robusta. First, we gather multi-dimensional masks of objects belonging to a particular set of classes $\mathcal{C}$. These classes may consist of various combinations, such as traffic signs and poles, traffic lights and poles, motorcycles and riders, and bicycles and riders. The collected masks are then sorted into sets $\mathcal{D}^{\text{OI}} \triangleq \{y_i^{\text{OI}}\}$, where each y_i^{OI} is an element of the interval $[\![0, N_C[\![^{H\times W}$. All pixels of y_i^{OI} except for the object's pixel are set to zero, whereas the object's pixel is assigned the label's object value. Then, we expand $\mathcal{D}^{\text{OI}}$ by applying translations to its masks. Next, we use a mixing function to combine the label maps of the original dataset $\mathcal{D}$, creating a new label $\mathbf{y}_i^{\text{mix}}$ for each element i in the range $[1, |\mathcal{D}|]$. The definition of this new label is as follows:

$$\mathbf{y}_i^{\text{mix}} = \left[\mathbf{y}_j^{\text{OI}} = 0\right] * \mathbf{y}_i + \mathbf{y}_j^{\text{OI}}, \text{ with } j \in [1, |\mathcal{D}^{\text{OI}}|] \quad (4)$$

This results in a new set of label maps denoted as $\mathcal{D}_y^{\text{mix}}$. Finally, we combine $G(\mathcal{D}_y^{\text{mix}})$ for the images and $\mathcal{D}_y^{\text{mix}}$ for the labels to create the final dataset $\mathcal{D}^{\text{mix}} = \{G(\mathbf{y}_i^{\text{mix}}), \mathbf{y}_i^{\text{mix}}\}_{i=1}^{|\mathcal{D}|}$.

To improve the generalization of the semantic segmentation model, we enrich the original dataset $\mathcal{D}$ with $\mathcal{D}^{\text{mix}}$, and we keep the same training procedure.

Improving outlier detection with *Outlier-CS*: To enhance the detection of outliers, we craft Outlier-CS. We apply the same mixing strategy as before, except that we no longer extract known objects to construct $\mathcal{D}^{OI}$. Instead, we choose a set of outlier objects from Cityscapes, such as bears, cows, and rocks. The shape of these instances is derived from the MUAD dataset [26]. Since the outliers do not possess labels corresponding to the Cityscapes label space, we value them as "humans" using the aforementioned outlier shapes.

To improve out-of-distribution detection, we can also use the outlier labels to train the networks directly with a binary cross-entropy (BCE) loss [15] and refer to this method as "Robusta + BCE". In the following, we test the efficiency of adding this dataset for Obsnet-like [5] architectures on OOD detection tasks.

Contrary to previous work by Ghiasi et al. [33], which requires finely labeled object instances, with Robusta, we can generate instances even from unknown shapes, eliminating the need for expensive fine-grained annotations of OOD instances. These annotations can be more expensive than segmentation labels, especially in our scenario.

6. Experiments

We conduct various experiments to validate the effectiveness of Robusta and our methods. One such experiment, described in section 6.1, demonstrates that the images generated by Robusta are of comparable quality to those produced by the state-of-the-art (SOTA) label-to-image translation. For a more detailed discussion on this topic, please refer to Appendix **C**. In addition, we assess the quality of Robusta-generated images under different input corruptions in section 6.2. We also evaluate the impact of using Robusta-generated data on the robustness of semantic segmentation algorithms and OOD detection in section 6.3.

6.1. Quality of Robusta's images

To evaluate the quality of the images generated by our cGAN-cascade, we adopt the evaluation protocol used in previous studies on label-to-image translation [84, 101]. Specifically, we convert label maps into RGB synthetic images and measure the FID and mIoU, expressed in %. We report experimental details and exhaustive comparisons with current SOTA methods and baselines [63, 71, 84, 87, 103], including OASIS [101], on multiple datasets [10, 14, 57, 84, 117] in Appendix **C**. Our experiments show that we achieve equivalent or superior performance compared to current methods on most datasets. We also provide qualitative assessments of the synthesized images in Appendix **G** using numerous visual examples to provide a more comprehensive evaluation.

6.2. Robustness of the cGAN cascade

We evaluate the robustness of our pipeline using the protocol proposed in Section 4.2 on Cityscapes [14] and ADE20K [117]. The results are summarized in Table 1 and show that our GAN-cascade performs at least as well as other SOTA methods on most datasets. Qualitatively, in Figure 1, we observe that the SOTA algorithms struggle to

Method	Corrupted-CS		Corrupted-ADE20K		Outlier-CS		Outlier-ADE20K		Morphological-CS		Morphological-ADE20K	
	FID ↓	mIoU ↑	FID ↓	mIoU ↑	FID ↓	mIoU ↑	FID ↓	mIoU ↑	FID ↓	mIoU ↑	FID ↓	mIoU ↑
SPADE	77.43	42.70	46.25	33.2	91.84	59.21	52.01	37.7	61.75	59.62	40.12	38.5
OASIS	76.88	41.79	45.86	44.1	90.04	62.02	67.29	47.1	51.90	62.89	39.03	48.7
Robusta G_{coarse}	76.48	42.04	40.84	42.7	84.66	60.43	45.08	45.5	53.57	58.09	35.67	46.1
Robusta $(G_{\text{coarse}}, G_{\text{fine}})$	**75.27**	**43.31**	**38.08**	**44.6**	**79.48**	**62.78**	**43.21**	**47.3**	**50.60**	**63.37**	**33.91**	**48.8**

Table 1. Comparative results for the **robustness tasks** presented in 4.2.

generate images when given perturbed label maps, indicating that these algorithms overfit strongly to their training dataset and lack robustness to changes in the input data. This highlights the importance of investigating the robustness of image-generation algorithms.

6.3. Evaluation of the segmentation robustness

We conduct experiments over five datasets to assess the network's robustness to uncertainties. First, we do experiments on StreetHazards [44] and BDD anomaly [44] to evaluate the network's ability to detect OOD classes. The test set includes some object classes absent from the training set. To perform this task, we trained a deep neural network called Obsnet [5] using the *Outlier-StreetHazards* and *Outlier-BDD anomaly* datasets generated by Robusta. We provide more information on these datasets in subsection 6.3.1. *Furthermore*, we investigate the network's reliability for the aleatoric uncertainty by training a DNN on *Corrupted-CS*, generated by Robusta. We then evaluate the network's performance on three datasets used for this task: Rainy [52] and Foggy Cityscapes [96], and Cityscapes-C [47].

Metrics. We use several criteria to evaluate the performance of the segmentation models. The first criterion, the mIoU, measures the predictive performance of the networks in segmentation accuracy, as introduced by Jaccard [56]. Another criterion is the negative log-likelihood (NLL), which is a proper scoring rule based on the aleatoric uncertainty, similar to the approach described by Lakshminarayanan et al. [60]. Additionally, we employ the expected calibration error (ECE) [38] to evaluate the grounding of the DNN's top-class confidence scores. Furthermore, we assess the DNN's ability to detect OOD data using the AUPR, AUROC, and the False Positive Rate at 95% recall (FPR95), as defined by Hendrycks et al. [48]. By considering multiple metrics, we obtain a more comprehensive and precise assessment of the DNN's performance in accuracy, calibration error, failure rate, and OOD detection. Although it may be challenging to achieve top performance on all metrics, we argue that evaluating multiple metrics, as supported by research [23, 80], is more practical and convincing than optimizing for a single metric at the potential expense of other factors.

6.3.1 Out-of-Distribution detection experiments

Datasets. We conduct a study on outlier detection using StreetHazards [44], BDD Anomaly [44] , and Road Anomaly [11, 70]. StreetHazards is a large-scale dataset comprising various sets of synthetic images of street scenes. It contains $5,125$ images for training and $1,500$ test images, with pixel-wise annotations for 13 classes in the training set. The test set comprises 13 training classes and 250 out-of-distribution (OOD) classes unseen in the training set. This enables us to test the algorithm's robustness when facing diverse scenarios. The BDD Anomaly dataset is a subset of the BDD [113] dataset and contains 6,688 street scenes in the training set and 361 in the testing set. The training set includes 17 classes, and the test set consists of these 17 training classes as well as 2 OOD classes. Finally, RoadAnomaly contains 100 high-resolution test images of unusual dangers encountered in real life, such as giraffes, cattle, boats, etc. The images are associated with binary per-pixel labels for the background and the anomalies.

Architecture. In this experiment, we employ Obsnet [5] and DeepLabv3+ [13] with a ResNet50 [41] as backbone, following the experimental protocol presented in [44]. Specifically, we adopt the evaluation criteria introduced by Hendrycks et al. [44]. Obsnet is a deep neural network designed to detect faults, and we train it using the *Outlier-StreetHazards* and *Outlier-BDD Anomaly* datasets generated by Robusta. To create the new datasets, we introduce additional OOD objects labeled as "cars" to the original label maps and use Robusta to generate the corresponding images, as explained in Section 5. In order to evaluate Robusta's performance against other baselines, we also train a Binary segmentation model (denoted BCE + Robusta) to identify outliers using *Outlier-BDD Anomaly* generated by Robusta. To ensure consistency with the other baselines, we employ the same DeepLabv3+ segmentation model.

Baselines. For this experiment, we evaluate our algorithm against several SOTA methods, including Deep Ensembles [60], BatchEnsemble [110], LP-BNN [24], TRADI [25], MIMO [39], and Obsnet [5], on epistemic uncertainty. In addition, we include the baseline, Maximum Class Probability (MCP), which considers the maximum probability as a confidence score.

Results. We show in Table 2 that our method outperforms all other methods on three out of four performance measures, achieving SOTA results on epistemic uncertainty. Moreover, our approach is faster than Deep Ensemble and LP-BNN, as it only requires a single inference pass, while the others need four. We also include results on RoadAnomaly [11, 70] in

Method	StreetHazards			
	mIoU ↑	AUROC ↑	AUPR ↑	FPR95 ↓
Baseline (MCP) [48]	53.90	86.60	6.91	35.74
TRADI [25]	52.46	87.39	6.93	38.26
Deep Ensembles [60]	55.59	87.94	8.32	30.29
MIMO [39]	55.44	87.38	6.90	32.66
BatchEnsemble [110]	**56.16**	88.17	7.59	32.85
LP-BNN [24]	54.50	88.33	7.18	32.61
Obsnet [5]	53.90	94.96	10.58	16.74
Obsnet + LA [5]	53.90	95.37	10.91	15.78
BCE + Robusta (Ours)	53.90	95.91	13.58	**13.05**
Obsnet + Robusta (Ours)	53.90	**96.27**	**15.60**	14.81
	BDD Anomaly			
Baseline (MCP) [48]	47.63	85.15	4.50	28.78
TRADI [25]	44.26	84.80	4.54	36.87
Deep Ensembles [60]	**51.07**	84.80	5.24	28.55
MIMO [39]	47.20	84.38	4.32	35.24
BatchEnsemble [110]	48.09	84.27	4.49	30.17
LP-BNN [24]	49.01	85.32	4.52	29.47
Obsnet [5]	47.63	87.66	1.01	19.50
Obsnet + LAA [5]	47.63	88.16	1.71	21.71
BCE + Robusta (Ours)	47.63	92.99	2.45	20.06
Obsnet + Robusta (Ours)	47.63	**96.86**	**5.53**	**14.51**

Table 2. **Comparative results on the OOD task for semantic segmentation.** The architecture is a DeepLabv3+ based on ResNet50.

Method	AUPR ↑	FPR95 ↓	SIoUgt ↑	PPV ↓	mF1 ↑
Baseline (MCP) [48]	28.0	72.1	15.5	15.3	5.4
Deep Ensembles [60]	17.7	91.1	16.4	20.8	3.4
Obsnet + LAA [5]	**75.4**	**26.7**	44.2	52.6	45.1
BCE + Robusta (Ours)	70.3	45.3	**48.2**	**57.6**	**48.2**

Table 3. **Comparative results on RoadAnomaly [70].** These methods are trained without any OOD data.

Table 3, where we show that Robusta improves over standard Obsnet on the component-level metrics, all defined in the benchmark accompanying paper [11].

6.3.2 Aleatoric uncertainty experiments

The objective of our study is to evaluate the ability of DNNs to handle aleatoric uncertainty in semantic segmentation. We utilize Rainy [52] and Foggy Cityscapes [96], which involve the introduction of rain or fog to the Cityscapes validation dataset. We generate aleatoric uncertainties on the Cityscapes validation set to create Cityscapes-C [75] with varying levels of perturbations using different methods such as Gaussian noise, shot noise, impulse noise, defocus blur, frosted, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic, pixelate, and JPEG. We used the code of Hendrycks et al. [47] to generate these perturbations.

To assess the reliability of the DNNs in the presence of aleatoric uncertainty, we measure the ECE, mIoU, and NLL in Table 4. We obtain results comparable to current methods, with Deep Ensembles performing better than the other approaches. However, Deep Ensembles require training multiple DNNs, which is more time-consuming for both training and inference.

Method	Cityscapes			Rainy Cityscapes		
	mIoU ↑	ECE ↓	NLL ↓	mIoU ↑	ECE ↓	NLL ↓
Baseline (MCP) [48]	76.51	0.1303	-0.9456	58.98	0.1395	-0.8123
MIMO [39]	77.13	0.1398	**-0.9516**	59.27	0.1436	-0.8135
BatchEnsemble [110]	77.99	0.1129	-0.9472	60.29	0.1436	-0.7820
LP-BNN [24]	77.39	**0.1105**	-0.9464	60.71	0.1338	-0.7891
Deep Ensembles [60]	77.48	0.1274	-0.9469	59.52	**0.1078**	-0.8205
MCP + Robusta (Ours)	**78.41**	0.1211	**-0.9546**	**62.31**	0.1254	**-0.8382**
	Foggy Cityscapes			Cityscapes-C		
Baseline (MCP) [48]	69.89	0.1493	-0.9001	40.85	0.2242	-0.7389
MIMO [39]	70.24	0.1425	-0.9014	40.73	0.2350	-0.7313
BatchEnsemble [110]	**72.19**	0.1425	**-0.9132**	40.93	0.2270	-0.7082
LP-BNN [24]	**72.39**	**0.1358**	**-0.9131**	**43.47**	0.2085	-0.7282
Deep Ensembles [60]	71.43	0.1407	-0.9070	**43.40**	**0.1912**	-0.7509
MCP + Robusta (Ours)	**72.01**	0.1438	**-0.9110**	**43.49**	0.2497	**-0.7578**

Table 4. **Evaluation of the influence of the quality of the training dataset.** We trained Deeplabs V3+ with ResNet101 backbones on CS (and CS + Corrupted-CS for MCP + Robusta) and tested the models on different corrupted variants of the original validation set.

7. Discussions and details

A comprehensive exposition of Robusta, including its implementation details, can be found in Appendices **A** and **B**. Appendix **C** evaluates Robusta as a label-to-image translation model and compares its performance with other state-of-the-art algorithms. Appendix **D** provides additional details on morphology and the *morphological* dataset derivation process. Extensive ablation experiments are conducted in Appendix **E** to demonstrate the significance of each contribution of Robusta. Finally, Appendix **F** highlights the trade-off between computational cost and performance.

8. Conclusion

In this work, we propose a novel approach to enhancing the robustness of deep semantic segmentation networks by generating new textures and outlier objects from existing label maps. For best performance, we propose a new robust label-to-image model called Robusta, which incorporates transformers and two conditional GANs. We verify that it is able to produce high-quality images even when asked to generate outliers, thanks to a new robustness evaluation framework for generative models, on which Robusta performs better than current methods.

Furthermore, we demonstrate that using the datasets generated with Robusta during training helps improve the robustness of semantic segmentation algorithms. We show significant improvements in the performance of segmentation models when given perturbed inputs as well as for out-of-distribution detection.

Acknowledgments

This work was supported by AID Project ACoCaTherm and performed using HPC resources from GENCI-IDRIS (Grant 2022-AD011011970R2).

References

[1] Samaneh Azadi, Michael Tschannen, Eric Tzeng, Sylvain Gelly, Trevor Darrell, and Mario Lucic. Semantic bottleneck scene generation. *arXiv preprint arXiv:1911.11357*, 2019. 3

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 2017. 1

[3] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *NeurIPS*, 2021. 4

[4] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. *arXiv preprint arXiv:2110.02797*, 2021. 4

[5] Victor Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. In *ICCV*, 2021. 2, 6, 7, 8

[6] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP*, 2020. 2

[7] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *DAGM GCPR*, 2019. 2

[8] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *ICCV*, 2021. 4

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[10] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 6

[11] 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. In *NeurIPS - Datasets and Benchmarks Track*, 2021. 1, 7, 8

[12] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *ICCV*, 2021. 2

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 7

[14] 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. In *CVPR*, 2016. 5, 6

[15] David R Cox. The regression analysis of binary sequences. *JRSS*, 1958. 6

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[17] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *CVPR*, 2021. 2

[18] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 3

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[20] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *ICLR*, 2022. 2

[21] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, 2018. 1

[22] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 2012. 1

[23] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. In *arXiv preprint arXiv:1912.02757*, 2019. 7

[24] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Séverine Dubuisson, and Isabelle Bloch. Encoding the latent posterior of bayesian neural networks for uncertainty quantification. *arXiv preprint arXiv:2012.02818*, 2020. 2, 7, 8

[25] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Séverine Dubuisson, and Isabelle Bloch. Tradi: Tracking deep neural network weight distributions. In *ECCV*, 2020. 7, 8

[26] Gianni Franchi, Xuanlong Yu, Andrei Bursuc, Rémi Kazmierczak, Séverine Dubuisson, Emanuel Aldea, and David Filliat. Muad: Multiple uncertainties for autonomous driving benchmark for multiple uncertainty types and tasks. *BMVC*, 2022. 3, 6

[27] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2

[28] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *NeurIPS*, 2017. 2

[29] Silvio Galesso, Max Argus, and Thomas Brox. Far away in the deep space: Nearest-neighbor-based dense out-of-distribution detection. *arXiv preprint arXiv:2211.06660*, 2022. 2

[30] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *NeurIPS*, 2018. 2

[31] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[32] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022. 4

[33] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 6

[34] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[35] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Dense open-set recognition with synthetic outliers generated by real nvp. *arXiv preprint arXiv:2011.11094*, 2020. 3

[36] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Dense anomaly detection by robust learning on synthetic negative data. *arXiv preprint arXiv:2112.12833*, 2021. 3

[37] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *ECCV*, 2022. 2

[38] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 1, 7

[39] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *ICLR*, 2020. 2, 7, 8

[40] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 7

[42] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023. 2

[43] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 2019. 1

[44] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. A benchmark for anomaly segmentation. In *arXiv preprint arXiv:1911.11132*, 2019. 1, 7

[45] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2019. 3

[46] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 1, 5

[47] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 7, 8

[48] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 7, 8

[49] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 2

[50] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 2

[51] Jonathan Howe, Kyle Pula, and Aaron A Reite. Conditional generative adversarial networks for data augmentation and adaptation in remotely sensed imagery. In *Applications of Machine Learning*, 2019. 3

[52] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *CVPR*, 2019. 1, 7, 8

[53] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *ICLR*, 2017. 2

[54] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[55] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[56] Paul Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 1912. 7

[57] Jinyong Jeong, Younggun Cho, Young-Sik Shin, Hyunchul Roh, and Ayoung Kim. Complex urban dataset with multi-level sensors from highly diverse urban environments. *IJRR*, 2019. 6

[58] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. Tsit: A simple and versatile framework for image-to-image translation. In *ECCV*, 2020. 2

[59] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 2

[60] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 2, 7, 8

[61] Olivier Laurent, Adrien Lafage, Enzo Tartaglione, Geoffrey Daniel, Jean-Marc Martinez, Andrei Bursuc, and Gianni Franchi. Packed ensembles for efficient uncertainty estimation. In *ICLR*, 2023. 2

[62] Alexander Lavin, Ciarán M Gilligan-Lee, Alessya Visnjic, Siddha Ganju, Dava Newman, Sujoy Ganguly, Danny Lange, Atílím Güneş Baydin, Amit Sharma, Adam Gibson, et al. Technology readiness levels for machine learning systems. *Nature Communications*, 2022. 1

[63] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, 2020. 6

[64] Guillaume Le Moing, Tuan-Hung Vu, Himalaya Jain, Patrick Pérez, and Matthieu Cord. Semantic palette: Guiding scene generation with class proportions. In *CVPR*, 2021. 3

[65] Yann LeCun, Sumit Chopra, M Ranzato, and F-J Huang. Energy-based models in document recognition and computer vision. In *ICDAR*, 2007. 2

[66] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 5

[67] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018. 2

[68] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J. Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, Michael Sokolsky, Ganymed Stanek, David Stavens, Alex Teichman, Moritz Werling, and Sebastian Thrun. Towards fully autonomous driving: Systems and algorithms. In *IV*, 2011. 1

[69] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual transformer networks for visual recognition. *TPAMI*, 2022. 4

[70] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *ICCV*, 2019. 2, 7, 8

[71] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *NeurIPS*, 2019. 6

[72] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. *NeurIPS*, 2019. 2

[73] Kira Maag, Matthias Rottmann, and Hanno Gottschalk. Time-dynamic estimates of the reliability of deep semantic segmentation networks. In *ICTAI*, 2020. 1

[74] RT McAllister, Yarin Gal, Alex Kendall, Mark Van Der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. In *IJCAI*, 2017. 1

[75] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 5, 8

[76] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 2

[77] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015. 1

[78] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. In *MIDL*, 2018. 1

[79] Marin Oršić and Siniša Šegvić. Efficient semantic segmentation with pyramidal fusion. *PR*, 2021. 1

[80] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019. 7

[81] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *TMultimedia*, 2021. 2

[82] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *JMLR*, 2021. 3

[83] Namuk Park and Songkuk Kim. How do vision transformers work? *ICLR*, 2022. 4

[84] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2, 4, 6

[85] Zachary Pezzementi, Trenton Tabor, Samuel Yim, Jonathan K. Chang, Bill Drozd, David Guttendorf, Michael Wagner, and Philip Koopman. Putting image manipulations in context: Robustness testing for safe perception. In *SSRR*, 2018. 1

[86] Fabio Pizzati, Pietro Cerri, and Raoul de Charette. Comogan: continuous model-guided image-to-image translation. In *CVPR*, 2021. 3

[87] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *CVPR*, 2018. 6

[88] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[89] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In *NeurIPS*, 2019. 2

[90] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *CVPR*, 2021. 2

[91] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 1

[92] Oren Rippel and Ryan Prescott Adams. High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125*, 2013. 3

[93] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[94] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[95] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 4

[96] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 2018. 1, 3, 7, 8

[97] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning (s) from a synthetic imagenet clone. *arXiv preprint arXiv:2212.08420*, 2022. 2

[98] Juergen Gall Shijie Li, Ming-Ming Cheng. Dual pyramid generative adversarial networks for semantic image synthesis. In *BMVC*, 2022. 2

[99] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *ECCV*, 2018. 2

[100] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[101] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *ICLR*, 2021. 1, 2, 6

[102] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[103] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*, 2020. 6

[104] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *ECCV*, 2022. 2

[105] Maxime Tremblay, Shirsendu Sukanta Halder, Raoul De Charette, and Jean-François Lalonde. Rain rendering for evaluating and improving robustness to bad weather. *International Journal of Computer Vision*, 2021. 1, 3

[106] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2, 4

[107] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly detection by partial image reconstruction with segmentation coupling. In *ICCV*, 2021. 2

[108] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 2

[109] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2, 4, 5

[110] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In *ICLR*, 2019. 2, 7, 8

[111] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *ECCV*, 2020. 2

[112] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 4

[113] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 7

[114] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *CVPR*, 2019. 1

[115] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash-creating hazard-aware benchmarks. In *ECCV*, 2018. 1

[116] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1

[117] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 6

[118] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *CVPR*, 2021. 2

[119] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2